Data Mining and Its Application in Traditional Chinese Medicine

Xinyou ZHANG

Jiangxi University of Traditional Chinese Medicine

ISCI **Published by ISCI Publishing**

Data Mining and Its Application in Traditional Chinese Medicine

Jiangxi University of Traditional Chinese Medicine

ISCI **Published by ISCI Publishing**

First published in August 2020

ISBN 978-1-78900-104-4 (online)
ISBN 978-1-78900-105-1 (print)

ISCI Publishing LTD.
Kemp House
160 City Road
London EC1V 2NX, UK
www.iscipublishing.org

Author Brief Introduction

Xinyou ZHANG, PHD, Professor, PHD advisor, Dean of Graduate School of Jiangxi University of Traditional Chinese Medicine, Vice Director of Graduate Education Research Association of National CM Higher Education Association, Deputy Director of CM Committee of College Physics Curriculum Teaching Steering Committee of the Ministry of Education, Director of Traditional Chinese Medicine Physics and Chinese Medicine Engineering Committee, Executive Director of Association of Fundamental Computing Education in Chinese Universities, Expert of Education Supervision and Evaluation of Jiangxi Province.

Xinyou ZHANG is engaged in the areas of research in traditional Chinese medicine of quantum information, data analysis and data mining, as well as the research and teaching in Physics and Medical Literature Retrieval. He has been in charge of more than 20 National Natural Science Fund Projects and National Educational Reform Topics. 3 courses which he teaches are regarded as high-quality courses, high-quality resource sharing courses and graduate student high-quality courses in Jiangxi province. In addition, he is the manager of Innovative Teaching Studio of Teachers and Masters. He is the chief editor of more than 20 national planned textbooks of “13th Five-Year Plan” such as Physics, Medical Image Processing and Literature Retrieval of Pharmaceutical Science, *etc*. He has written more than 40 monographs. Among them, Science of Medical Physics, whose chief editor is him, was awarded the first prize of outstanding teaching material. In addition, he won the first prize of Jiangxi provincial teaching achievement for 4 items and the second prize of Jiangxi provincial teaching achievement for 2 items. More than 150 academic papers written by Xinyou Zhang have been published in SCI journals and national core journals.

(***E-mail***: xinyouzhang@jxutcm.edn.cn **OR** xinyouzhang@163.com)

Preface

With the development of information technology, the data mining technology has been widely applied in the field of traditional Chinese medicine. In order to promote the informatization, modernization and internationalization of traditional Chinese medicine, the author wrote the book of "Data Mining and Its Application in Traditional Chinese Medicine". This book introduces the research results of National Natural Science Foundation project (Research of Data Mining Method for the Herbal Nature and Efficacy of Traditional Chinese Medicine Based on Strategy Pattern, Xinyou Zhang, Grant NO. 81660727) undertaken by the author. The book comprehensively introduces the data mining, as well as the technology, methods and latest research results of its application in the field of traditional Chinese medicine. In addition, techniques and methods described in the book can be used as the reference for researchers of traditional Chinese medicine information.

The book has six chapters. Chapter 1 Basic Knowledge of Data Mining in Traditional Chinese Medicine. Chapter 2 Application and Evaluation of Data Mining Algorithms in Traditional Chinese Medicine. Chapter 3 Application and Examples of Algorithms in Traditional Chinese Medicine. Chapter 4 Application and Examples of Neural Network Algorithm in The Field of Traditional Chinese Medicine. Chapter 5 Research and System Design of Traditional Chinese Medicine Data Mining Based on Strategy Pattern. Chapter 6 The Latest Research Achievements of Data Mining in Traditional Chinese Medicine.

Xinyou Zhang

1 December, 2019

Acknowledgements

The research was supported by the major program of National Natural Science Foundation of China (Xinyou Zhang, Grant No. 81660727).

Table of Contents

Chapter 1 Basic Knowledge of Data Mining in Traditional Chinese Medicine

1.1 The overview of data mining overview

1.1.1 The emergence of data mining

Data mining is a new thinking method and technical approach for large-scale data processing. It emerges under the condition that the amount of various data in real life exponentially increases and the information technology with database technology as the core gradually matures. Data mining can help users find the hidden rules and patterns in large databases. It integrates theories, methods and technologies of many disciplines, such as artificial intelligence, statistics, machine learning, pattern recognition, and database theory. In recent years, it has been widely used in many different types of organizations and fields such as business, enterprise, government, scientific research and sports, especially in the field of Traditional Chinese Medicine (TCM). In daily life, data mining technology has been imperceptibly involved in the improvement of people's quality life.

One of the most famous examples of data mining, the story of "diapers and beer," describes the several characteristics of data mining. In fact, data mining was originally applied in the field of business. The story of "diapers and beer" is just a well-known and interesting example. To analyze the product that customers most likely to buy together, Wal-Mart, the world's largest retailer, used the method of data mining to analyze a large amount of data in the database. They found that beer is the product with the highest probability that is bought with the diapers. Why they are bought together? The analysis showed that the wives often ask their husbands to buy diapers for their children after work, and the husbands have the most chance to buy diapers with beer. Thus, the market put them together. Then, sales of both diapers and beer have increased significantly.

The rise of data mining has the application background. As the world moves towards an information-based society, the abilities of humans to collect, process, organize, and produce information using information technology had been greatly improved. This results in the birth of tens of thousands of various types of databases, which plays an important role in the scientific research, technology development, production management, market expansion, commercial operation, and government office. However, with the increasing amount of information, especially the rapid expansion of network information resources, human beings are facing new challenges. For example, how does we not to be overwhelmed by massive amounts of information? how can we quickly obtain useful data from massive amounts of information? how can we fully improve the utilization of information? Then, data mining technology came into being. From the current development trend, the research and application of data mining technology show strong vitality.

Data mining appeared in the late 1980s, and it started with the research of Knowledge Discovery in Database (KDD). The term KDD first appeared at the International Conference

on Artificial Intelligence in 1989, and has gradually become a hot topic. Because of the expansion of this research object, it was named as the data mining. In 1995, the first International Conference on Knowledge Discovery and Data Mining was held, and this conference has been held once a year. The research on data mining in China started rather late, around the mid-1990s. In recent years, many universities and research institutes have conducted research in this area and achieved results. Data mining research results from not only the processing need of the large amount of information, but also the urgent need of various aspects of social development. For example, to improve the competitiveness of enterprises and develop good business operations, the organization of network information by information providers needs to study data mining technology.

After decades of development, the database system has saved a lot of daily business data. With the continuous deepening of the application of databases and various information systems, the amount of data is accumulating. A large amount of data is accumulated every year, which is in an incremental development trend. A large amount of information is a feature of today's information society and our precious wealth. However, facing huge amounts of data, we are confused, cannot find the relationships and rules among the data, and cannot predict future development trends based on the existing data. This results in the phenomenon of “People are drowning in a sea of information and lack of knowledge”. People began to think: “How can we not be overwhelmed by information, and find useful knowledge in time to improve the information utilization rate?” We hope to use data mining technology to mine knowledge from these data. A lot of information with decision-making significance is hidden behind a large amount of data. Through the analysis of massive data, potential connections among data are found, and automatic decision support is provided.

Data mining technology is the result of long-term research and development of database technology. Database technology was first used for online transaction processing to achieve unified storage of large amounts of data, and to provide transactional operations of query, insertion and deletion on data. With the accumulation of a large amount of historical data, people are no longer satisfied with simply querying and modifying data, but hoping to discover potential relationships among data. Therefore, new requirements have been put forward on database technology. With the maturity of some related disciplines and research fields, and the cruelty of commercial competition, enterprises are eager for quickly processing these data to obtain information that is good to the further development of enterprises. Whether the maximum utility of information resources had been achieved to manage and influence the decision-making process of the enterprise will decide whether the enterprise can have the maximum competitive advantage. Then, data mining technology has emerged and has been applied rapidly.

Data mining can be applied in various fields, and predict future trends and behaviors. Thus, it can support people's decisions well. For example, banks can use data mining to discover valuable customers, and insurance companies and securities companies can use it to detect fraud. The predictive information can be automatically discovered in a large amount of data by data mining. Therefore, problems that required a lot of manual analysis by domain experts and analysts in the past can now be rapidly made knowledge-based decisions directly from the data itself.

1.1.2 The concept of data mining

The definition of data mining is to mine information from a large amount of data, that is the non-trivial process to discover hidden, regular as well as finally understandable information and knowledge. Prior information means that the information is unexpected or novel one. Data mining information cannot be discovered by intuition, even information and knowledge that is contrary to intuition. The more unexpected information is, the more valuable it may be. The types of knowledge include models, laws, rules, patterns, constraints, *etc*. Potential usefulness refers to that the discovered knowledge has the usefulness in the future, which means the information or knowledge is effective, and achievable for the business or research area. Common-sense conclusions or facts that have already been grasped by people or unachievable speculations are meaningless. Ultimate understandability requires that the pattern can be understood by the user, which is expressed by the simplicity. The knowledge must be acceptable, understandable and applicable. It is better to express the findings in natural language. Non-trivial process means that the data mining is nonlinear. There are repetitions and cycles in the mining process. The mined knowledge is often difficult to be obtained through the simple analysis. This knowledge may be hidden inside the surface phenomenon and requires a lot of comparative analysis of data, using some tools that specialize in processing large amounts of data for mining.

The definition of data mining remains unclear, and depends on the perspective and background of the definer. Therefore, data mining does not have a unified definition. Definitions provided by Usama M. Fayyad *et al*. have been widely used. (1) Fayyad defines data mining as an important process for determining valid, new, potentially useful, and ultimately understandable patterns in data. (2) Zekulin defines data mining as a process of extracting unknown, understandable, and executable information from a large database and using it to make key business decisions. (3) Ferruzza defines data mining as the method used in the process of knowledge discovery to identify unknown relationships and patterns that exist in the data. (4) Jonn defines data mining as the process of discovering good patterns in data. (5) Parsaye defines the data mining as a decision support process of studying large data sets for unknown information patterns. These definitions are mainly based on the commercial application of data mining. From this perspective, the main characteristics of data mining are the extraction, transformation, analysis, and model processing of a large amount of transactional data in a commercial database. The key knowledge of business decisions is extracted, relevant business models from the database are discovered.

Data mining is a process that uses various analysis tools to discover the relationship between the model and data in massive data sets. These models and relationships can be used to make predictions. It helps decision makers to discover the potential association among data, and neglected factors. Thus, it is considered to be an effective method to solve the problem of data explosion and lack of information in the modern era. Data mining methods and mathematical tools include statistics, decision trees, neural networks, fuzzy logic, linear programming, etc.

1.1.3 Classification of data mining

The data mining can be classified from different perspectives. For example, data mining can be classified according to the type of database being mined, the type of knowledge being discovered, and the type of technology being used. In the following text, we only classify the data mining according to the type of database being mined and the type of knowledge being discovered.

(1) The data mining is classified according to the type of database being mined

(a) Relational database mining. The relational database mining is the data mining from a relational database to discover knowledge. The relational data model was proposed by E.F. Codd in 1970. Then, he made important contributions to the development of relational algebra, relational calculus and relational normalization theory, and laid the foundation for the theory and practice of relational database systems. Relational data theory refers to the theory of data correlation and normalization in relational models. Data correlation is a type of integrity constraint that describes the relationship among data in a relational model, and is usually reflected by the equality of data values. Data correlation includes the function correlation, multi-value correlation, interconnection correlation, connection correlation, etc. Among these, the function correlation and multi-value correlation are the most commonly used one.

(b) Object-oriented data mining. Each entity can be regarded as an object. For example, a table can be an object involved in reading a book, a customer can be an object in a shopping mall, etc. An object class is a collection of objects with certain common characteristics. Object-oriented database is a type of database designed based on object-oriented thinking. It introduces object-oriented concepts such as object identification, encapsulation, inheritance, and polymorphism in traditional databases to support data modeling in complex application fields.

Data mining in the object-oriented database can be used to discover object-based knowledge. In fact, the class hierarchy of object-oriented databases provides natural support for describing the background knowledge of knowledge discovery, and its inheritance and encapsulation mechanisms can support the modularity, reusability, and polymorphism of data mining.

(c) Data mining of transaction database. The data mining of transaction database can be considered to discover rules from the business transaction data. A general transaction database consists of a file, where each record represents a transaction with the unique identifier. Data mining of transaction database is often used for data analysis of “cargo basket”. For example, you can find products suitable for sale together in the shopping mall.

(d) Data mining of multimedia database. The multimedia database stores not only text and data, but also information such as images, audio and video. Different from the case of mining of other databases, the objects stored in the multimedia database occupy a large space, and the processing depends on special technologies such as image recognition and speech recognition technology. For example, speech data is a complex multimedia data, which contains a lot of information, such as frequency information, duration information, amplitude information, position information, and accent information. In simple terms, the same syllable will show different information characteristics in different sentences. The attribute values of contextual intent syllables will change. The speech data is a kind of time series data. The arrangement of

syllables in a sentence has a sequence. There is a strong "syllable" relationship between speech syllables. All these information characteristics have a great impact on the understandability and naturalness of the output of the entire synthesis system. In addition, because the data mining technology places high demands on the objects being processed. Therefore, the speech file must undergo a strict pre-processing process, such as syllable segmentation and syllable labeling of the recorded waveform. This requires a lot of human and material resources, and it is impossible without the accumulation of powerful speech processing capabilities. Applying data mining technology to speech signal processing can solve some of the difficult speech technology problems at the current stage. In addition, the impact of human experience on speech processing should be minimized to complete the transformation from qualitative to quantitative speech processing. Therefore, the application of data mining methods to speech synthesis has important significance and broad prospects.

Following example is a possible application. You only remember a melody of a song and don't know other information. However, you want to find this song in the karaoke library soon. Then, you need to perform data mining in a multimedia database, and can find songs stored as multimedia format data in the music library quickly and accurately.

(e) Data mining of spatial database. The spatial database contains information related to spatial data, which originates from the telemetry systems, remote sensing system, global positioning system, and geographic information system. Spatial data mining refers to the extraction of spatial pattern and feature, the general relationship between spatial and non-spatial data, as well as other common spatial data features implicit in the database from spatial databases that the user is interested in. In the mining process of spatial database, the purposes can be the characteristics of housing in a certain area, the urbanization situation, the correlation between the rich and the poor, etc. For example, the residents are divided into regions according to the case of ATM machines distributed in different geographical locations. According to this information, the setting plan of the ATM machine can be effectively performed. The waste and the loss of business opportunity are avoided. The research content of spatial data mining includes the following aspects. 1 Spatial distribution law. The geospatial distribution law of the target includes the horizontal distribution law, the vertical distribution law, the joint distribution law between the horizontal direction and the vertical direction. 2 Spatial association rules. Spatial association rules for adjacent, connected, symbiotic, and inclusive spatial objects. For example, villages are connected to roads, the intersections of roads and rivers are bridges, and large towns close to highways are usually adjacent to water. 3 Spatial clustering rules. Spatial objects with similar characteristics are clustered into the upper class. For example, scattered residential points that are close to each other are clustered into residential areas. 4 spatial classification rules. A category description rule for a space target. For example, waters and regions can be distinguished by the classification of remote sensing images, and then further are subdivided into rivers, lakes, reservoirs, grasslands, drylands, orchards. 5 Spatial feature rules. The general characteristics of the geometric and properties of spatial targets. 6 Space differentiation rules. The different characteristics of the geometric or attribute between two or more types of targets can be distinguished. 7 Spatial evolution rules. The spatial evolution rule refers to the rule that the spatial target varies with the time. For example, characteristics of some regions are inconstant, characteristic of some targets are inconstant and how they vary with time.

(f) Data mining on the Internet. Data mining on the Internet can sometimes be called Web mining. It can be used to filter news on the Internet, block spam emails, discover user browsing preferences, and speed up the network. The information on the Web can be considered as a database, the amount of data is large, and the data types are complicated. Every site on the Web is a data source. The information and organization of each data source are different, which constitutes a huge heterogeneous database environment. Using this data for data mining, we must first study the integration of heterogeneous data between sites. Only by integrating the data from these sites and providing users with a unified view, it is possible to access huge data resources. Second, it is also necessary to solve the data query problem on the Web. If the required data cannot be obtained effectively, it is impossible to analyze, integrate and process the data.

(g) Data mining of deductive database. The deductive database is the introduction of deductive rules in the traditional relational database, which makes it logically inferential. It is a combination of logical programming and relational database. The deductive database is composed of a database management system and a rule management system. The stored data includes fact data for inference and logical rules for deriving facts. The deductive database mainly studies the way to effectively calculate logical rule inference, including the optimization of recursive queries and the consistency maintenance of rules, *etc*. This mainly used as an environment for developing large-scale knowledge systems. There are two types of data mining in the deductive database. 1 Discover knowledge on the data defined by the deductive rules. 2 Further discover new knowledge in the mined knowledge and gradually refine it. This may be a more potential area of data mining in the deductive database.

(h) Data mining of time/time series database. This kind of database stores data related to time attributes, such as transaction changes in the stock market. Mining such databases can reveal trends in some of the objects and make predictions in decision making.

Time series pattern mining has a great relationship with time series databases. Its focus is on analyzing the contextual relationship between data. It can find knowledge in the database such as "a certain period of time, the customer purchases the product A. Then, the customer purchases the product B, and the product C. The sequence of $A \rightarrow B \rightarrow C$ appears frequently". The problem described by series pattern mining is that each series is a set of transactions arranged according to transaction time, and the series with higher frequency appearing are mined in a given transaction series database.

(i) Data warehouse. Because the data warehouse stores a large amount of historical data and comprehensive data, it provides a rich data resource for data mining. The data in the data warehouse is not up-to-date, proprietary, but derived from other databases. Data warehouse is based on database and is a new application of database technology. So far, most data warehouse systems use relational database management systems to manage their data.

Data mining technology has become an important and fairly independent tool in data warehouse applications. Data mining and data warehouse are both fused and interactive. The data in the data warehouse is collated and integrated, which simplifies the important steps in the data mining process (*i.e*., data integration and preprocessing). This improves the efficiency and capability of data mining, and ensures the data source in data mining. In addition, data mining goes one step further than multidimensional analysis in data warehouse. For example,

if managers ask to compare sales of a product in each region over the past year, they can obtain an answer through multidimensional analysis. If managers want to predict sales of the product in the coming year, they must use data mining tools. Furthermore, the particularity of the data warehouse also puts forward higher requirements for data mining, such as algorithm execution efficiency and dynamic maintenance of knowledge.

(2) Classification according to the type of knowledge

Discovering various rules is one of the results that data mining can provide. If the purpose is to discover association rules, it can be called association rule mining. Similarly, there are the characteristic rule mining, classification rule mining, time series rule mining, and deviation rule mining.

(a) Association Rules mining. Association rules are one of the first research objects in data mining. In 1993, Agrawal *et al.* first proposed the mining of association rules between "item sets" in the customer transaction database. After that, thousands of research papers in this area have been published. The mining of association rules has mature applications. The purpose of mining extends from discovering the association relationship between the original items to different types items.

(b) Characteristic rule mining. An expression of the data is extracted from a set of data related to the learning task to describe the overall characteristics of the data set. Feature rule mining has the wide application in business. For example, after customers are divided into different classes, the characteristics of each customer are further analyzed. The specific methods are as following: analyzing the characteristics of different types of customers through the analysis of the purchasing power of the customer, and developing the consumer with high purchasing power into the fixed customer; obtaining the characteristic rules of the customer's recent purchase through the time period analysis; analyzing the customer characteristic rules with different purchase frequencies through the mining of purchase frequency; mining the characteristic rule of customers spending the amount of money within a certain range based on the analysis of the amount of customers' purchase; mining characteristic rules of the most reasonable customer based on the purchase day, purchase frequency, the total purchase amount of money, etc.

(c) Classification rules mining. Classification is one of the major contents of data mining, which is analyzing experimental data samples to produce an accurate description of the categories. The result of the classification usually is a classification rule that can be used to predict future data. For example, the database of credit card company generally stores the record of cardholders. The company classifies the cardholders into three categories according to the credit grade: good, average, and poor. The classification rules mining is to find rules with different credit grades. For example, "customers with annual incomes above 50,000 yuan and between 40 and 50 years old have good reputation", and other customers can be classified according to the obtained rules.

(d) Time series rules mining. Timing rules refer to time-related rules, sometimes is called series patterns. For example, timing rules can be used to discover the time sequence and rules of purchasing insurance during the customer's life cycle in the insurance industry. This can assist insurance companies in discovering the best solution for re-sale of existing policyholders.

(e) Deviation rules mining. Deviations include anomalous instances in the classification,

exception patterns, deviations of observations from expected values, and their variation over time. The basic idea of deviation rules mining is to discover meaningful differences between observations and reference quantities.

1.1.4 Objects of data mining

In principle, data mining can be performed for any type of database, including text data sources, web data sources, and complex multimedia data sources that are not organized by the database. The following is a brief introduction to relational databases, data warehouses, text databases, multimedia databases, etc.

(1) Relational database

A relational database is a collection of tables. Each table has a unique identifier (table name). Each column of the table represents an attribute (also called a field), which is identified by a unique field name. Each row in the table is a tuple (also called record), all records are assigned the record number in sequence. Software that manages, accesses, maintains, and controls integrity and security for a database is called a Database Management System (DBMS).

Relational database is the most important, popular and rich information data source for data mining. It is one of the main forms of data mining research. The Structured Query Language (SQL) queries of relational database are transformed into a series of relational operations, such as selection, connection, projection, and so on. These operations can solve many problems, and generate new relational tables. Almost all resources can be expressed by relational tables (relational models). When data mining is used in relational databases, you can discover knowledge and potential information through techniques, such as association analysis, the links between products sold in supermarkets, and the purchase trends of customers at different age levels.

(2) Data warehouse

Data warehouse can be an advanced stage of data technology development. It is a topic-oriented, integrated, fairly stable, time-varying data set that can be used to support the management decision making process. The data warehouse system allows for the integration of various application systems and multiple databases to provide a solid platform for unified historical data analysis.

Data warehouse is derived from the needs of the decision support process. Therefore, it is first oriented to decision support. Its purpose is to establish a highly integrated data storage processing environment that will analyze the large amount of data required for decision making. Separated from the traditional operating environment, the scattered, inconsistent operational data is converted into integrated, uniform, and fairly fixed information. The most effective data mining tool for data warehouse is Multidimensional Data Analysis (MDA), also known as Online Analytical Processing (OLAP).

Data mining requires good data organization and "pure" data. The quality of data directly affects the effect of data mining. The characteristics of the data warehouse precisely meets the requirements of data mining. It captures data from various data sources. After cleaning, integration, selection, conversion and other processes, it provides high-quality data for data mining. The data mining provides an effective analytical processing method for data warehouse,

and data warehouse prepares a good data source for data mining. Therefore, data warehouse must become the best environment for data mining with the coordinated development of data warehouse and data mining. Compared with traditional databases, data warehouse has the following characteristics. 1 Subject-oriented. For example, policy data warehouse, customer data warehouse. 2 Integration. Data warehouse is not a process of simple accumulation of data. The multiple data sources are cleaned up, undergo irredundant procedure, and are integrated into the data warehouse. 3 The read-only characteristic of data. For the user, the data in the data warehouse is only for querying, retrieving, and extracting, and cannot be modified or deleted. 4 The historical nature of the data. Historic nature mainly refers to the accumulation of past data. 5 Variability over time. The data in the data warehouse is updated periodically over time. 6 Data warehouse has other characteristics. However, compared with database, these characteristics are not obvious. In sum, these characteristics of the data warehouse are suitable for data mining.

(3) Text database

The contents of the text database are texts. These words are not simple keywords, but have sentences, paragraphs and even text. The text databases are mostly unstructured and some are semi-structured (e.g., bibliographic data and full text; HTML, EMAIL mail, *etc.*). WEB pages are also text information. The database composed of many WEB pages is the largest text database. If the text data has a good structure, it can be implemented using a relational database.

What can data mining do in the text database? To answer this question, we first analyze it from the perspective of the user. Users obtain information from a large number of text sources. to obtain all the texts that reflect a topic or all the texts of one type information. However, because of the many discovered texts, the length of texts may be long. Thus, the long text is condensed into short texts (summary) that can reflect main content, and the information is further filtered by reading the short text. Therefore, the main content of data mining of text database includes the extraction of topic characteristic of text, text classification, text clustering, and text summary.

(4) Complex type of database

A complex type of database is not a simple text or a database that can represent dynamic sequence data. It mainly has the following types.

(a) Spatial database. This mainly refers to a database storing space information, where the data may be provided in raster format or represented by vector graphics data. For example: geographic information database, satellite image database, urban underground pipelines, and various underground building distribution databases of sewers. The mining of spatial databases can provide decision support for urban planning, ecological rules, and road construction.

(b) Time series database. It is mainly used to store time-related data, which reflects real-time data that varies with time, or events that occur at different times. For example, stock trading information, satellite orbit information, *etc.* are continuously stored. The mining of time series data can reveal the development trend of events, the evolution process of issues and hidden characteristics. This information will be useful for planning, making decisions and warnings about events.

(c) Multimedia database. It is a database for storing image, sound, and video information. Because of the development of multimedia technology and the achievements of related research

(such as visual information retrieval and virtual reality technology), multimedia databases have become popular and applied to many important research fields. At present, the mining of multimedia data is mainly on the retrieval and matching of image data. As the research progresses, it will expand to the mining of sound and video information, such as, the excerpt processing of video information.

1.1.5 Data mining and knowledge discovery

In the past, when I talked about the source of data mining, I have already mentioned the relationship between data mining and knowledge discovery. In many cases, two words can be mixed indiscriminately, but there are still some differences between them. The former is a stage of the latter.

(1) Definition of knowledge discovery

Knowledge discovery in a database is the process of identifying valid, novel, potentially useful, and ultimately understandable patterns that exist in a database. Knowledge discovery can be defined as follows using mathematical languages.

F is a data set, E is an expression describing F_E that is a subset of F. Knowledge value is an effective (C), novel (N), potentially useful (U), simple (S) complex, *i.e.*, $I(E,F,C,N,U,S)$. If there is a threshold i, it is called knowledge if $I(E,F,C,N,U,S) > i$.

(2) Steps of knowledge discovery

The whole process of KDD includes extracting models with data mining algorithms in a specified database, and a series of steps around preprocessing and result expression around data mining. Although data mining is at the heart of the process, it typically only accounts for 15% to 25% of the workload. KDD is a process of human-computer interaction centered on knowledge users.

The main steps are as follows. 1 Familiar with the application field, background knowledge and the nature of the user's KDD task; 2 Data acquisition: extract data from various types of data sources, and determine relevant data sets through operations such as selection, sampling, mapping, and aggregation; 3 Data cleaning and pre-processing: check the integrity of the data, remove errors and redundant data, organize inconsistent data, process lost data, update data and timing information, and prepare it as the expression form required by the data mining tool; 4 Data conversion: Data is unified into a form suitable for mining through data conversion methods such as aggregation, and reduction of dimensions, reducing the amount and complexity of data; 5 Determining the tasks of data mining, such as clustering, classification, regression analysis, etc.; 6 Selecting the data mining algorithm: select the appropriate model and parameters; 7 Performing the data mining process: discovering the potential patterns existing in the data and expressing them in an easy-to-understand form; 8 Evaluating and interpreting the discovered patterns, and repeating steps 1-7; 9 Using visualization methods and knowledge representation techniques to submit the discovered patterns to the user.

These nine steps can be simply attributed to three processes of knowledge discovery. 1 Data preparation: including three sub-steps of data integration, data selection, and data pre-processing. Data integration combines data from multiple files or database runtime

environments to resolve semantic ambiguity, process missing data, and clean dirty data. The purpose of data selection is to identify the data set that requires to be analyzed, reduce the scope of processing, and improve the quality of data mining. Preprocessing is to overcome the limitations of current data mining tools and transform data into the format required by data mining tools. 2 Data mining: first determine the mining task, then select the appropriate tools, mine knowledge, and finally confirm the discovered knowledge. 3 Expression and interpretation of the results: The extracted information is analyzed according to the decision-making purpose of the end user, the most valuable information is distinguished and submitted to the decision maker through the decision support tool.

A knowledge discovery system is a complex of a series of structures that discovers and extracts unknown and valuable patterns from data stored in real-world databases. It consists of a set of subsystems with different functions and interactions. The basic structure includes: 1 The controller processes requests for other components. 2 The interface generates data queries and returns the findings. 3 knowledge base stores domain information. 4 Concentrator determines the relevant data. 5 The mode extractor includes several knowledge extraction algorithms. 6 The evaluator evaluates the value of the discovered knowledge.

The input includes three pieces of information: 1 The user provides a high-level mining command to the controller; 2 DBMS provides source data; 3 Knowledge base provides domain knowledge. The source data selected by the DBMS is processed by the decimation algorithm to generate a candidate mode. The mode that is considered to be valuable by the evaluator is returned to the user. It is also stored in the knowledge base to support future data mining.

(3) The difference between knowledge discovery and data mining

The difference between knowledge discovery and data mining lies in the following two points: (1) Knowledge discovery is a cyclical process of using specific data mining algorithm, extracting valuable knowledge and patterns according to specified methods and thresholds, and performing evaluation and interpretation. The knowledge is constantly deepened and made easy to understand. The data mining is only a specific step of knowledge discover, generates patterns using specific data mining algorithms, not including data preprocessing, domain knowledge combination and steps, such as evaluation of results. 2 Data mining is mostly used by statisticians, data analysis experts and management information systems. The knowledge discovery is a promising field formed by the integration of many disciplines, such as artificial intelligence, machine learning, pattern recognition, statistics, databases, knowledge bases, and data visualization.

1.2 Common techniques for data mining

1.2.1 Features and processes of data mining

Data mining is an interdisciplinary subject that integrates knowledge in disciplines such as databases, artificial intelligence, and statistics. It has received much attention in recent years, and its definition is different with that of knowledge discovery (KDD). Some important features of data mining are in the example of “beer and diaper”, and are summarized below. 1 The size of the processed data is large, otherwise it is enough to simply use statistical methods to process

the data. 2 Query is generally random and proposed by decision makers (users). It often fails to form accurate query requirements. It relies on data mining technology to discover things that may be of interest and cannot be predicted. 3 Data mining must be responsible for the task of discovering potential rules, as well as managing and maintaining rules. In some applications, because the data changes rapidly, the rules can only reflect the characteristics of the current database. As new data is added continuously, the rules are constantly updated, and the original rules are required to be modified based on the new data to quickly respond. This can be referred to as "incremental" data mining. 4 In data mining, the discovery of rules is based on the statistical law of large samples. When the confidence reaches a certain threshold, the rule can be considered to be valid.

The basic process of data mining generally has the following steps: data collection, collation, mining, mining results evaluation, as well as analysis and decision making. This requires multiple iterative processes to achieve the desired effects. In particular, the steps are different in different application fields, and it has its uniqueness in the field of medicine. 1 Understanding the meaning of the problem to be solved in the field, determine the target and success criteria; 2 Understanding the familiar data; 3 Processing the data according to the refined theme, and building the data warehouse, which is also a dynamic cyclic process; 4 Data mining, which includes the choice of data model, training and verification process, modeling and quality evaluation of the model. Different algorithms can be used for the same process. This is a reasonable possibility to understand each algorithm for different aspects of data. In practice, it is necessary to repeatedly verify and compare; 5 Evaluating results and application, interpreting the extracted new knowledge, and requiring interpretation to have certain application value.

1.2.2 Common data mining methods

Different methods are used for analysis according to different purposes, and it is difficult to meet the requirements by using a single analysis method. Current data mining tools often use decision trees, neural networks, association rules, OLAP online analytical processing, genetic algorithms, Knearest neighbors, data visualization, and traditional statistical methods. These methods contain a large number of mining models, and commonly used in the medical field are correlation analysis, trend analysis, classification analysis, cluster analysis, sequence analysis, deviation detection and visualization technology.

(1) Decision tree

Decision tree is an effective method for classification in pattern recognition. It can help people to solve a complex multi-class classification problem into several simple classification problems. It is a tree structure in form, consisting of intermediate nodes, leaf nodes, and branches. The construction of the decision tree starts from the root node, selects the appropriate attribute to divide the sample data set into several subsets, establishes the branch of the tree, and repeats the process of constructing the lower nodes and branches in each branch subset until the condition is satisfied. A trained decision tree is used to predict the category of the new sample.

The decision tree is based on information theory and is a method to solve the classification

problem. It uses the information "gain" to discover the attribute of the largest amount of information in the database to establish a node, and recursively builds the branch of the tree from top to bottom according to the different values of the node attributes. Thus, scholars construct the tree model. It realizes the visualization of data rules, and has the advantages of high precision, high calculation speed, easy understand results and good benefits. 1 The decision tree method does not need to assume a priori probability distribution. This non-parametric feature makes it more flexible and robust. 2 Decision tree methods can not only use continuous real or discrete numerical samples, but also use “semantic data”, such as discrete semantic data: East, South, West, North, *etc*. The decision tree or production rule set generated by the decision tree method has the characteristics of simple and intuitive structure, easy understanding, and high computational efficiency. The decision tree method can effectively suppress the training sample noise and solve the problem of the attribute missing. Thus, it can solve the problem that the classification accuracy is reduced because of the noise from the training samples.

(2) Association rules

Association rule mining is the discovery of interesting associations or related relationships between sets of items in a large amount of data. It is an important topic in data mining and has been widely studied in the industry in recent years. The association rule is to describe the rules and patterns of certain attributes in things through association analysis. It establishes relevant rules according to the expected credibility, support and degree of action, and to speculate on unknown problems. A typical example of association rule mining is the analysis of shopping baskets. A well-known example is the Online Analytical Processing (OLAP) for beer and diaper relationships in market analysis. Association rule research helps to discover the connections between different commodities (items) in the transaction database. We discover the patterns of customer behavior, such as the impact of purchasing a certain product on the purchase of other commodities. The results of the analysis can be applied to merchandise shelf layout, inventory scheduling, *etc.*, as well as classifying users according to the purchase model. It collects databases from different levels and dimensions, and uses multi-dimensional methods to analyze, query, and report data to achieve the purpose of analyzing current and historical data.

The work of Agrawal *et al*. includes optimization of the original algorithm, such as introducing random sampling and parallel ideas to improve the efficiency of algorithm mining rules; and promoting the application of association rules. The domain target attribute value is converted into a discrete value by domain division, and the association rule method can be used to predict the attribute value of the numerical target. The association rule describes the nature of the target attribute by testing the condition attribute, and the association rule can be used to predict the value of the target attribute. Setting Y as the target attribute, X as the condition attribute test, the association rule $X \Rightarrow Y \Rightarrow a$ means that if the tuple t satisfies the test, the probability that Y (the value of attribute of t) is equal to a is high.

The types of association rules are: 1 Based on the categories of variables processed in the rules, the association rules can be divided into Boolean and numeric type. The values processed by Boolean association rules are discrete and categorical, showing the relationship between these variables. Numeric association rules can be combined with multi-dimensional associations or multi-layer association rules to process numeric fields, dynamically perform

segmentation on them, or directly process the original data. Numeric association rules can also contain category variables. 2 Based on the abstraction level of data in the rules, it can be divided into single-level association rules and multi-level association rules. In the multi-level association rules, the "multilayering" of data has been fully considered. 3 Based on the "dimension" of the data involved in the rule, the association rules can be divided into single-dimension and multi-dimension. In the one-dimensional association rule, we only cover one dimension of the data. For example, if the user purchases an item in a multi-dimensional association rule, the data to be processed will involve multiple dimensions.

(3) Neural network

The Neural Network (NN) is an information processing system constructed to simulate the structure and function of the biological brain. It is a non-linear prediction model based on self-learning mathematical model and learned through training. It is commonly used as a classifier in machine learning. Its classification accuracy is much higher than that of decision trees. It can analyze a large number of complex data and complete complex pattern extraction and trend analysis for human brain or other computers. However, this takes a longer time and the extracted rules are less visually inferior to decision tree.

The neural network is formed by a large number of simple units interconnected in a certain way, and the information is processed by the dynamic response of its state to external input information. Neural networks have the capabilities of self-learning, self-organization, self-adaptation, association, fuzzy inference, massively parallel computing, nonlinear processing, robustness, distributed storage, and association. This can help people effectively solve many nonlinear problems. Neural network technology is an important method in the field of soft computing. It is the result of researchers' function simulation of human brain neuron learning for many years and has been successfully applied in many industries and fields. The research and application of neural networks have penetrated into machine learning, expert systems, intelligent control, pattern recognition, computer vision, information processing, intelligent computing, associative memory, coding theory, medical diagnosis, financial decision making, nonlinear system identification. The combinatorial optimization, real-time language translation, enterprise management, market analysis, decision optimization, material handling, adaptive control, neurophysiology, psychology, and cognitive science research. Data mining is one of the application fields of artificial neural networks. Artificial neural networks are based on self-learning mathematical models. Through the coding of data and the iterative solutions of neurons, complex pattern extraction and trend analysis functions are completed.

The neural network system consists of a series of processing units (called Nodes) similar to human brain neurons. The nodes are interconnected and divided into an input layer, an intermediate (hidden) layer, and an output layer. The main neural network models are: 1 Feedforward network. For example, mathematical models of formal neurons—old neuron models, perceptron, backpropagation models, radial basis function networks, Madaline networks, and multilayer feedforward networks—for prediction and pattern recognition; 2 Feedback networks, such as Hopfield discrete models and continuous models, *etc.*, for associative memory and optimization calculations; 3 Self-organizing networks, such as ART model and Kohonen model, *etc.*, used in the fields of intelligent control, pattern recognition, signal processing and optimization calculation.

The neural network system has the advantages of nonlinear learning and associative memory. However, there are also some problems. First, the neural network system is a black box, and the intermediate learning process cannot be observed. The final output result is also difficult to explain, affecting the credibility of the result and acceptable level. Second, neural networks require longer learning times, and performance can be severely problematic in the case of large amounts of data. The use of neural network techniques is effective when it is difficult to derive concepts and determine trends from complex or imprecise data. A trained neural network is like an "expert" with some kind of expertise, and can learn from experience like a human.

(4) Genetic algorithm

Genetic algorithm is a kind of randomized search method, which is derived from the evolutionary law of the biological world (the survival of the fittest). It was first proposed by Professor J. Holland of the United States in 1975. Its main features are as follows. 1 Direct manipulation of structural objects. There is no limitation of derivation and function continuity; 2 Inherent implicit parallelism and better global optimization capabilities; 3 Adopting probabilistic optimization method, which can automatically acquire and guide the optimized search space, adaptively adjust the search direction, and does not require to determine the rules. Due to these properties, the genetic algorithm has been widely used in the fields of combinatorial optimization, machine learning, signal processing, adaptive control and artificial life. It is a key technology in modern intelligent computing.

Genetic algorithm is a kind of self-organizing and adaptive artificial intelligence technology that simulates the evolution process and mechanism of natural biological processes. It is a search optimization algorithm based on biological evolution theory and molecular genetics. It encodes the possible solutions of the problem and randomly selects them. Several chromosomes (coded solutions) are used as the initial population, and the fitness value of each chromosome is calculated according to the pre-evaluation function. Then, the chromosomes with greater adaptation values are selected to generate new populations that are more adaptive to the environment. Finally, they are convergence to the most adaptable individual, and the process of optimizing the solution is obtained.

Genetic algorithm is a kind of search algorithm, and can be used for complex system optimization and robustness. Compared with the traditional optimization algorithm, it has the following advantages. 1 Genetic algorithm uses the coding of decision variables as the operation object. Traditional optimization algorithms often directly determine the actual value of variables, and the genetic algorithm deals with certain coding forms of decision variables. Thus, we can learn from the concepts of chromosomes and genes in biology, and imitate the genetic and evolutionary mechanisms of natural organisms. It is convenient to apply genetic manipulation operators. 2 Genetic algorithm directly uses the fitness as search information without other auxiliary information. 3 Genetic algorithm uses multiple points of search information with implicit parallelism. 4 Genetic algorithm uses probabilistic search techniques rather than deterministic rules.

(5) K-nearest neighbor method

The K-nearest neighbor method treats data that is close to each other as neighbors. Based on the principle of do as your neighbors, the average data of K neighbors is used to predict a

certain attribute or behavior of a specific data.

(6) Traditional statistical method

The traditional statistical method is to analyze the functions and correlations between data attributes, using regression, correlation, and principal components.

In addition, there are rough set theory, visualization technology, and Bayesian network. In actual data processing, it should be appropriately selected according to the specific conditions and characteristics of methods.

1.2.3 Data mining and decision support system

(1) Data mining system

The structure of the data mining system is very complex. It is an application software system with high technical content, which integrates information management, information retrieval, expert system, analysis and evaluation, and data warehouse. It is usually composed of a database management module, a pre-excavation processing module, a mining operation processing module, a mode evaluation module, and a knowledge output module. According to the organic composition of these modules, the architecture of the data mining system is formed (Figure 1-1).

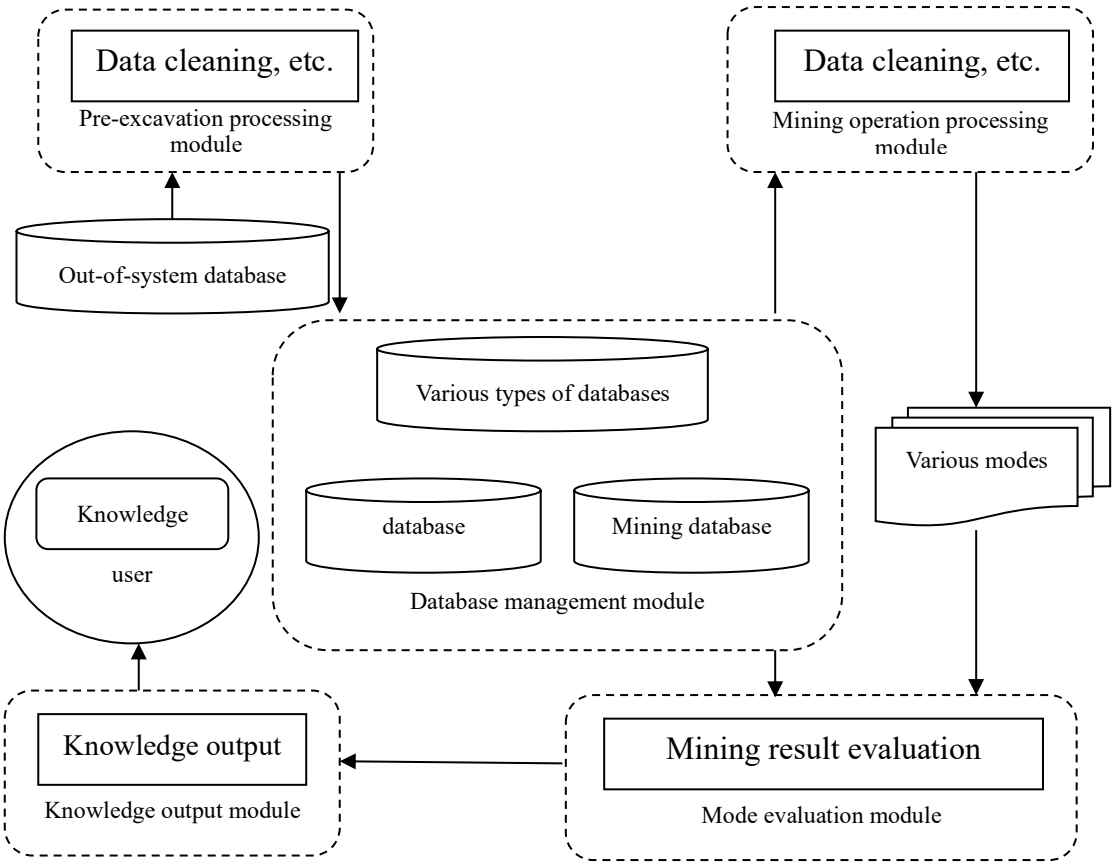


Figure 1-1. Data mining system architecture.

(a) Database management module. It is responsible for the maintenance and management of the database, data warehouse, and mining knowledge base in the system. These are the

foundations of data mining by transforming, cleaning, and purifying external databases to obtain the database and data warehouses required by the system. The mining knowledge base records those experiences, rules, techniques, methods, theories, factual data, and the knowledge used in the mining process. Its role is to guide mining and model evaluation.

(b) Pre-excavation processing module. The collected data is cleaned, integrated, selected, converted to generate a data warehouse and a data mining library. The cleaning process is mainly to clear the noise. The integration process is to combine multiple data sources. The selection is to select the data related to the problem under study. The conversion is to convert the selected data into a form that can be mined. If the prepared data problem affects the mining mode, the mode evaluation will discover the problem and return to the data mining pre-processing process or program.

(c) Mining operation processing module. Using various data mining algorithms, knowledge is mined and discovered for different databases, data warehouses, and data mining libraries by mining rules, methods, experiences, and facts in the knowledge base. The mining operation processing module is the core of the data mining system. It involves many algorithms and techniques, such as decision tree method, association analysis method, neural network method, online analysis processing, text mining technology, and multimedia data mining technology.

(d) Mode evaluation module. Data mining results is evaluated. There may be many models to be mined. Thus, it is necessary to analyze and compare the user's interest with these models, evaluate the value of the model, and analyze the cause of the deficiency. If the model is mined differs greatly from the user's interest, the operation will be returned to the corresponding process and re-executed, such as the mining operation or the pre-excavation processing module.

(e) Knowledge output module. Tasks such as translating and interpreting data mining patterns are completed, the best result is provided for decision makers who are truly eager to get the knowledge they require. The knowledge output module is a bridge between the user and the data mining system. Users can directly interact with the data mining system to develop data mining tasks, provide information, help mining focus, and conduct exploratory data mining based on the results of each step on data mining knowledge.

(2) Decision support system

The Decision Support System (DSS) proposed in the 1970s can provide various decision-making information and solutions for many business problems according to users' needs. Thus, it reduces the burden on managers to engage in low-level information processing and analysis. You can focus on the work that requires the wisdom and experience of decision making, improving the quality and efficiency of decision making. DSS is based on the Management Information System (MIS). MIS is to use database technology to realize the management business of managers at all levels, and to carry out various transaction processing on the computer. DSS is to provide the ability for assistant managers to make decisions. In 1980, Sprague proposed the DSS three-part structure, namely the dialogue component, the data component (database and database management system) and the model component (Model Library (MB) and Model Library Management System (MBM), etc.), which greatly promotes the development of DSS. In 1981, Bonczak *et al.* proposed the DSS three-system structure, including the Language System (LS), Problem Processing System (PPS) and Knowledge

System (KS). The DSS is mainly based on the model library system, and assists decision making through quantitative analysis. The models in the model library have been extended from mathematical models to data processing models and graphical models. It can be summarized as generalized models. The essence of the DSS is to organically combine multiple generalized models and process the data in the database to form a large model of decision problems. The decision-making ability of decision-making support system has evolved from the single-model decision-making in operations research and management science to multi-model comprehensive decision making. This has brought the decision-making ability to a new level. Executive Information Systems (EIS, also known as Manager Information Systems), similar to DSS, are also analytical systems. These are systems built on data warehouses that provide decision support for users. EIS is primarily used to provide an effective mean of accessing, creating, and disseminating information to senior executives who lack experience in computer use. The information here refers to the information of the decision class. EIS typically provides a high-level, comprehensive view of the data, since senior executives need to "slice" and "cut" the same type of data, and are less to see the details.

In the late 1980s and early 1990s, the DSS was combined with the Expert System (ES) to form an intelligent decision support system (Intelligent DSS). The expert system is a qualitative analysis and assistant decision-making. It is combined with the DSS with quantitative analysis to assist decision-making, which further improves the decision-making ability. Intelligent DSS is a new stage in the development of DSS. The Group Decision Support System (GDSS) is beneficial to overcome the subjective judgment errors in some individual decisions. However, the decision-making process is relatively long. To achieve high-efficiency group decision making, many people have done a lot of work in theoretical methods and application software development. They have obtained some results, such as the multi-person multi-objective decision theory, master-slave decision theory, negotiation system and conflict analysis.

The comprehensive DSS formed by combining data warehouse, query reporting tool, OLAP, data mining and model library is a more advanced form of the DSS, which can effectively improve the system's decision support ability. This is manifested in the following aspects. (1) The data warehouse includes basic data, historical data, comprehensive data, and metadata describing the characteristics of the data, and realizes storage and synthesis of the data of the decision-making theme. (2) Query report tools achieve daily transaction operations such as data query and management. OLAP analyzes the data to form professional reports. (3) Data mining is used to mine knowledge in databases and data warehouses for predictive analysis and decision-making. (4) Model library implements combined auxiliary decision making of multiple generalized models. (5) Expert systems use knowledge reasoning for qualitative analysis. The integrated decision support systems they integrate can complement each other and rely on each other to leverage their respective decision-making advantages to achieve more effective decision-making. It can be said that data mining technology provides a new method for the development of decision support systems.

1.3 Data mining tools

1.3.1 Classification and selection of data mining tools

With more and more software vendors joining the ranks of data mining, the performance of existing mining tools has been further enhanced. It is more convenient. The price threshold has been rapidly reduced, which has brought about the popularity of applications. At present, more than 50 related software is available, and their selection should be based on the analyst's analytical ability, purpose, data type, analytical methods and ease of use provided by the software.

(1) Classification of data mining tools

In general, data mining tools are classified into two categories based on their scope of application, including the specialized data mining tools and general data mining tools. Specialized data mining tools provide solutions to problems in a specific domain. When considering algorithms, they take into account the specificity of data and requirements, and optimize them. General data mining tools do not distinguish the meaning of specific data, and mining algorithms are used to handle common data types.

(2) Selection of data mining tools

Data mining is a process. Only when the technology and implementation experience provided by data mining tools are closely integrated with the business logic and requirements of the enterprise, then the success can be achieved. Thus, the selection of data mining tools is based on many factors, mainly including the following aspects.

(a) The function and method of data mining. That is, whether it can complete various data mining tasks, such as association analysis, classification analysis, sequence analysis, regression analysis, cluster analysis and automatic prediction. The process of data mining includes data sampling, data description and preprocessing, data transformation, model building, model evaluation and release. Thus, a good data mining tool should provide a corresponding feature set for each step. The data mining tool should easily export the mined model for the use in future applications.

(b) The scalability of data mining tool. To solve complex problems, a good data mining tool should be able to handle as much data as possible. This can improve the processing efficiency, and make the result of the processing more effective. If the time of mining increases linearly with the increase in data volume and mining dimension. Then, the mining tool can be considered having better scalability.

(c) The ease of operation. A good data mining tool should provide users with a friendly visual operation interface and graphical reporting tool. This should maximize the automation level in the process of data mining. In sum, it is for the majority of users rather than skilled professionals.

(d) The visualization of data mining tool. This includes visualization of source data, visualization of mining models, visualization of mining processes, and visualization of mining results. The degree of visualization, quality, and flexibility of interaction seriously affect the ability to use and interpret data mining systems. After all, 80% of the information received by the outside world is obtained visually, and the visualization capabilities of natural data mining

tools are very important.

(e) The openness of data mining tool. It is the ability to combine data mining tools with databases. Good data mining tools should be able to connect as many database management systems and other data resources as possible, as well as integrate with many other tools. Although data mining does not necessarily have to be done on top of a database or data warehouse, mining data collection, data cleaning, data transformation, *etc*. consume huge time and resources. Therefore, data mining tools must be closely integrated with the database to reduce the time of data conversion and make full use of the processing power of the entire data and data warehouse. Data mining is performed directly in the data warehouse. The development model, test model, and deployment model must fully utilize the processing power of the data warehouse. In addition, multiple data mining projects can be performed simultaneously.

The mentioned above is just some common reference indicators. When selecting the mining tool, users need to analyze the actual situation.

1.3.2 Tools for current data mining

Famous tools are IBM Intelligent Miner, SAS Enterprise Miner, SPSS Clementine, *etc.* They are able to provide regular mining processes and mining patterns.

(1) Intelligent Miner

Intelligent Miner, a data mining software developed by IBM in the United States, is a software series for data mining of database and text information, including Intelligent Miner for Data and Intelligent Miner for Text. In the Intelligent Miner portfolio, there are many different features, two of which commonly used are the "cluster" mining feature and the "association" mining feature. These two functions can be used for customer risk analysis and marketing activities for banking products. Intelligent Miner for Data can mine hidden information contained in databases, data warehouses, and data centers to help users exploit data from traditional databases or structured data in common documents. It has been successfully applied to market analysis, fraud monitoring and customer contact management. Intelligent Miner for Text allows companies to mine data from textual information. Text data sources can be text documents, web pages, emails, Lotus Notes databases, and more.

(2) Enterprise Miner

This is a data mining tool that has been adopted in enterprises in China. The typical applications include the Shanghai Baosteel's ore distribution system and the railway department in the study of passenger transport in the Spring Festival. SAS Enterprise Miner is a general-purpose data mining tool that performs data mining in accordance with the method of "sampling-exploration-conversion-modeling-evaluation". It has the ability to integrate with SAS data warehouse and OLAP to achieve "end-to-end" knowledge discovery from presenting data, capturing data, to obtaining answers.

(3) SPSS Clementine

SPSS Clementine is an open data mining tool that has twice received the UK Government's SMART Innovation Award. It supports the entire data mining process from data acquisition, transformation, modeling, evaluation to final deployment, as well as the standard of data mining-CRISP-DM. Clementine's visual data mining makes "thinking" analysis possible,

focusing on the problem itself, rather than being limited to doing some technical work (such as writing code). It provides a variety of imaging techniques to help understand the critical connections between data and guide users in the most convenient way to find the final solution to the problem.

Other commonly used data mining tools include LEVEL5 Quest, MineSet (SGI), Partek, SE-Learn, SPSS data mining software Snob, Ashraf Azmy's SuperQuery, WINROSA, XmdvTool and so on.

After more than ten years of development, the performance of data mining tools has been significantly improved. Both the degree of automation and the scope of application have changed dramatically, the price threshold has been rapidly reduced. The application of data mining in enterprises and e-commerce has been promoted. However, there are still many shortcomings in the current data mining tools. The 1999 survey shows that most data mining tools use only a limited number of technologies and focus on the relatively simple types of data mining technologies.

1.3.3 Characteristics of medical data mining

Although the data mining technology has only been produced for more than ten years, it has been widely used in commercial, industrial engineering, telecommunications and other fields. It has achieved considerable economic and social benefits. However, it is still in its infancy in the medical field. The medical and health system has its dual characteristics of complexity and time-varying. In addition, medical technology is practical, experimental and statistical. It is a confirmatory science. This leads to the uniqueness of medical data mining, and the data mining in this field has strong practical value and broad prospects. Medical information contains all the data resources of medical procedures and doctors' patient activities, including clinical medical information and hospital management information, especially the former reflects the uniqueness of medicine. This information has model polymorphisms (pure data, images, signals, transcripts, etc.), incompleteness (objective incompleteness of disease information and subjective incompleteness of disease description), strong temporality, complexity and redundancy. In addition, its mathematical characteristics, non-standardized form, and the asymmetry of medical information of doctors and medical materials involve more ethical and legal issues. Thus, this determines the uniqueness of medical data mining. Because of the wide range of sources and the large capacity of medical data, medical data often contains fuzzy, noisy and redundant information. This requires the use of unique technologies in mining. For example, pre-processing data, cleaning and filtering data are used to ensure the certainty of data; information fusion technology is used to make different patterns of data convergent or consistent in terms of attributes. Considering the efficiency of mining, the mining algorithm of medical data should have strong fault tolerance and robustness. In addition, how to ensure the accuracy, reliability, and scientificity of the mining results and reduce the risk of mining are the key to provide scientific decisions for medical activities and management, and to obtain practical applications.

1.4 Data preprocessing

The data in the data source may be incomplete (such as the uncertainty or vacancy of certain attributes), noisy and inconsistent (such as an attribute has a different name in a different table). The quality of data mining for these incomplete, noisy, and inconsistent data is difficult to guarantee. Moreover, the data required for data mining may involve multiple data sources. The amount of data for each data source is large, and include redundancy. The data is scattered. It is often stored in a heterogeneous environment lacking unified design and management, and is difficult to comprehensively query. A large amount of historical data is offline, and queries cannot be stored centrally online. These factors can affect the efficiency of data mining. Therefore, pre-processing methods such as cleanup, integration, transformation, reduction, *etc*. can be used to improve data quality before data mining, thereby improving the efficiency and quality of data mining.

1.4.1 Data cleanup

(1) Elimination of incompleteness

(a) Fill with a global constant. For example, the age attribute in a customer table is populated with "unknown."

(b) Fill with the attribute average. For example, the salary attribute in a customer table is populated with the salary attribute average.

(c) Fill with the attribute average of the same class. For example, in classification rule mining, you can use the attribute averages of other samples that belong to the same class as a given sample.

(d) Fill with the most likely values. For example, the salary attribute in a customer table is a predictive attribute, and a prediction algorithm is used to predict and fill the most likely value of the salary attribute of a given sample.

(2) Elimination of noise

(a) Eliminate noise by smoothing the data. For example, the binning technique sorts the data, distributes the data to different bins according to the distribution rules, and replaces the data in the same bin with the corresponding data according to the smoothing rule. The distribution rules can be equal depth and equal width. The same depth means that the number of data in each box is equal. The equal width means that the value interval of each box is equal. Smoothing rules can be average smoothing, median smoothing, and boundary smoothing. The average smoothing means that all the data in the same box is replaced by the average value of the data in the box; the median smoothing means that all the data in the same box is replaced by the median value of the data in the box; boundary smoothing means that the same data in the box is replaced with the nearest boundary value in the box.

Example 1-1 The values of an attribute are 18, 12, 3, 9, 7, 6, 15, 21. 16. The binning technique is used to smooth the data to eliminate noise. The distribution rule is equal depth with the depth of 3, and the smoothing rule is the average smoothing.

First, sort the values of the attributes to 3, 6, 7, 9, 12, 15, 16, 18, 21

Then, the distribution rules (equal depth with depth of 3) distribute the data to

Box 1: 3, 6, 7

Box 2: 9, 12, 15

Box 3: 16, 18, 21

Finally, according to the smoothing rule

Box 1: 5.3, 5.3, 5.3

Box 2: 12, 12, 12

Box 3: 18.3, 18.3, 18.3

(b) Eliminate noise by identifying isolated points. For example, using a clustering algorithm to obtain classes (or clusters), data outside the class can be treated as isolated points (or noise) and eliminated.

(3) Eliminate inconsistency

For example, the inconsistency of the unary data is eliminated by describing the data.

1.4.2 Data integration

Data integration is the combination of data from multiple data sources in a consistent data store (such as a data warehouse). These data sources may include multiple databases, data cubes, or general files. There are many issues should be considered when the data is integrated. For example, how to match the real-world entities from multiple sources of information.

Redundancy is an important issue. An attribute is redundant if it can be "exported" by another attribute, such as annual salary. In addition, inconsistent naming of attributes or dimensions can result in redundancy in the data set.

Some redundancy can be detected by correlation analysis. For example, for two numerical properties A and B, the correlation between them can be calculated according to the following formula

$$r_{AB}=\frac{\sum(A-\overline{A})(B-\overline{B})}{(n-1)\sigma_A\sigma_B} \tag{1-1}$$

where n is the number of tuples, $\overline{A}$ and $\overline{B}$ are the average of the value A, B ($\left[\overline{A}=\frac{\sum A}{n}\right]$) respectively, σ_A and σ_B are the standard deviation of A and B ($\sigma_A=\sqrt{\frac{\sum(A-\overline{A})^2}{n-1}}$), respectively.

If $r_{AB}>0$, A is related to B, A is increasing with the increase of B; If $r_{AB}<0$, A is inverse proportional to B, A is increasing with the decrease of B; If $r_{AB}=0$, A and B are independent. Therefore，when $|r_{AB}|$ is great, we can eliminate A or B.

In addition to detecting redundancy between attributes, "repetition" should also be detected at the tuple level. Repetition means that there are two or more identical tuples for the same data.

Another important issue in data integration is the detection and processing of data value conflicts. For example, attribute values from different data sources may be different for the

same entity in the real world. This may be because the representation, scale or coding is different. This semantic inconsistency of data is a huge challenge for data integration.

Careful integration of data from multiple data sources can reduce or avoid redundancy and inconsistency in the data in the resulting data set. This helps to improve the accuracy and speed of subsequent excavation.

1.4.3 Data transformation

The data transformation transforms the data into a form suitable for mining. The most common method of data transformation is to normalize data. The attribute data is scaled so that it falls into a small specific interval, such as -1.0 to 1.0 or 0.0 to 1.0.

(1) Minimum-maximum normalization

For a given numerical property, [minA, maxA] is the value range before A normalization, [new_minA, new_maxA] is the value interval after A normalization. According to Equation (1-2). The value v of A is normalized to a value v'

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A \tag{1-2}$$

Example 1-2 Assume that the value range [-100,100] before normalization is normalized, and the normalized value range [0,1] is normalized by the minimum-maximum rule.

$$v' = \frac{88-(-100)}{100-(-100)}(1-0)+0 = 0.94$$

(2) Zero-average normalization

A, $\overline{A}$, σ_A are the average value, standard deviation of A, zero-mean normalization, respectively. According to Equation (1-3), v value of A is normalized to v'

$$v' = \frac{v - \overline{A}}{\sigma_A} \tag{1-3}$$

Example 1-3 Assume that the mean and standard deviation of an attribute are 90 and 30, respectively, and zero-mean normalization is used.

$$v' = \frac{88-90}{30} = -0.67$$

(3) Standardization of decimal calibration

For A, $\max_A$ is the greatest absolute value of A, J is the smallest integer to satisfy the Equation (1-4)

$$\frac{\max_A}{10^j} \le 1 \tag{1-4}$$

According to the decimal scale normalization, v value of A is normalized to v'

$$v' = \frac{v}{10^j}$$

Example 1-4 Assumes that the value range before normalization of an attribute is $[-120,110]$, and the decimal scale is normalized to 88. The greatest absolute value of A is 120, A is 3. The value of 88 is normalized

$$v' = \frac{88}{10^3} = 0.088$$

1.4.4 Data reduction

Data reduction techniques can be used to derive a reduced representation of a data set that is much less redundant than the original data. However, it is still close to maintaining the integrity of the original data. In this way, mining on the reduced data set will be more efficient and produce the same (or almost identical) analysis results. The reduction methods mainly include:

(1) Attribute reduction (dimensional reduction)

For example, a decision tree is obtained according to a classification algorithm of information gain-based decision tree. Attributes outside the decision tree can be regarded as ones that are not related to the data mining task and should be deleted.

(2) Record reduction

For example, data generalization usually adopts attribute-oriented induction, hierarchy according to the concept of attributes. This replaces the lower-level attribute values with the corresponding high-level concepts, and merges the same records. Thus, it achieves the record reduction concept layering. A set of mapping sequences is defined that maps more specific low-level concepts to more general high-level concepts. Concept hierarchy is divided into four types of pattern-defined stratification, collection grouping stratification, operation-derived stratification, and rule-based stratification. The collection grouping stratification is one of the most commonly used type. This can generally be described by a tree structure called a concept hierarchy tree. The same attribute can define multiple concept layers based on different aspects.

The threshold control is oriented to the attribute induction process and can be divided into two types of attribute threshold and relationship threshold. The attribute threshold is the most commonly used types. Generally, you can set the same attribute of value for all attributes, or you can set different attribute thresholds for different attributes.

According to the relational table, the concept hierarchy tree of each attribute and the attribute threshold, the attribute-oriented induction treats each attribute as follows. First, according to the concept hierarchy tree of attribute A, the attribute value of A in the relational table is converted into the corresponding concept of the lowest level, which has known as the leaf concept. The number of different leaf concepts of A in the statistical relationship table should be counted. If the number of different leaf concepts of A is greater than the attribute width of A, then according to the concept hierarchy tree of A, the leaf concept of A is in the relational table. We convert it to the corresponding concept of the previous layer, and repeat the process until the number of different concepts of A in the relational table is less than or equal to the attribute threshold of A. Finally, we merge the same records and count the number of duplicate records.

1.4.5 TCM information quantization processing

In recent years, Chinese scholars have focused on the research of Chinese medicine information database systems. The information of Chinese medicines is divided into two

categories. One is the increasing information, such as new natural product components, pharmacological activity data, *etc.* The other one is information that is basically stable but needs to be standardized, such as the traditional efficacy of Chinese medicine, processing methods, and so on. Most of the research on Chinese medicine information database is aimed at the short-form data of the phrase type with exclusive effect and relatively independent connotation, such as efficacy, jinging, usage and dosage, toxicity, chemical composition and other information on the establishment of TCM information. The computer performs direct quantization processing. The other type is TCM data or pictures of complex large texts composed of multiple Chinese characters, such as pharmacological effects, clinical diagnosis, various discourses, and pictures of TCM. This kind of data has an important relationship with the deep information of Chinese medicine. However, most of the content is not suitable for computerized quantitative processing without special treatment. The traditional method cannot effectively solve the extraction of Chinese medicine information that describes vagueness, and improve Chinese medicine. The accuracy of quantitative data extraction needs to address the accurate description of the fuzzy concept. The author took the lead in introducing the concept of "quantization" of TCM information in the design of TCM database for the first time. This concept refers to the refinement of the original complex Chinese medicine data through reasonable analysis, which has become composed of several Chinese characters or numbers. The smallest unit of information that is relatively independent and cannot be subdivided and exclusive. This minimum unit of information is called the "quantization" of TCM information. After quantification of Chinese medicine information, a large number of ambiguous, noisy and redundant data in Chinese medicine data is cleaned and filtered to ensure the certainty, consistency and uniqueness of Chinese medicine data to meet data mining requirement. For example, research on TCM processing is an important part of the theory of TCM. Different medicines have different processing methods. The separation and removal of non-medicinal parts of TCM are quantified, which are refined into the removal of root, stem, branch, stalk, peel, shell, hair, heart, core, reed, flesh, head, skin, tail, bone, feet, wing, meat, impurities, mildew. The information about the Chinese medicine picture is quantized by the first letter plus the number of the pinyin. For example, the picture of the yam can be coded as SY001. By "quantizing" the original data of TCM, the TCM data with complex original content can be subdivided into much "quantum" information with independent concepts. The quantum information of TCM can be established based on the "quantum" information of these TCM. The database lays the foundation for deep mining of Chinese medicine data.

1.5 Data mining in the field of TCM

Data mining technology has been application-oriented since its inception. It is not only a simple query for a specific database, but also micro, meso and even macro statistics, analysis, synthesis and reasoning to guide the actual problem solving. It attempts to find the interrelationship among events, and even uses existing data to predict future activities. Data mining can be performed on any type of data, and can be commercial data. It can be data generated by social sciences, natural science processing, or satellite observations. The data form and structure are also different. It can be a hierarchical, networked, relational database. It can

be an object-oriented and object-relational advanced database system. It can be a database for special applications, such as a spatial database or a time series database, text database and multimedia database. In the field of pharmacy, it is also widely used in database technology. Literature search, online inquiry, retrieval of information on the Internet, office information system, hospital pharmacy information management system, all of these application technologies are backed by database technology. Nowadays, there are very few data that are not stored in the form of a database. The application of database technology has brought great convenience to the storage, management and query of a large amount of data.

The following is mainly to explain the application and prospect of data mining in the field of TCM from the aspects of bioinformation knowledge discovery, data mining in scientific drug sales and statistical data mining, chemical data mining, and web mining.

1.5.1 Bioinformation knowledge discovery

Bioinformatics is a new interdisciplinary subject that emerged with the launch of the Human Genome Project in the late 1980s, often refers to as genomic informatics. Bioinformatics is a discipline that uses mathematical and information science perspectives, theories, and methods to study life phenomena, organization to analyze biological data that exhibit exponential growth. This is the first study to the carrier DNA of genetic material and its encoded macromolecular protein. It uses computer as its main tool to develop various software to collect, organize, store and release the sequence and structure of the rapidly growing DNA and protein. Extraction, processing, analysis and research are aimed to gradually understand the origin, evolution, inheritance and developmental nature of life. This is achieved through the analysis and decipher of the genetic language hidden in DNA sequences, which reveals the molecular basis of human physiological and pathological processes, and the most reasonable and effective method for the prediction, diagnosis, prevention and treatment of diseases. Bioinformatics has become a powerful driving force in the development of biomedicine, agronomy, genetics, cell biology and other disciplines, as well as an important part of drug design and environmental monitoring. It is one of the major frontiers of life science and natural science, and will be one of the core areas of natural science in the 21st century. The research of bioinformatics focuses on two aspects of genomics and proteomics. At present, there are several shifts in the focus of genomics research. One is the functional genomics research that links the sequence and function of known genes. The second is the shift from mapping-based gene separation to sequence-based gene isolation. The third is to explore the pathogenesis from the cause of the disease. The fourth is the shift from disease diagnosis to disease susceptibility research. The application of Biochip provides the most basic and necessary information and basis for the above research. This becomes the main technical support for genomic informatics research. The development of bioinformatics provides an opportunity for further breakthroughs in life science and revolutionary change in the drug development process. Obtaining a sequence is only the first step in the case of the human genome. The latter step is the task of the post-Genome Era, which collects, organizes, retrieves and analyzes the structure and function of proteins expressed in the sequence. To find out the rules, bioinformatics plays a vital role in it. At present, the role of data mining in bioinformatics mainly includes the following aspects.

(1) Establishment and query of biological information database

(a) Gene and genome databases. Such as the US gene database Genbank, the European Molecular Biology Laboratory (EMBL) nucleic acid sequence database, the genome database GDB, DNA database of Japan (DDBJ) and so on. The Genbank contains all known nucleic acid and protein sequences, as well as literature and biological annotations associated with them. It was established and maintained by the National Center for Biotechnology Information (NCBI). Its data comes directly from the sequence submitted by the sequencing workers, a large number of EST sequences and other sequencing data submitted by the sequencing center, and data exchange with other data organizations. Genbank exchanges data with the database of the EMBL and DDBJ every day to synchronize the data of these three databases. The EMBL nucleic acid sequence database consists of nucleic acid sequence data maintained by the European Bioinformatics Institute (EBI). It is a comprehensive database of nucleic acid sequences because of data exchange with Genbank and DDBJ. The genomic database stores and processes genomic map data for the Human Genome Project (HGP). The DDBJ is also a comprehensive database of nucleic acid sequence that work with Genbank and EMBL nucleic acid libraries to exchange data.

(b) Protein database. Such as Protein Identification Resource (PIR), International Protein Sequence Database (PSD), Protein Data Bank (PDB), Structural Classification of Proteins (SCOP), Clusters of Orthologous Groups of proteins (COGs), *etc.* The PIR International Protein Sequence Database is the largest international public protein sequence database maintained by the Protein Information Resource, the Munich Protein Sequence Information Center and the Japan International Protein Sequence Database. PDB is the only international archive of biological macromolecular structure data, established by the Brookhaven National Laboratory. The SCOP database details the relationship among known protein structures. The COGs database is a protein encoding 21 complete genomes of bacteria, algae and eukaryotes, organized according to phylogenetic relationships.

(c) Function database. Such as the Kyoto Encyclopedia of Genes and Genomes (KEGG), the Alternative Splicing Database (ASDB), the Transcriptional Regulatory Region Database (TRRD) and so on. KEGG is a knowledge base that systematically analyzes gene functions as well as links genomic information and functional information. The interacting protein database collects experimentally validated protein-protein interactions. ASDB includes both protein libraries and nucleic acid libraries. TRRD is constructed based on the accumulated structure-function characteristics of eukaryotic gene regulatory regions.

(2) Comparison and analysis of biological information

(a) Sequence alignment. An alignment between protein sequences or between nucleic acid sequences. This includes pairwise alignments and multiple sequence alignments. Pairwise alignment is to find possible molecular evolutionary relationships by comparing similar regions and conserved sites between the two sequences. Sequence search services provided by sequence databases, such as Genbank and SWISS-PROT, are based on sequence pairwise alignment. Multiple sequence alignment is a method of comparing two or more sequences that may have a systematic evolutionary relationship. At present, the research on multi-sequence alignment is still in progress. Most of the existing algorithms are based on the idea of progressive comparison. The results of multi-sequence alignment are gradually optimized based on the pairwise

alignment of sequences. After performing multiple sequence alignments, the results can be further processed, such as constructing a profile pattern profile, clustering around the sequences to construct a molecular evolution tree.

(b) Predictive analysis of nucleic acid and protein structure and function. Prediction of nucleic acid sequences is the process of finding genes in nucleic acid sequences, finding the location of genes and the location of functional sites, and labeling known sequence patterns. In the prediction method for proteins, traditional biology believes that the sequence of a protein determines its three-dimensional structure and its function. Because X-ray crystal diffraction and NMR nuclear magnetic resonance techniques are used to determine the three-dimensional structure of proteins, the functional efficiency of biochemical methods for studying proteins is not high enough to adapt the rapid growth of protein sequence numbers. Many scientists have devoted themselves to research in recent decades. The theoretical calculation method predicts the three-dimensional structure and function of proteins, and has achieved certain results after years of efforts.

(c) The analysis of genomic sequence information and functional genomic related information. It includes the large-scale gene expression profiling, comprehensive prediction of protein function at the genomic level. Sequence analysis, gene expression, protein structure prediction, drug discovery and design are the focus of biology. There are many meaningful mining models and mining algorithms applied in these areas and achieved corresponding results.

Research has proven that data mining is a powerful tool for biological information processing. Specifically, data mining techniques applied in bioinformation knowledge discovery include the following aspects. 1 Constructing a gene database or a data warehouse. Because of the high degree of dispersion of a wide variety of DNA data, to facilitate the systematic analysis of DNA databases, data cleansing and data integration methods in data mining are required to construct integrated data warehouses and develop distributed databases. 2 Data cleansing, data integration, heterogeneous, semantic integration of distributed databases. Many countries and research organizations have established biological sequence databases, protein structures and functional databases to provide people with a wealth of information. However, these data are scattered and the storage medium is diverse. There are a large number of sequences with repeated information and some highly similar data in the same database, resulting in data redundancy. Thus, the semantic integration of this heterogeneous and widely distributed database becomes an important task. Data cleansing and data integration methods in data mining can help solve this problem. 3 DNA sequence similar searches and alignments. To identify the evolutionary relationship between a newly discovered gene and a known gene family, sequence alignment is usually required to discover the largest match between them, determining their homology or similarity. Thus, this gives a similar degree. Because of genetic data is non-digital, the non-digital sequence pattern analysis method in data mining can play an important role in gene sequence alignment. Therefore, exploring efficient search and alignment algorithms is important in sequence analysis. 4 Genomic features and analysis of simultaneous gene sequences. For a set of sequences of a gene family, the relationship among multiple sequences must be elucidated to reveal the characteristics of the entire gene family. Multiple sequence alignments have important biological significance in identifying a set of related sequences. 5 Association analysis. Most diseases are caused not by one gene, but by the result

of a combination of multiple genes. Therefore, the use of correlation analysis in data mining can help to discover the connections between genomes. Then, we reveal the genetic causes behind human diseases. 6 Biometric data visualization and visual data mining. Because of the complexity and high dimensionality of biological data, it cannot be represented by numerical formulas or logical formulas. It can be represented by graphs, trees, chains, and various visualization tools. Commonly used biological data visualization chemicals have semantic mirror technology, information wall technology, gene regulation grid and so on. In addition, the data results obtained by the data mining tool are presented to the user in the form of graphics and images. Therefore, the user can discover the rules and relationships between the data.

1.5.2 Data mining in drug sales

Commodity sale is the earliest application field of data mining. It is the most important application field at present. By collecting and processing a large amount of information related to consumer consumption behavior, we can determine the interest, consumption habits, consumption propensity and consumer demand of a particular consumer group or individual. This help us infer the consumer behavior of the corresponding consumer group or individual. Based on this, we can target marketing of specific content to the identified consumer groups. Compared with the traditional large-scale marketing methods that do not distinguish the characteristics of consumer objects, this greatly saves marketing costs and improves marketing results. Thus, it brings more profits to the enterprise.

Medicine is a special commodity. As people's demand for life and health increases, the drug sale market expands and market competition is increasing. Successful applying the model of data mining in the sales of goods to increase the sales of drugs and expand market share is an important issue for pharmaceutical manufacturers and sellers. Drug sale has a direct impact on pharmaceutical production and drug research. Therefore, applying the data mining results in drug sales to adjust the direction of drug production and drug research and development is a topic worthy of attention. As a special commodity, medicine has something in common with goods in the retail industry. The marketing of pharmaceutical supermarkets is similar to that of retail. In the marketing of retail industry, data mining technology analyzes customers' purchase baskets, assists merchants in discovering customers' consumption habits, predicts customers' purchase patterns, divides customer groups, and discovers goods. The correlation between sales volume optimizes the shelf layout, researching the best promotional time and merchandise portfolio, and understanding the status of slow-moving and best-selling merchandise. Analyzing the market share of a commodity in various chain stores, customer statistics and historical conditions is beneficial for companies to determine effective strategies for sales and advertising.

Supermarkets can collect information based on different types of customers. One is the anonymous customer, and the other is the customer who registers or has customer card (some are called honorary customer card, membership card, etc.). For the first type, the supermarket can analyze the association rules of the customer to purchase the product. When a customer purchases a certain product, the supermarket can issue a coupon for the product related to the product. For customers with customer cards, the supermarket can analyze the sequence pattern

of the purchase history and issue coupons for the items that may be purchased next time. Now, the customer database of many large supermarkets has evolved from the mode of anonymous customer management to customer card management. This realizes the marketing focus of the supermarket from commodity to customer. Promotion is a common means for shopping malls to attract customers and increase sales profits. The traditional promotion method is based on statistics. For customers who spend more than 400 yuan per month, the supermarket provides a free gift. This method has a flaw. It is mainly not for the customer, only for the fixed cost in the supermarket.

According to consumption patterns, customers can be classified into three types. One type of customers will not spend more than 400 yuan regardless of how the supermarket is promoted. The second type of customers spends more than 400 yuan per month. This type of customers cannot be attracted by the free gift. The last type of customers is the one who spend more than 400 yuan per month due to the promotion. Only the last type of customers is the real goal of this promotion strategy. To attract more customers to participate, the supermarket should use the data mining technology to design the promotion activities according to each customer's consumption level and purchase of goods. For example, it is found that a product that may be of interest to a customer is used as a promotional item through data mining. Thus, this can increase the customer's buying interest.

As people's demand for life and health increases, the drug sales market expands, and market competition is increasing. Applying the successful model of data mining in the sale of goods as a special commodity, we increase the sales volume of drugs and expand market share. These are important issues for pharmaceutical manufacturers and sellers. Drug sales have a direct impact on pharmaceutical production and drug research. Therefore, applying the data mining results in drug sales to adjust the direction of drug production, research and development is a topic worthy of attention.

1.5.3 Scientific research and statistical data mining

At present, the importance of data mining and knowledge discovery on scientific data is increasing rapidly. Many fields, such as bioinformatics, geophysics, astronomy, medicine, meteorology, particle physics, *etc.*, are experiencing exponential growth in data volumes. Experts in all fields need to define metadata standards and propose mining goals.

Data mining has been applied in different fields of information technology, geography and astronomy, including classification of cosmic image data, monitoring of crustal activity and discovery of biological models. For example, SKICAT, developed by the California Institute of Technology's Jet Propulsion Laboratory, helped astronomers discover 16 new quasars. SKICAT uses the decision tree method to construct the classifier. As a result, the resolved stars are an order of magnitude lower than that discovered by the previous method. The new method is 40 times more efficient than the previous method. In the field of biomedicine, scientists use sequence pattern analysis and similar search techniques in data mining to analyze DNA data, complete semantic integration of heterogeneous and distributed gene databases, and similar search and comparison between DNA sequences. They use correlation analysis to identify genes that appear at the same time. Then, they use path analysis to identify disease-causing genes at

different stages of the disease.

Both data mining and statistics attempt to discover a structure from the data to obtain valuable information. Thus, the data mining has been inextricably linked to statistics since the birth of it. Statistics, databases and artificial intelligence constitute the three pillars of data mining technology. Statistic is the science of collecting, presenting, analyzing, and interpreting data. Statistic is not a list of methods, but a science that processes data. Data mining is the process of extracting implicit, previously unknown, potentially useful information and knowledge from large, incomplete, noisy, fuzzy, and random data. The implementation of most core functions of data mining is supported by measurement and statistical analysis methods. Statistical thinking has a heavy responsibility in all stages of data mining. Statistics had made a significant contribution to the innovation of data mining methods.

In the process of data mining, statistics played a vital role. The tools developed by statistical methods can be used for data extraction, cleaning, transformation, integration, *etc.* Statistical logic reasoning can enable data analysts to stand. At a higher level, pattern recognition of data is performed from a broader perspective. The statistical and probability theory is used in statistical analysis method to analyze each attribute. This can find out the relationship and law between them. One of the foundations of data mining technology is statistics. Statistical analysis methods are one of the most basic data mining techniques.

While applying the general methods of data mining, scientific and statistical data mining must combine relevant subject background knowledge to improve the quality of mining results. Constructing an integration information system of database (or data warehouse), a domain-specific knowledge base and data mining system is a recommended choice.

1.5.4 Chemical data mining

(1) Overview of chemical data mining

With the support of modern analytical testing technology, the data information in the chemical industry is accumulating. The quantity and variety are growing rapidly. The application of these massive data is a serious challenge faced by chemical information workers. However, theoretical development often lags behind practice, especially in disciplines closely related to experiments, such as chemistry and chemical industry. Sometimes only a small number of data results can be explained or derived by mechanism. A large number of results still need effective data processing methods to assist in the exploration and discover the laws contained in them. With the development of information technology, many methods have been developed. Good results have been obtained. Thus, chemometrics and various methods of modeling, optimizing and controlling chemical process based on data have been developed. However, these processing methods usually only rely on the data collected through experiments under specific purposes. The amount of data is relatively small. It is not suitable for using a large number of related databases, especially for high-dimensional data, noise data, mixed type data, means of handling. In addition, the analytical capabilities of these methods are limited. The exploration of data inclusion rules (knowledge patterns) depends on the researchers' understanding of the problems. It is not beyond this understanding. It is automatically obtained from a large number of existing data sources and a wide range of useful knowledge. Therefore,

a large number of chemical information retrieval systems (chemical databases) have emerged as the times require. The various chemical databases (databases of various chemical substances, chemical map databases, etc.) have been established. In recent years, with the accumulation of chemical data and the widespread use of databases, researchers have gradually realized that the use of massive data is very difficult and inadequate, and more valuable regular information and knowledge are concealed. The way to discover more and more valuable chemical laws from chemical data is gradually becoming the focus of chemists. Therefore, chemical data mining has begun to be applied to computer chemistry as a new information technology. The key of this technology is using software to automatically discover non-obvious, potentially useful information and knowledge from vast amounts of chemical data.

Chemical data mining is a multidisciplinary research field in computer science, statistics, artificial intelligence, chemistry, and chemometrics. The key to technology is to use computer technology, mathematical models and chemical background knowledge, reveal and characterize new knowledge. The new information that is not obvious and have potential application value from massive chemical data. From an implementation perspective, it is often tightly integrated with data screening, classification, analysis, characterization, and knowledge discovery. From the perspective of achieving the goal, chemical data mining is a theory and implementation technology that studies comprehensively use various methods exploiting valuable information (or knowledge) hidden in massive chemical data. The main contents include data cleaning and data reduction, feature detection and feature reduction, classifier and category detection, and fraud detection, recognizing unusual patterns, novelty detection, construction of sample databases, knowledge discovery and rule detection from large databases, and related technologies.

(2) Computer processing of chemical structure and chemical reaction

The first difficulty that chemists have encountered in applying computers to solve chemical problems has long been the problem of computer processing of chemical structures. All areas of chemistry are closely related to the structure of compounds. Over the past 30 years, this issue has attracted extensive attention and in-depth research, thus forming an important research field of computer chemistry. After years of hard work, the theory and most of the technical problems in the computer processing of chemical structures have been basically solved. However, these methods are still limited and difficult to apply to such issues as phylogenetic structure processing, structure-activity related automation research, and reaction mechanism research. Even with the problem of determining structural processing, existing solutions are not acceptable to all chemists. Therefore, there are still some difficulties in determining the computer processing of structures, such as inorganic compounds, organometallic compounds, and tautomeric chemical structures, which require more in-depth research. In addition, these problems are the basis of many fields in computer chemistry, and their complete solution will be conducive to the development of computer chemistry.

The chemical reaction can be regarded as the conversion of some chemical structures to other chemical structures. The treatment of chemical reactions is ultimately the treatment of chemical structures. However, computer processing of chemical reactions has its own specific problems, such as identification of reaction centers, discovery of reaction knowledge, organization and utilization, and generation of similar reactions. These problems are the main

research directions in the field of computer processing chemical reactions. Their solutions will promote the development of chemical reaction databases to a higher level. In addition, through the combination of data mining technology and discovery of reaction knowledge, computer-assisted organic synthesis route design has a more solid foundation. Thus, it can be more reasonably resolved.

1.5.5 Application prospects of data mining in the field of TCM

(1) Application of data mining in the medical field

At present, the research and application of data mining are very popular. Although it has a short time, it has been widely used in other fields and has achieved great results. Although data mining in the medical field started slowly, there is bound to be broad prospects and value in the field of exploration. In addition, data mining applications and research in this field have achieved more results for subsequent research. This revealed many new ways and patterns.

(a) Application in the field of basic medicine. Many researchers are constantly exploring the use of data mining methods for DNA sequence classification, as well as analyzing the sequencing data of the genome. The nonlinear correlation statistics, such as AMI, neural network, classification and clustering algorithms, have significant value.

(b) Application in the field of clinical diagnosis and treatment. These data mining methods have more advantages than traditional statistical methods in forecasting and other applications, such as clinical diagnosis and treatment of the disease. The diagnosis, treatment and post-cure prediction of diseases, because of the intricacies of certain diseases, and medical technology itself, is a highly practical and statistical verification science. Thus, the diagnosis and treatment processes are the intertwining effect of physician knowledge and experience. Data mining can serve clinical decision making by processing and refining a large amount of valuable information contained in the patient data database. After trying and exploring, the classification analysis, rough set theory, artificial neural network and fuzzy logic analysis have their unique value in this respect. For example, Bayesian learning method for automatic diagnosis of CT images of patients, fuzzy neural network for analyzing ultrasound images of the heart, decision tree method for the choice of treatment methods, the rough sets theory in the detection of cervical cancer lesions and so on.

(c) Epidemiological studies and medical statistical methodologies. The data mining method is based on machine learning. Each of it and traditional data processing method has advantages. In practice, they are often combined and used to learn from each other. For example, the decision tree is combined with regression and neural network methods to analyze the factors affecting disease mortality. Thus, this makes up for the shortcomings of the neural network. In addition, data mining methods are effective in epidemiological research.

(d) Hospital and health management. The establishment of the Hospital Information System (HIS) provides a large amount of information resources for the scientific management of hospitals, and provides a source of information for decision support for the formulation of health care policies and the rational allocation of health resources. However, the traditional database methods and analysis methods mostly stay in the data entry, query and statistics functions. Thus, the emergence and development of data mining methods provide valuable

decision support information from the intricate and huge medical information database. With effective ways and methods, users can improve hospital service quality and management.

(e) Other aspects. For example, data mining also has exploration and applied research in drug research and development, toxicology research and so on.

In addition, data mining is used to analyze case and patient characteristics, prescription management, arrange medical plans, determine the effectiveness of prescriptions, verify the treatment mechanism of drugs, analyze health data, and determine deviations. Data mining of bioinformatics/genomics is a hot topic of current research. Incyte BioSoft uses the MineSet tool to mine gene databases, analyze the relationship between gene and disease, and successfully discover new gene patterns. The Heart Institute of San Francisco provides cardiac surveillance and treatment services to hundreds of patients from around the world each year. They use data mining technology to track the performance and efficacy of treatment physicians with more than 20,000 patients, which greatly improves the efficiency of hospital operations, improves the accuracy of doctors' diagnosis, and reduces the time for patients to stay in the hospital. For example, in the course of treatment of coronary contractions, the average time spent on medical treatment has been reduced from seven nights to three nights by analyzing the relevant factors. In this way, $500,000 is saved each year and the quality of treatment is effectively improved.

(2) Application of data mining in the field of TCM

(a) Application background of data mining in the field of TCM

In the field of TCM, database technology has been widely used. Data mining is the main link in database knowledge discovery. The establishment of back-end database is important. With the rapid development of computer technology, database technology has become more and more important. The construction of Chinese medicine database originated in the 1980s. After more than 20 years of construction, preliminary results have been achieved. TCM information databases of different types and sizes have been built in the field of Chinese medicine, such as ancient literature databases and structured databases, modern literature and factual databases, and various Chinese medicine data warehouses. In addition, since 2002, the Chinese Medicines Search Center of the State Administration of TCM has used the Virtual Research Center platform to jointly establish more than a dozen Chinese medicine universities, colleges and research institutions across the country to create a basic database of structural Chinese medicine technology. In recent years, data mining techniques for Chinese medicine treatment information have been studied by some researchers or research institutions. Since the TCM treatment system includes syndrome differentiation and disease differentiation, various researchers and research institutions have carried out a lot of research work mainly from the relationship between syndrome and disease. As a hidden and non-trivial knowledge acquisition technology, data mining is necessary for the establishment of decision support systems.

(b) Application of data mining in the study of the compatibility law of TCM

Data mining is the link of database knowledge discovery. Knowledge acquisition is its function, especially for the acquisition of implicit and non-trivial knowledge. Data warehouse is used as a data storage tool to related TCM with different structures and locations. The "multi-library integration" platform system for database integration is being built. An intelligent decision support system combining data mining with data warehouse, online analytical processing, knowledge base and model library should be established. OLAP and data mining

are used for data analysis and knowledge acquisition tools. The data in the data warehouse is purposefully analyzed. It is useful and understandable for the development of things, and the acquired knowledge is stored in the corresponding knowledge base. Based on the knowledge base and model library, the application of expert system to scientific prediction of the future development trend of issues will be the necessary way to accelerate the process of informatization, standardization and knowledge of TCM. Data mining has been widely used in the modernization research on the theory of Chinese medicine and the scientific connotation of compound compatibility. The theory of Chinese medicine is one of the core contents of the compatibility of TCM. The degree of perfection of the content affects the accuracy and effectiveness of the treatment. The understanding of properties of TCM is a process of gradual accumulation and improvement. There are still some medicinal properties of medicines are imperfect. The classification of the efficacy of TCM is based on the degree of understanding of the different category drugs by experts. Data mining technology can assist in the perfect medicinal properties research of Chinese medicine on the basis of analyzing a large amount of historical data. For example, the classification method in data mining can classify some TCMs that have not been classified according to the identification results of drug characteristics. In addition, clustering methods can be used to cluster the drug tastes. According to the similar properties of drugs, the classification should be similar. According to the characteristics of the classification and prediction, association rules can be used to correlate the drug properties of each drug in the association pattern or rule research. The data mining research on the pharmacological characteristics of Chinese medicine is of great significance for the study of the scientific law of Chinese medicine compound compatibility. TCM compound plays an important role in the science of TCM. The hundreds of thousands of TCM compounds accumulated over thousands of years. The established database of many Chinese medicine compounds are the most valuable resource and wealth of TCM. It has realized the excavation and reproduction of expert knowledge. However, the situation of “Hundred Flowers Blossoming and Hundred Schools of Contention” makes the compound often have different medicinal tastes and doses. This cannot form a deep understanding and research on the disease and compatibility law of Chinese medicine. Applying data mining methods to comprehensively collate and mine hundreds of thousands of Chinese herbal medicines provides a comprehensive understanding of the basic theories and clinical practice of Chinese medicine. Then, we study the compatibility law. The TCM compound is an organic community composed of drugs of two or more flavors based on the principle of “Jun-chen-Zuo-zhi” on the basis of syndrome differentiation. We Use the data mining method to conduct intelligent analysis of the historical data of Chinese medicine compound compatibility, achieving the understanding of the essential laws of TCM illness and compound compatibility. It provides theoretical support for effectively streamlining compound and reasonable compatibility. Among them, frequency mode, association rule, Bayesian network and other methods can analyze the pattern or rules of drug compatibility at different compatibility levels. This can realize the understanding of the common law of compound. In addition, data mining technology is widely used in the study of the relationship between the symptom and compound prescription, TCM syndrome and compound prescription. The data mining research on the compatibility law of TCM compound is in its infancy.

(c) Application of data mining in the medical field.

Mining knowledge from the huge and complicated data in the field of medicine to guide future work is rich in prospects and application value. However, it is a difficult and complex subject. The mining of medical data is the product of the fusion of computer technology, artificial intelligence, statistics and modern medicine, the process of extracting knowledge for the entire medical information database, and an important component of the scientific decision-making of the overall medical service. Due to the wide range of medical data mining objects, the algorithm requires efficient extraction of knowledge, and the decision-making recommendations require higher accuracy, and the existing medical information database is still incomplete with respect to the requirements of data mining. These require a computer, mathematics, statistics, and multi-party collaboration among medical staff, which can make greater breakthroughs in key technologies such as multi-party information fusion, efficient algorithms, and accuracy of knowledge acquisition. It is believed that with the widespread application of data mining technology, the continuous improvement of methods, and the development of achievable software, the application of data mining in the medical field will be more extensive and deeper, thereby bringing greater social and economic benefits.

References

1. J.W. Zhang. University Information Retrieval. China Water & Power Press: Beijing, China, 2004.
2. Z.X. Liu, R.S. Li, W. Ye. Introduction to Practical Information of Retrieval Technology. Tsinghua University Press: Beijing, China, 2006.
3. R. Tang, T. Yi, Y. Zhang. Modern Pharmacy Information Technology. Military Science Publishing House: Beijing, China, 2003.
4. B. Hu, Y.G. Jiang. Literature Retrieval of Traditional Chinese Medicine. Shanghai Science & Technical Publishers: Shanghai, China, 2006.
5. J.W. Yan. Literature Retrieval of Traditional Chinese Medicine. Academy Press: Beijing, China, 1995.
6. A.G. Fan, D.J. Xue. Modern Information Retrieval. Peking University Press: Beijing, China, 2006.
7. Z.Y. Yu. Retrieval and Utilization of Pharmacy. China Medical Science Press: Beijing, China, 2005.
8. Z.J. Sun, H. He. Retrieval of Pharmaceutical Information Resources. Southeast University Press: Nanjing, China, 2002.
9. Z.C. Hu. Basic Knowledge of Patents. Intellectual Property Publishing House: Beijing, China, 2004.
10. L.Q. Wo. Retrieval of Pharmaceutical Literature. China Medical Science Press: Beijing, China, 2000.
11. A.J. Zhao, Y.Y. Zeng, B.H. Xu. Network Security Technology & Application. Posts & Telecom Press: Beijing, China, 2007.
12. X.H. Ge, H. Tian, S.M. Jin. The Management of Computer Network Security. Tsinghua University Press: Beijing, China, 2008.

13. R. Tang. Modern Pharmaceutical Information Technology. Military Science Publishing House: Beijing, China, 2004.
14. X.J. Zhang. Writing and Submission of Chinese and English Medical Research Papers. People's Medical Publishing House: Beijing, China, 2008.
15. X.Y. Zhang. Retrieval of Traditional Chinese Medicine Literature. People's Medical Publishing House: Beijing, China, 2018.
16. X.Y. Zhang. Retrieval of Pharmaceutical Literature. China Press of Traditional Chinese Medicine: Beijing, China, 2017.

Chapter 2 Application and Evaluation of Data Mining Algorithms in Traditional Chinese Medicine

With the skyrocketing data in TCM, Chinese medicine data mining came into being. Under the guidance of TCM theory, scholars use Chinese medicine data mining to conduct in-depth systematic research on new drugs, prescriptions of TCM, mechanism of action, effective components and group-effect relationship. Data mining is a series of methods for exploring and calculating data mining models based on data analysis or the needs of data users. Commenting on the application status of data mining algorithms in TCM research cannot only explain the current situation of the use of data mining algorithms in TCM, but also explore the characteristics and fields of the data mining algorithm in more detail, and expand the mining algorithm in TCM research. The application provides useful reference for the in-depth study of the algorithm.

2.1 Application of data mining algorithms in TCM

The data mining task is the classification basis for statistical analysis and application analysis of common data mining. The tasks that data mining can be divided into two types, including the descriptive task and predictive task, as shown in Figure 2-1. Descriptive tasks describe the general nature of the data in the target data. Predictive tasks are summarized on the current data to predict. The two types of tasks can be classified into six categories, including characterization and differentiation, association rule analysis, classification analysis, regression analysis, cluster analysis, and outlier detection analysis. The “characterization and differentiation” task can establish database and data preprocessing, and the task of “outlier detection” is mostly done by cluster analysis method.

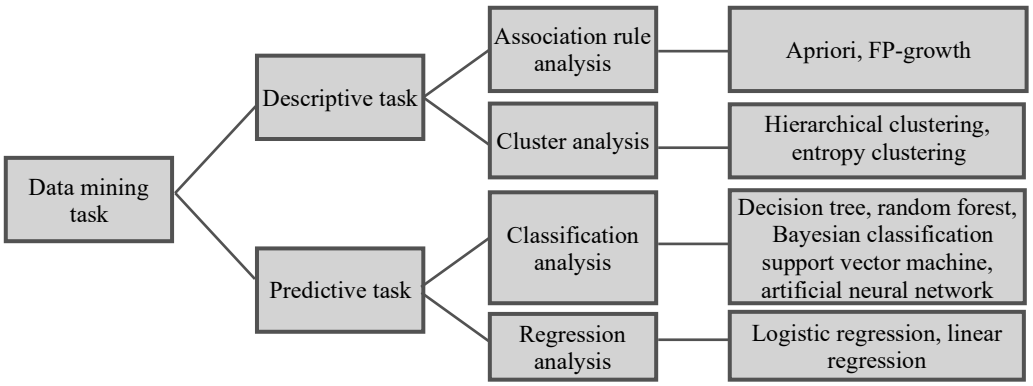


Figure 2-1. Common mining algorithms for completing Chinese medicine data mining tasks.

In this study, we used “Chinese medicine”, “Apriori”, “FP-growth”, “hierarchical clustering”, “entropy clustering”, “decision tree”, “random forest”, “Bayesian”, “support vector

machine", "artificial neural network", "Logistic regression" and "linear regression" as the key words. The data mining algorithms, published in the database from January 2000 to May 2018, relevant in Chinese medicine research, and included in the database of China Knowledge Network, Wanfang and Weipu, as well as, were searched. The name of the algorithm retrieved was based on the book "Data Mining-concepts and techniques" (third edition) edited by Jiawei Han, published by Machinery Industry Press, China. The related literature reviews are more commonly used in various data mining algorithm industries. 11 document databases were classified by the data mining algorithm, and the literature in the single literature library was classified according to the TCM research sub-area used in the algorithm. The algorithm was applied to more than 3 papers in the same field of TCM research as the algorithm. In the sub-area of TCM, if there are only one or two documents in the same field, it will not be discussed. The merger will be classified in the "other" item. Since some documents involve more than two mining algorithms, such as a document applied to both decision trees and random forests. The literature is classified into the "decision tree" literature database and also summarized in the "random forest" literature from the library. Finally, based on the retrieved literature results, the author classified according to the TCM research sub-area used in the algorithm. Then, they combined the characteristics of each algorithm to comprehensively review the commonly used algorithms for the four types of tasks, including association rule analysis, classification analysis, regression analysis and cluster analysis of the application status and characteristics of TCM research.

2.1.1 Literature search results of Chinese medicine research using the data mining algorithm

A total of 607 related articles were retrieved, of which 573 were valid articles, and include into 11 document databases, as shown in Table 2-1.

Table 2-1. Search results of effective literature of Chinese medicine data mining algorithm.

Algorithm	Number of documents (Article)
Apriori	39
FP-growth	14
Hierarchical clustering	28
Entropy clustering	113
Decision Tree	44
Random forest	22
Bayes	34
Support vector machine	87
Artificial neural network	129
Logistic regression	51
Linear regression	46
Valid literature	573

Note: Some documents involve more than two algorithms and are therefore repeatedly classified.

2.1.2 Classification of TCM research sub-area using data mining algorithm

According to statistics, the mining algorithms that can be used in TCM researches are shown in Table 2-2. In the study of the compatibility law of prescriptions, the algorithms and the application of data mining technology in this research sub-field is the most widely. In addition, artificial neural networks and support vector machine algorithms can be used in the widest range and can be applied to multiple professional sub-areas.

Table 2-2. The application of mining algorithms in the TCM sub-area.

TCM sub-area	Number of documents (Article)	Mining algorithm
Formula compatibility	189	Apriori, FP-growth, hierarchical clustering, entropy clustering, decision tree, artificial neural network, Bayesian classification, logistic regression
Medicine analysis	83	Hierarchical clustering, random forest, decision tree, artificial neural network, linear regression, support vector machine
TCM research	48	Apriori, decision tree, random forest, artificial neural network, Bayesian classification, support vector machine
Formulation process research	36	Support vector machine, artificial neural network
TCM security evaluation	25	Decision tree, random forest, logistic regression, support vector machine
Research on the efficacy of TCM	20	Bayesian classification, support vector machine, artificial neural network
TCM identification	14	Support vector machine, artificial neural network
Drug efficacy evaluation	13	Logistic regression, decision tree, Bayesian classification
Study on active ingredients of TCM	10	Support vector machine, artificial neural network
Medical case study	4	Apriori
Spectral effect relationship	3	Linear regression
Quantitative diagnostic criteria	3	Logistic regression
Patent research	3	Hierarchical clustering
Drug target prediction	3	Random forest
Drug cost-effectiveness analysis	3	Bayesian classification
Drug dose study	3	Linear regression
Pharmacy Dispensing Research	2	FP-growth
Others	145	-

Note: Some documents involve more than two algorithms and are therefore repeatedly classified.

2.1.3 Status and characteristics of data mining algorithms in TCM

(1) Association rule analysis

Association rule analysis is primarily used to discover meaningful connections that are hidden in large data sets. It mainly consists of two phases. One is to discover all the frequent itemsets from the dataset, and the other one is to generate the association rules from the frequent itemsets. In TCM research, the algorithms commonly used to complete association rule analysis

tasks are Apriori algorithm and FP-growth algorithm.

(a) Apriori algorithm. The Apriori algorithm is a basic algorithm for discovering frequent itemsets. It uses an iterative method of layer-by-layer search to scan the database. It collects items that satisfy the minimum support. This help to discover frequent itemsets, and directly generate strong association rules. According to the 39 related literatures retrieved, Apriori algorithm is mainly applied to several research sub-areas such as prescription compatibility, medical research, and TCM research. (I) Study on the compatibility of prescription. Collecting clinical prescriptions, famous or classics treatment for a specific disease, we establish a database after standardization, and use the Apriori algorithm to mine the core drug pairs in the database ("core drug pairs" are more common noun in the mining of prescription compatibility rules), and high-frequency drug combination. Combined with the characteristics of drug efficacy and under the guidance of TCM theory, we can explain the compatibility of the prescription. The analysis results have been important guiding significance for clinical drug use or new drug development. (II) medical research. The Apriori algorithm was used to mine the association rules of TCM in typical cases, which provides reference for the clinical treatment of TCM, the teaching of Chinese medicine and the development of proprietary Chinese medicines. (III) Chinese medicine research. The Apriori algorithm is used to dig out the organic relationship between drug and efficacy. The theory of TCM is explained by modern scientific theory to promote the modernization of TCM.

(b) FP-growth algorithm. The FP-growth algorithm is an optimization process of Apriori algorithm. This solves the problem that Apriori algorithm generates a large number of candidate sets during operation. It has strong generalization and robustness in large-scale prescription dataset mining. It is more suitable for large-scale recipe database mining. According to the 14 related literatures retrieved, the FP-growth algorithm is mainly used in pharmacy dispensing research and prescription compatibility research. (I) pharmacy dispensing research. The FP-growth algorithm is used to count the common use of drugs, small-package Chinese medicine decoction pieces, and optimize drug position management, thereby improving the efficiency and accuracy of pharmacy dispensing prescriptions. (II) prescription compatibility research. The FP algorithm is used to mine the frequent itemsets in the dataset to discover the core drug pairs and high-frequency drug combinations in the prescription pool, and to explain the prescription compatibility from a scientific perspective.

(2) Cluster analysis

Cluster analysis is the process of dividing a data set into several sub-data sets. Thus, the objects in the sub-data set are similar to each other and different from the objects in other sub-data sets. The methods of cluster analysis mainly include clustering, hierarchical clustering, complex system entropy clustering, density-based clustering, grid-based clustering and so on. After searching the keywords of the algorithm, the application methods of "division method", "density-based method" and "grid-based method" are few. Therefore, we focus on the application of hierarchical clustering method and another combination with information entropy and complex system entropy clustering algorithm.

(a) Hierarchical clustering. Hierarchical clustering divides data into groups on different layers. According to the 28 related literatures retrieved, the hierarchical clustering algorithm is mainly applied to TCM research sub-areas such as prescription compatibility, drug analysis and

patent research. (I) Prescription compatibility. We use the hierarchical clustering algorithm to form a single disease syndrome and a drug combination under a single prescription, or to form a cluster based on natural taste, based on which the association rule analysis is continued, and what results have been obtained by mining the core combination of drugs. (II) Drug analysis. The hierarchical clustering method is used to classify the various active ingredients or molecular skeletons of TCMs, and to analyze the distant and close relationship of chemical components. (III) Patent research. The hierarchical clustering method is used to reveal the law of the development of Chinese patent compound and how the patents drive the internal mechanism of industry development.

(b) Complex system entropy clustering. Complex system entropy clustering is based on the "information entropy" theory proposed by Shannon. By calculating the correlation coefficient between each drug and other variables, we judged whether the variables are positively related. If any two variables between the three variables are positively related, then these three variables are clustered together. This resulting in a series of greater related drug core combinations. The algorithm is an unsupervised pattern discovery algorithm. The clustering task is completed by calculating the entropy value, which is especially suitable for the highly discrete Chinese medicine formula data. According to the relevant literatures of the search, the entropy clustering of complex systems mainly uses the TCM inheritance platform software for mining analysis. The entropy clustering algorithm is used to mine the core combination of the prescriptions to analyze the compatibility rules. In addition, this can be used as a mining tool to discover new prescriptions for TCM.

(3) Classification analysis

Classification analysis is an important form of analysis of predictive data, consisting of two phases, including the learning phase of building the model and the modeling phase used to predict the characteristics of the data. Common algorithms for performing classification analysis tasks include the decision tree, Bayesian classification, random forests, support vector machines, and artificial neural networks.

(a) Decision tree. The decision tree is a tree structure similar to the flow chart. Based on the known probability of occurrence of various situations, the decision tree is used to obtain the probability that the expected value of the net present value is no less than 0. The project risk is evaluated, and the feasibility is determined. According to the 44 related literatures obtained from the search, the decision tree algorithm is mainly applied to several TCM research sub-fields, such as TCM research, drug safety research, drug analysis, drug efficacy evaluation and prescription compatibility. (I) TCM research. Researchers first summarize the nature effects of known drugs, then use the decision tree algorithm to predict unknown or missing medicinal properties, or to study drug-law and natural-effect relationships. (II) TCM safety evaluation. Epidemiological indicators, drug-related factors or chemical compositions of TCM are used as variables in decision tree analysis to predict the probability of adverse reactions or whether the drugs have organ toxicity. (III) Drug efficacy evaluation. Based on the four type of diagnostic information and disease diagnosis indicators of patients of TCM, the decision tree model is used to explore the relationship between clinical detection indicators and disease information as well as treatment effects. This has important reference value for clinical treatment. (IV) Drug analysis. We use the decision tree to build a mining model to predict the characteristic

parameters. This has a great influence on fingerprint evaluation or drug quantitative analysis. Alternatively, we can use the decision tree to establish a digital evaluation system for multi-dimensional and multi-interest features of fingerprints, thereby guiding optimization of experimental operating conditions and improving the accuracy of drug analysis results.

(b) Random forest. Random forest refers to a classifier that trains and predicts multiple trees through the idea of integrated learning. The basic unit is the decision tree. From an intuitive point of view, each decision tree is a classifier. For an input sample, each tree will give its own classification choice and vote for it. The random forest integrates all the classification voting results, specifying the category with the most votes as the final output. According to the 22 related literatures retrieved, the random forest algorithm is mainly applied to TCM research sub-fields such as drug analysis, TCM research, drug target prediction and TCM security evaluation. (I) Drug target prediction. A series of protein databases is used to establish a random forest model of drug composition-targets for predicting and identifying the target of active ingredients of TCM. Compared with traditional pharmacological experimental methods, model identification is not only efficient but also easy to operate. (II) Drug analysis. The application process is similar to the decision tree. (III) TCM security evaluation. Its application process is similar to a decision tree. (IV) TCM research. Its application process is similar to a decision tree. In these three research sub-areas of (II), (III), and (IV), the analysis process of the random forest and the decision tree algorithm is roughly the same. However, in view of the fact that the decision tree is prone to over-fitting, the established classification model can be compared with the existing experimental data set, cannot adapt to other datasets, and predicts poorly for unknown datasets. For random forests, although a single tree is prone to over-fitting, because of the existence of multiple trees, the increase in the breadth of attributes can also eliminate over-fitting. Therefore, the model based on forests is more generalized and more practical.

(c) Bayesian classification. Bayesian classification is a statistical classification method based on Bayes' theorem. Its definition expression is $P(H/X)=P(X/H)\times P(H)/P(X)$. The items to be classified are given. The Bayesian theorem is used to solve the probability that the item appears in different categories. The item with the highest probability is defined as the category, to which the item to be classified belongs. According to the 34 related literatures retrieved, the Bayesian classification algorithm is applied to TCM research sub-fields such as TCM research, prescription compatibility, TCM efficacy research, drug efficacy evaluation, and drug cost-effectiveness evaluation. (I) TCM research. The collected drug clinical data, physiological and biochemical indicators, and "primitive elements" are used as network nodes (also called variables) to establish a network topology map and a conditional probability table to predict the four flavors of TCM or its components and efficacy. (II) Prescription compatibility. It is inconsistent with Table 2-2. We collect key data, such as prescriptions, syndromes, appearances and efficiencies of medicinal odors, perform statistics, and then establish Bayesian classification to predict the efficacy of drugs, or combine with other algorithms to study the prescriptions, and analyze the relationship among "Disease-Syndrome-Prescription-Medicine". (III) TCM efficacy study. The collected pharmacological efficacy indicators will be used to establish a Bayesian model to predict the efficacy of TCMs and components. (IV) Drug efficacy evaluation. It is inconsistent with Table 2-2. Mainly based on the Bayesian theory to establish a mesh meta-analysis, we collect the basic functions of drugs, indications, safety, evaluation

status and other indicators, establish a new systematic evaluation method, and perform complex synthesis of drug efficacy and safety comparative analysis. (V) Drug cost-effectiveness evaluation. Based on the theory of Pharmacoeconomics, the Bayesian mixed treatment comparison method is designed to provide the decision-making basis for clinical rational drug use, and provides a methodological reference for pharmacoeconomic evaluation.

(d) Support vector machine. The support vector machine is a two-class model, which aims to find a support vector from the sample, and can construct the best classification hyperplane to segment the sample to maximize the interval. According to the 87 related literatures obtained from the search, the support vector machine algorithm is mainly applied to several fields of TCM such as drug analysis, formulation process research, TCM research, TCM identification, active ingredients of TCM, TCM security evaluation, and TCM efficacy. (I) Drug analysis. The use of support vector machine and infrared spectroscopy to establish a rapid drug detection model provides a new method for real-time monitoring and quality control of product quality. (II) Formulation process research. The support vector machine is used to establish the prediction model for the preparation conditions such as drug extraction time, solvent amount and extraction rate, to obtain the optimal process parameters. (III) TCM research. Taking the statistical results of the content of elements or chemical components as the characteristic index of the drug classification, the support vector machine is used to establish the drug identification model, to clarify the internal mechanism of the drug. (IV) TCM identification. The support vector machine is combined with the spectroscopy technology to perform non-destructive rapid identification of TCM and improve the rate of identification of TCM. (V) Active ingredients of TCM. Using support vector machine to construct the relationship between chemical composition and pharmacodynamics, we can establish a “group-effect relationship” model to predict the activity of compounds, which is of great significance for the in-depth study of new drugs. (VI) TCM security evaluation. A support vector machine model based on toxicity markers or related physicochemical properties to establish toxicity discrimination provides a new method for the study of security of TCM. (VII) TCM efficacy study. Using the support vector machine to establish the efficacy classification model, predicting the efficacy of different combinations of compounds, has a good application value for the secondary development of TCM compound.

(e) Artificial neural network. The artificial neural network is based on the basic principles of neural networks in biology and the knowledge of network topology, and simulates a mathematical model of the complex information processing mechanism of the human brain’s nervous system. According to the 129 related literatures retrieved, the artificial neural network is mainly used in several research fields of TCM, such as drug analysis, formulation process research, prescription compatibility, TCM efficacy research, TCM identification, TCM research, and research of active ingredient of TCM. (I) Prescription compatibility. The artificial neural network is used to correlate the prescriptions of different ratios with the pharmacodynamic indicators, and the combination of the best efficacy is better than the prescription. (II) Drug analysis. Using artificial neural networks or combining with infrared spectroscopy to establish a rapid drug detection model, provides a new method for real-time monitoring and quality control of product quality. (III) Formulation process research. Based on the formulation conditions such as drug extraction time, solvent amount and extraction rate, a prediction model

is established, and the artificial neural network is used to optimize the process parameters. (IV) TCM efficacy study. The collected compounds, medicinal properties, pharmacological effects and other indicators are used to predict and classify the efficacy of new or unknown components of artificial neural networks. (V) TCM research. The drug-related information collected is used to establish a drug identification model using an artificial neural network. Thus, the prediction of the drug properties of unknown drugs has obtained reliable results. (VI) TCM identification. The combination of artificial neural network and spectroscopy technology can quickly identify TCM and improve the rate of identification. (VII) The research of active ingredient of TCM. The artificial neural network is used to construct a model of the relationship between chemical composition and pharmacodynamics, so as to accurately predict the activity of the compound and promote the development of new drugs. In research sub-areas such as drug analysis and TCM research, the application methods, steps and purposes of artificial neural networks are similar to those of support vector machines. Combining the characteristics of the two algorithms, the analysis of the output of the analysis results is suitable for multi-classification research, and the two-class study of multi-variable small samples is suitable for solving with the support vector machine algorithm.

(4) Regression analysis

Classification and regression are the two main types of prediction problems. The differences are from the type of output variables. Classification is a qualitative output used to predict discrete variables. Regression is a quantitative output used to predict continuous variables. Common algorithms used to perform regression tasks are linear regression and logistic regression.

(a) Linear regression. Linear regression is a statistical analysis method that uses the regression analysis in mathematical statistics to determine the quantitative relationship between two or more variables. According to the 46 related literatures obtained from the search, the linear regression algorithm is mainly applied to fields of drug analysis, spectral effect relationship, drug dose study, and other Chinese medicine research. (I) Drug analysis. The regression equation is used to establish the regression equation. The correlation coefficient and the line graph are obtained to reflect the relationship between the drug concentration and the instrument response value. This is mainly used to verify the accuracy of the analytical methodology. (II) Spectrum effect relationship. The regression algorithm is used to study the relationship between the information contained in the fingerprint and the efficacy of the drug, and reveal its correlation. (III) Drug dose study. The regression algorithm is used to optimize the dose ratio of each drug in the prescription. The dose study is also an important part of the prescription compatibility study.

(b) Logistic regression. Linear regression requires that the dependent variable must be a continuous data variable. The logistic regression requires that the dependent variable must be a categorical variable. The categorical variable is classified as two or more-category. Therefore, logistic regression solves the problem of classification.

According to the 51 related literatures obtained from the search, the logistic regression algorithm is mainly applied to TCM research sub-areas such as prescription compatibility, drug safety evaluation, drug efficacy evaluation, and quantitative diagnostic criteria. (I) Prescription compatibility. Most of the researchers use the logistic regression algorithm to establish a model

for disease typing, or to establish multiple regression models based on factors such as drug-induced taste return, and provide reference for doctors. (II) Drug safety evaluation. Logistic regression algorithm is used to analyze the related factors and occurrence of adverse reactions or toxicity induced by TCM, which promoting the clinical rational drug use. (III) Drug efficacy evaluation. Logistic regression is used to analyze factors that influence efficacy, complications, or compliance. (IV) Quantify diagnostic criteria. Firstly, the clinical and physiological factors of Chinese and Western medicine are collected, and then the logistic regression analysis is used to establish a diseasing syndrome classification model for disease prediction.

2.1.4 Discussion on the application of data mining algorithm in TCM

Through analysis and research, there are many types of data mining algorithms in TCM. The research scope covers more than a dozen research sub-areas such as prescription compatibility and drug analysis. Especially in the analysis of prescription compatibility rules, the application of mining algorithm is quite mature, and standard pattern operation has been formed. The research of Zou Jin et al. and Dong Junlong also reported the use of multiple algorithms combined with the compatibility of the other agents to further explore and promote the Chinese medicine. The discovery of prescriptions provides means for the development of new drugs. In the sub-areas, such as drug dose research, active ingredient research of TCM, and drug efficacy evaluation, data mining algorithms are gradually developed. However, there are few reports on related applications, and generally prefer to use traditional clinical trials or drug trials as research and analysis methods. Solving problems by cumbersome and complicated experiments is consumable. In addition, labor-intensive, inefficient, and the large amount of data obtained from experiments cannot be effectively processed. The author suggests combining data mining methods with traditional experiments, using results from artificial intelligence algorithms to guide drug testing. Therefore, we can reduce unnecessary losses or avoid the processing of cumbersome data.

In general, there are few reports on the application of two or more algorithms at present, which only appear in sub-areas such as prescription compatibility, drug analysis, and TCM research. If the TCM data can be processed, the advantages of each algorithm can be fully combined to promote the application of data mining technology in TCM research. In addition, in the research of TCM identification and formulation process technology, it is recommended to use support vector machine and artificial neural network to classify and predict multiple drug attributes or process conditions.

In summary, we study the application of various data mining algorithms in sub-fields of TCM, with certain reference value. In the future research, we will expand the application field of data mining algorithms, promote the application of mining algorithms in sub-domains, improve the credibility of results, and promote the application of data mining in TCM research. This provides strong technical support for the modernization of TCM.

2.2 Evaluation of evidence-based mining method based on TCM

2.2.1 Overview of TCM efficacy data through mining method and evidence-based science

Data mining is also known as Knowledge-Discovery in Databases (KDD). Its essence is to assist humans in analyzing and understanding data. By processing and transforming rich data, we summary internal rules of data, predict unknown attributes of data, and draw interesting rule patterns or laws. Thus, we can make better decisions. The basic steps of data mining include data cleansing, data integration, data selection, data exchange, data mining, pattern evaluation, and knowledge representation. The main tasks of data mining technology can be summarized as two types of description and prediction. At present, a large number of problems have been solved for various disciplines, and they themselves have been continuously developed in the process of solving problems. Statistics show that data mining has been widely used in TCM research. The application of TCM in terms of natural effects has been reported in the literature since 2000, and the literature publishing trend has increased year by year. The benign interaction brought about by the interdisciplinary integration has undoubtedly promoted the development of TCM and the development of data mining itself.

There are two kinds of data mining methods, including the traditional one and improved one. When the traditional mining model cannot meet the mining demands, it will prompt researchers to improve it. Researchers usually establish the model and prove its rationality to clarify the correlation and related strength between the research contents. Because of some shortcomings found in the practice research process, the researchers are continually improved and optimized the data mining methods, so that they can do it more quickly. More valuable information can be mined from vast amounts of information. This support decision makers to make more reliable decisions.

(1) Research status of commonly used data mining methods of TCM property efficacy

The important components of the tetralogy theory of TCM include not only the properties and efficacy, but also the primary biomass metabolites, secondary biomass metabolites, inorganic components and other material components. The application of data mining in the efficacy of TCM is related to the material basis, such as the relationship between the properties of TCM and monosaccharides, lipids, trace elements, and other pharmacological indicators, the efficacy of TCM or the efficacy and substance basis of TCM compound, the relationship between the pharmacological properties of TCM components and pharmacological indicators. In addition, many researchers explore the relationship between natural taste and drug efficacy, cold and hot drug properties and whether it has immunological properties from statistical analysis. The research on the effectiveness of data mining in TCM usually includes the study of the prescription compatibility, dose-effect relationship, the law of medication, the rule of Chinese medicine, prescriptions, drug characteristics, clinical use, clinical drug experience, prescription drug, quality evaluation, efficacy classification, fighting spectrum arrangement, drug placement law, prescription drug, main drug analysis, diagnosis and treatment law, Chinese medicine resources analysis, law of syndrome-symptom-rule-prescription, and channel tropism judgment research. In addition, data mining plays an important role in the processing

of experimental data in TCM analysis, TCM chemistry, and TCM pharmacology. It is an indispensable technical support for TCM.

(2) Common data mining methods of TCM efficacy

Data mining has long been used to solve the problems related to the efficacy of TCM with the wide range of applications. The research includes compound efficacy prediction, single-taste TCM efficacy prediction, component efficacy prediction, cold-hot drug discrimination, neutral-nonneutral drug discrimination, pharmacodynamic relationship, natural-effect correlation, pharmacodynamic index prediction, neuroimmunological prediction, neurological prediction, etc. The application in the prediction of the efficacy of TCM is mainly based on the prediction of the efficacy of TCM compound, the prediction of the efficacy of single TCM, and the discrimination of cold-hot TCM. The distribution of data mining methods in these three aspects is shown in Table 2-3.

Table 2-3. Application of data mining methods in the prediction of TCM efficacy.

Mining content	Single TCM efficacy	Compound effect	Cold-hot drug discrimination
Mining method	Neural Networks	High dimensional data reduction	Artificial neural networks
	Principal component analysis	Neural networks	High dimensional data reduction
	Decision tree	Rough set	Linear discriminant analysis
	Least squares	Support vector machine	Support vector machine
	Bayesian network	Space vector	Principal component analysis
	Support vector machine	-	Fisher discriminant analysis
	Correspondence analysis	-	Bayesian network
	-	-	Discriminant analysis
	-	-	Return tree
	-	-	Decision tree

(a) Support vector machine. Support Vector Machine (SVM) is a machine learning method with classification function. It has the good processing ability for both linear and nonlinear problems. The disadvantage is that it is difficult to implement large-scale training samples. The relationship between the properties as well as efficacy of TCMs and the basic information of TCMs is non-linear. Based on this commonality, support vector machine has been widely used in drug discrimination, efficacy prediction of single TCM, and efficacy prediction of compound TCM. There are a lot of research reports in detail. The support vector machine mining is relatively efficient for the TCM, and can be combined with appropriate data mining methods for specific problems. This helps achieve better mining results. For example, Wang Xiaoyan *et al.* used the support vector machine to establish a model to study the correlation between polysaccharides, and TCM cold-heat properties. The experimental prediction accuracy can reach 100%.

(b) Regression discriminant analysis. The ultimate goal of logistic regression model and discriminant analysis is to achieve classification effect, which has been widely used in enterprise management, economic crisis, risk management assessment, and meteorological prediction. In 2004, Kutner proposed a lot of useful conclusions about Logistic regression application. In 2009, Wang Guofu *et al*. proposed that the application of logistic regression to

discriminant analysis can improve the discriminant efficiency. They prove the validity through case data analysis. In 2014, Yin Jian *et al.* solved the two-class problem by discriminant analysis and Logistic regression combination method, and the correct rate was higher than that by both logistic regression and discriminant analysis. Logistic-Discrimina analysis can discriminate non-normal maternal, because it is applicable to both discrete and continuous variable attribute factors, and the modeling is not complicated. In addition, its application in TCM is dominated by efficacy classification or drug differentiation.

(c) Principal component analysis-linear-discrimina analysis. Principal component analysis (PCA) and linear discriminant analysis (LDA) are of great value in data processing. Linear discriminant analysis is classified by projection idea. Principal component analysis is to replace comprehensive index by fewer representative indicators. The thoughts are classified. Compared with principal component analysis, linear discriminant analysis is better at directly dealing with the relationship between classes. However, it is weaker than PCA in the processing of some problems. Therefore, it is necessary to improve and combine these two. The case study proves that Principal Component Analysis-Linear-Discrimina Analysis is beneficial to the solution of small samples and high-dimensional problems. This has been important guiding significance for solving practical problems. The combination of principal component analysis and linear discriminant analysis has been applied to comprehensively evaluate the survivability of backbone grids with classification information. The application value is high.

(d) Partial least squares discrimina analysis. Partial Least Squares-Discrimination Analysis (PLS-DA) is widely used in TCM research. Partial least squares have been gradually applied to biological information and informationomics. Discrimina analysis is the best method for determining partial least squares after dimension reduction. Chen Jiawei *et al.* used spectral imaging technology and PLS-DA to identify American ginseng. This is not only fast and non-destructive, but also verified with the recognition rate of 96.67%. Ran Jian et al. used the method of 1H-NMR and PLS-DA to accurately and quickly determine whether the deer tortoise wine is qualified, thereby controlling the quality of the deer tortoise wine. Wang Chenggang et al. established the animal cold-heat syndrome discriminant model by PLS-DA method, which has high accuracy and fitting degree. This has high application value in the evaluation of animal cold-heat syndrome. Liu Wenhui et al. established two models and verified the discriminant accuracy by using PLS-DA on the basic information of 1,725 TCMs in Chinese Materia Medica. The experimental results show that the accuracy of PLS-DA drug identification is as high as 85%. Wang Yun, Zhou Zhengli, Bin Nie, etc. also established the cold-hot drug discrimination model of TCM through PLS-DA method, and obtained good test results.

(e) Random forest model. The random forest model is used to determine the cold-hot drug properties through the indicator pathway model. It is characterized by the evaluation of decision variables, high accuracy, and small calculation. Nie Bin *et al.* used a random forest model to establish a cold-hot drug discrimination model for TCM through metabolomics data. The effective rate of model discrimination is greater than 90%. Zheng Tingting *et al.* used the random forest model to seek the optimal combination of lung cancer treatment and verified it through experiments. This proves that it can provide reference in compound combination optimization. Hui Na *et al*. established the Codonopsis classification model by random forest, decision tree and other data mining methods. According to the comparison among the model

prediction results, the random forest model has the best effect. Wu Siyuan *et al.* established a cold-hot drug discrimination model through random forest method. The model has good classification and prediction ability.

(f) Linear discrimina analysis. Linear discrimina analysis is a commonly used statistical method that plays a significant role in qualitative analysis. It can identify unknown sample categories. However, because of the complexity and efficacy of TCM, data often appear to be more complex in terms of high-dimensionality and multiple linearity. Linear discrimina analysis has certain limitations in the application of TCM. At present, linear discrimina analysis has been widely used in electronic nose identification, authenticity identification, face recognition, image extraction, feature extraction, motion recognition, *etc*. Many researchers have applied it to the judgment of TCM, and the accuracy rate is about 70%.

(g) Artificial neural network. Artificial neural network can be divided into BP neural network, probabilistic neural network, linear neural network, *etc.* It is composed of a large number of simple neurons. It is established by adjusting the number of neurons and the number of nonlinear network layers. Li Weiwei *et al*. established BP neural network efficacy prediction model of Chinese herbal compound, the prediction accuracy rate is as high as 92.5%. Liu Liping *et al.* used BP neural network to establish and use the prediction model of tonic efficacy. The prediction accuracy rate of the model reaches 83.33%. This indicates that BP neural network is better in the application of Chinese medicine compound prediction. There are many researches on neural network in TCM. In recent years, it has been gradually applied to the field of network pharmacology, which has a good application prospect.

(3) The combination application of data mining method of TCM efficacy

The joint use of data mining methods can achieve the complementary advantages of a single mining method. This can realize the purpose that a single mining method cannot achieve, and can greatly improve the mining efficiency. The following is a description of commonly used principal component analysis combined with artificial neural network, rough set, support vector machine, principal component analysis and support vector machine.

(a) The combination of principal component analysis and artificial neural networks. When principal component analysis is used in combination with artificial neural network, the data processing process can be completed on the computer through statistical software. The method is simple. It improves the classification efficiency, and the discriminant result is better than that by the single method. For example, Gao Jinhong *et al.* used the combination of principal component analysis and artificial neural network to classify the efficacy of 20 TCMs in 2009. The discriminating basis is the trace elements contained in TCMs, and the classification accuracy rate reaches 100%. The purpose of principal component analysis in this study is to achieve the effect of reducing the dimension. The artificial neural network is trained by the sample, and then used for classification research after passing the test.

(b) The combination of rough set and support vector machine. The purpose of the rough set in the study is to simplify the sample properties of TCM. The support vector machine is also classified and studied by establishing the model after continuous debugging. In the use of combination of rough set and support vector machine, the rough set reduces the computational complexity of the original support vector machine, and the support vector machine makes up for the limitation of the rough set single kernel function. The combination of the two eliminates

the redundant attribute and improves the model efficiency. For example, Wu Huimin *et al*. used the rough set reduction attribute, and then used the support vector machine to establish the efficacy model. The combination of the two improved the classification efficiency of the TCM for osteoarthritis. The accuracy rate reaches 80%.

(c) The combination of principal component analysis and support vector machine. When principal component analysis and support vector machine are used together, the main method is to use the principal component analysis method to select important variables in the model construction process. Then, it is combined with the small sample classification and the characteristics of better generalization ability of support vector machine. The advantages are complemented, and the discrimina effect is good. For example, Deng Jiagang *et al.* established a pharmacological flatness or non-flatness discriminant model through the combination of principal component analysis and support vector machine. The average discriminant accuracy rate reaches 85%.

In addition to the above-mentioned combination, there is the combination use of rough set and neural network, combination use of partial least square and linear discrimina analysis, combination use of least square and support vector machine as well as partial minimum. The combination of the two-multiple discrimina analysis and the Bayesian network model has been applied in the TCM efficacy mining and achieved better results. From related research, the combination use of data mining methods of TCM efficacy is almost matched with ones such as the rough set, principal component analysis, partial least squares method, which reduces the dimension or attribute. Attribute reduction can eliminate irrelevant or unimportant attribute, and maintaining knowledge classification or decision-making ability. It is widely used in data analysis, machine learning and knowledge discovery. We can use its features to simplify model complexity and improve models. Training speed and strengthen model recognition performance will make greater contributions to the exploration of TCM efficacy.

Data mining methods have been applied to a variety of fields. The combination of different fields and data mining methods can mutually promote each other. At present, the application of data mining methods in TCM is still increasing. In addition to the prediction of the TCM efficacy, it involves neuroimmune prediction and neurological prediction. It has been deeply researched with traditional fields, such as network pharmacology. The new research hotspots have helped solve many hot and difficult problems. This becomes an indispensable technical support for internationalization of TCM.

(4) Introduction to evidence-based science and its application

Evidence-based medicine was born in 1992. After more than 20 years of development, it has become a very influential new subject in the medical field. At present, applications of evidence-based clinical medicine have been extended to public health, and will be pushed to a wider field of discipline. The first issue of the “Evidence-Based Traditional Chinese Medicine” monograph published on October 12, 2018 promoted the production and application of high-quality evidence of TCM, and enriched its connotation. With the advent of the era of big data, the technology of evidence-based evaluation continues to develop. The application of evidence-based method digitizes the information generated in practice. With the help of big data management and tool, from different perspectives of thinking, scholars reproduce, analyze, and reconstruct the cumbersome data. This solves the shortcomings of small sample experiments.

It is more in line with the research characteristics of TCM.

The five steps of evidence-based practice are shown in Figure 2-2. Their purpose is to evaluate decisions made based on the best evidence. However, the evidence will be more abundant based on the improvement of the quality of basic research and the increase of the quantity. Therefore, the update of evidence needs to pay enough attention to strengthen the aftereffect evaluation and stop at the best. At present, evidence-based science has been applied to different aspects of many fields. Although the research objects are different, the basic principles are the same. The data generated by the data mining method in the experiment of TCM efficacy prediction usually are the total sample size, the number of training sets, the number of test sets, the accurate quantity predicted and the rate. The results of different authors using the same data mining method to predict the efficacy of TCM can be regarded as the comparison between the experimental group and the control group in the randomized trial. The research object is the discrimination and prediction effect of the data mining method on the efficacy of TCM. The measures are different data mining methods. The outcome indicator is the prediction accuracy rate. The application of data mining methods in the efficacy of TCM conforms to the PICOS principle constructed by evidence-based scientific issues. It is feasible to apply the evidence-based method to the data mining method in the evaluation of the efficacy of TCM. It can further utilize the research data of some basic experiments and present it in the form of knowledge.

(5) The research purpose and significance

At present, the use of data mining methods to mine TCM information is usually based on modeling, such as the natural taste, pharmacological indicator, spectral chromatogram, sugar, lipid and other chemical component. However, the TCM efficacy itself is more complex. The relationship among efficacies is not linear. The efficacy of TCM has a certain correlation with the primary biomass metabolite and secondary biomass metabolite. A large number of material components have become a complex study of the efficacy of TCM. The task of compound contains a variety of TCMs. The compound efficacy is not the sum of the efficacy of single TCM. Thus, the research task of TCM is arduous and significant. The application of data mining methods in the efficacy of TCM can quickly and efficiently explore TCM information, and promote the modernization of TCM. However, this also requires a large amount of basic research to support, such as the detection of various chemical components, the statistical analysis of the efficacy of natural taste, and the detection of pharmacological indicators. The evidence-based method is based on the evaluation of the best evidence to make the decision. The evidence-based method is used to evaluate the data mining method of TCM efficacy, which can make full use of the experimental data generated, saving manpower, material resources, financial resources and optimizing data Mining methods for reference.

2.2.2 The distribution law of data mining in TCM efficacy

(1) Literature

(a) Source of information. Based on consulting data mining related books and literatures, we determined search terms, and carried out the professional search in CNKI (China Knowledge Resource Integrated Database), VIP (Chinese Science and Technology Journal Full-text Database) and WanfangData. Then, we collected data mining methods in the relevant literature on the application of TCM efficacy. Finally, we manually retrieved full-text documents that cannot be obtained through professional search.

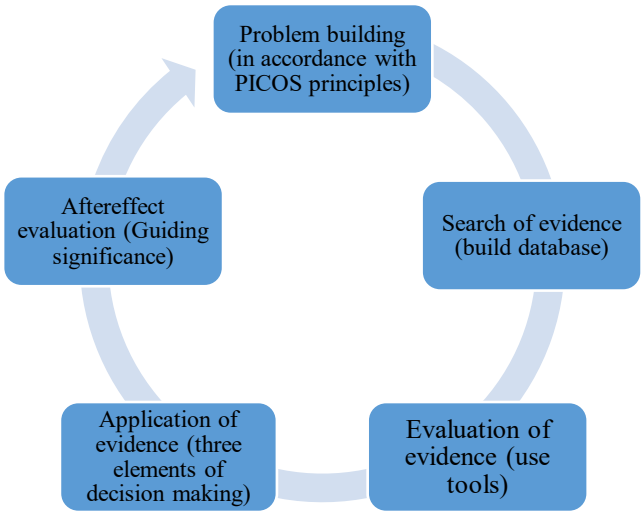


Figure 2-2. Five steps in evidence-based practice.

(b) Search method. The literature search procedure is determined according to the research requirements and the conditions of the subject, which are mainly divided into five steps. The specific process is shown in Figure 2-3. First, the search requirements are clarified. The search terms are determined. The search formula is written according to the search strategy. The professional search is carried out in the databases of CNKI, VIP and WanfangData. The relevant titles are obtained and exported in NoteExpress format. Then, we create a new database and folder in the NoteExpress software, and import the original document title into the new database of NoteExpress. Then, the establishment of the literature database is completed.

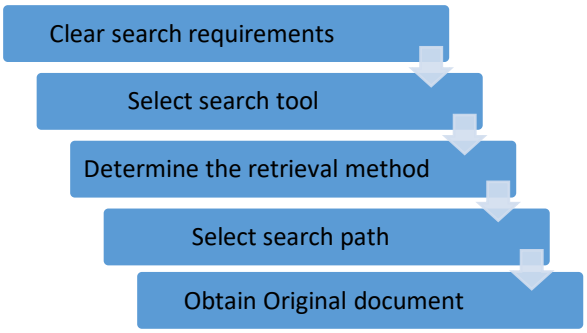


Figure 2-3. General procedure for document retrieval.

(c) Determination of the retrieval strategy. Based on the data mining related books and literatures, the principal component analysis, factor analysis, Bayesian network principal component analysis, factor analysis method, regression model, multiple regression, logistic

regression, naive Bayesian estimation, naive Bayesian classification, Bayesian network, genetic algorithm, regression analysis, cluster analysis, neural network, artificial neural network, decision tree, decision tree analysis, association analysis, association rule, association law, rough set theory, rough set method, fuzzy set method, integrated learning, support vector machine, correlation analysis, prediction, association mining, information entropy theory, business intelligence, manufacturing intelligence, digital decision making, business analysis and optimization classification, descriptive statistics, parameter test, non-parametric test, k-means, and dichotomy clustering, and other data mining methods are analyzed by statistical methods. The retrieval strategy in CNKI is shown in Table 2-4. The retrieval terms and strategy in WanfangData and VIP are consistent with those in CNKI, respectively.

(d) Specification of the mining method name. According to the data mining related books, the names of data mining methods are standardized and unified. The different expressions of the same data mining method are classified into one category, which is convenient for accurate statistics. For example, neural network is the abbreviation of artificial neural network. Thus, artificial neural network and neural network are unified into artificial neural network. Association analysis, association rule, association law, and association mining are different expressions of the same data mining method, so they are unified. The association analysis, naive Bayesian estimation, naive Bayesian classification, and the Bayesian network are unified into naive Bayesian.

(2) Analysis method

Firstly, the mathematical statistics method is used to classify the information in the literature database. The main contents are the characteristics of the included documents, the analysis of application trends, the distribution of application content and other contents distribution. The various types of information are presented in a separate form. The statistics and results are visualized and reproduced through Cytoscape software, bars and pie charts in Excel 2016.

(3) Research result

The literatures related to the application of data mining methods in the efficacy of TCM in the three major databases of electronic retrieval are statistically analyzed. The distribution of the top nine data mining methods in the total number of literatures is shown in Figure 2-4. It is found that the number of articles based on mining, correlation analysis, prediction, cluster analysis, frequency analysis, regression analysis, correspondence analysis, support vector machine, neural network and other mining method is the most. However, the number of documents using time series analysis, time data analysis, in time series mode analysis, shopping basket analysis, multi-scale method, business analysis and optimization classification is almost zero. The current data mining methods, such as correlation analysis, cluster analysis and regression analysis are more mature in this respect.

Table 2-4. Document retrieval strategy.

Database		**CNKI**
	#1	TCM property [subject]
	#2	TCM efficacy [subject]
	#3	TCM effectiveness [subject]
	#4	#1 OR #2 OR #3
	#5	Data mining [subject]
	#6	Principal component analysis [subject]
	#7	Factor analysis [subject]
	#8	Regression model [subject]
	#9	Multiple regression [subject]
	#10	Logistic regression [subject]
	#11	Naive Bayes estimates [subject]
	#12	Naive Bayes classification [subject]
	#13	Bayesian network [subject]]
	#14	Genetic algorithm [subject]
	#15	Regression analysis [subject]
	#16	Cluster analysis [subject]
	#17	Neural network [subject]
	#18	Artificial neural network [subject]
	#19	Decision tree [subject]
	#20	Association analysis [subject]
	#21	Association rule [subject]
	#22	Association law [subject]
	#23	Rough set theory [subject]
	#24	Integrated learning [subject]
	#25	Support vector machine [subject]
	#26	Related analysis [subject]
	#27	Prediction [subject]
	#28	Association mining[subject]
	#29	Information entropy theory [subject]
Search process	#30	Frequency analysis [subject]
	#30	Correspondence analysis [subject]
	#32	Shopping basket analysis [subject]
	#33	Multivariate scale method [subject]
	#34	Deviation analysis [subject]
	#35	Visualization technology [subject]
	#36	Evolution analysis [subject]
	#37	Complex system entropy clustering [subject]
	#39	Decision tree analysis [subject]
	#40	Rough set method [subject]
	#41	Fuzzy collection method [subject]
	#42	Business intelligence [subject]
	#43	Manufacturing intelligence [subject]
	#44	Digital decision [subject]
	#45	Business Analysis and Optimization Classification [subject]
	#46	Descriptive statistics [subject]
	#47	Parameter test [subject]
	#48	Nonparametric test [subject]
	#49	K-mean [subject]
	#50	Dichotomy clustering [subject]
	#51	Time series analysis [subject]
	#52	Time data analysis [subject]
	#53	Timing pattern analysis [subject]
	#54	Anomaly analysis [subject]
	#55	Abnormal point analysis [subject]
	#56	Specific group analysis [subject]
	#57	Specific group mining [subject]
	#58	Web data mining [subject]
	#59	#5 OR #6 OR #7 OR #8 OR #9 OR #10 OR…OR#53 OR#54
	#60	#4 AND #59

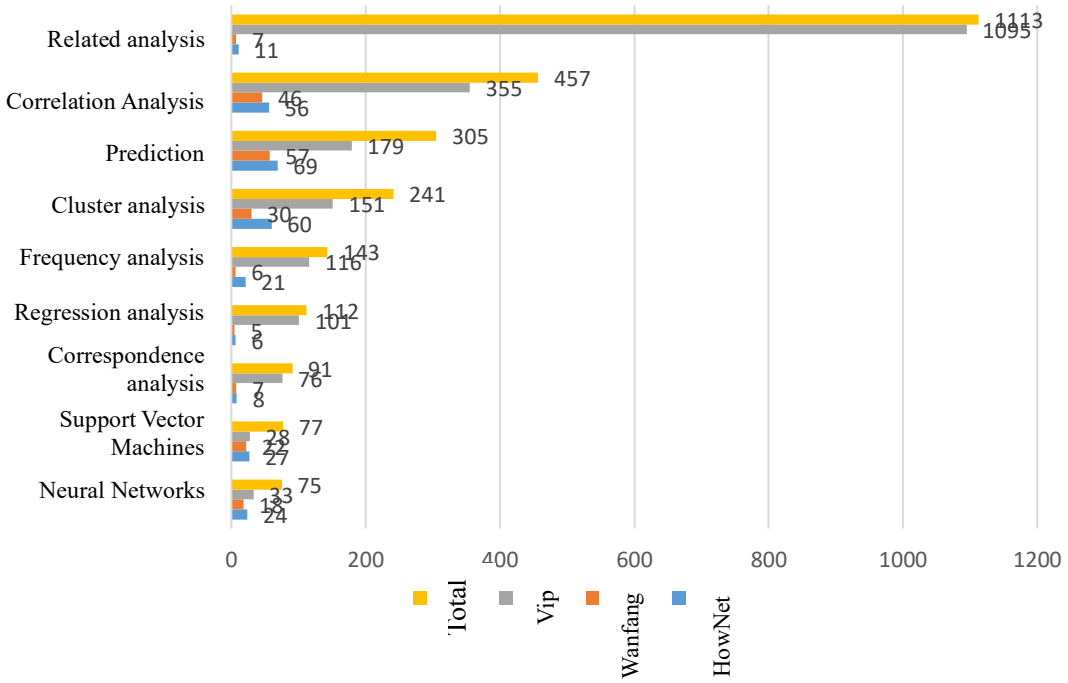


Figure 2-4. The distribution of the top nine data mining methods in the total number of literatures.

(a) Characteristics of the included literature. We Check the literature database in NoteExpress software, and delete 226 duplicate documents. 1813 related documents remain. The various types of information on the 1813 articles are statistically analyzed, and the results are visually reproduced. The statistical results are as follows. The number of data mining methods in the database is shown in Table 2-5. The number of association rule and cluster analysis reaches 507 and 213, respectively, which accounts for a large proportion in the total number. The method is usually used for the analysis and research of TCM prescription, such as the core drug pair or core drug combination mining and classifying the efficacy of the prescription. Principal component analysis has 72 applications. This cannot only independently mine TCM information, but also often use data mining methods, such as support vector machine and neural network. In addition, most data mining methods in the table have been applied by researchers. For example, the data mining method is quite mature in the application of drug discrimination and has been recognized by a large number of drug researchers.

The number and percentage of literatures of each data mining methods are shown in Figure 2-5 to 2-6, respectively. Among them, 507 literatures are about association rule, accounting for 42%; 213 literatures are about cluster analysis, accounting for 17.65%; 72 literatures are about principal component analysis, accounting for 5.97%; 69 literatures are about neural network, accounting for 5.72%; 61 literatures are about support vector machine, accounting for 5.05%; 52 literatures are about frequency analysis, accounting for 4.31%; 48 literatures are about factor analysis, accounting for 3.98%; 42 literatures are about Bayesian network, accounting for 3.48%; 35 literatures are about partial least square, accounting for 2.90%; 34 literatures are about decision tree, accounting for 2.82%; 28 literatures are about corresponding analysis, accounting

for 2.32%; 22 literatures are about regression model, accounting for 1.82%; 10 literatures are about linear discrimination, accounting for 0.83%; 9 literatures are about rough set, accounting for 0.75%; 5 literatures are about random forest, accounting for 0.41%.

Table 2-5. The number of literatures related to each data mining method.

Data mining method	Number of literatures
Association rule	507
Cluster analysis	213
Principal component analysis	72
Neural network	69
Support vector machine	61
Frequency analysis	52
Factor analysis	48
Bayesian network	42
Partial least squares	35
Decision tree	34
Correspondence analysis	28
Regression model	22
Linear discrimination	10
Rough set	9
Random forest	5

Statistical analysis is done for the included literatures. As shown in Table 2-6 and Figure 2-7, the literatures are classified into four types of the journal paper, dissertation, conference paper, and newspaper article. Among them, the number of journal papers is 1684, accounting for 92.86%; the number of dissertations is 98, accounting for 5.41%; the number of conference papers is 30, accounting for 1.66%; the number of articles is one, accounting for 0.06%. The research in this area is mainly reported through journal papers. The number of dissertations obtained by the retrieval is relatively small. The reason is that a large number of dissertations has not been officially disclosed. With the increase in the number of dissertations, the research result will be more reliable. The publication of dissertations will not only facilitate the further improvement of the quality of this research, but also facilitate the development of various disciplines and provide reference for researchers.

According to the statistics of the journals for the retrieval literatures, the number and percentage of journals is shown in Table 2-7. We visualize the number of journals, and results are shown in Figure 2-8. Among them, “Chinese Journal of Traditional Chinese Medicine” contains 106 papers, accounting for 5.85%; “China Experimental Formulation Journal” contains 76 papers, accounting for 4.19%; “World Science and Technology: Modernization of Traditional Chinese Medicine” contains 58 papers, accounting for 3.20%; “Chinese Medicine” contains 43 papers, accounting for 2.37%; “Chinese herbal medicine” contains 40 papers, accounting for 2.21%; “China Journal of Traditional Chinese Medicine” contains 32 papers, accounting for 2.04%; “Liaoning Journal of Traditional Chinese Medicine” contains 32 papers, accounting for 1.77%; “Journal of Shandong University of Traditional Chinese Medicine” contains 24 papers, accounting for 1.32%; “World Traditional Chinese Medicine” contains 20 papers, accounting for 1.1%; “Chinese Pharmacovigilance” contains 18 papers, accounting for 0.99%; “Journal of Crops” contains 17 papers, accounting for 0.94%; both “Chinese Agricultural Science” and “Journal of Liaoning University of Traditional Chinese Medicine”

contain 16 papers, each accounting for 0.88%; “pharmacology and clinical practice of Chinese medicine” contains 16 papers, accounting for 0.83%. The above-mentioned journals are among the top in the number of literatures on the exploitation of Traditional Chinese Medicine. In addition, there are more than 10 papers in the journals such as “Chinese Journal of Traditional Chinese Medicine”, “Chinese Pharmacy”, “Chinese Pharmacy” and “Chinese Herbal Medicine”. These journals are in the list of “2017 Chinese core journals”. Thus, the research result is reliable, and the literature quality is high. The data mining has a high degree of research recognition and attention in the efficacy of TCM.

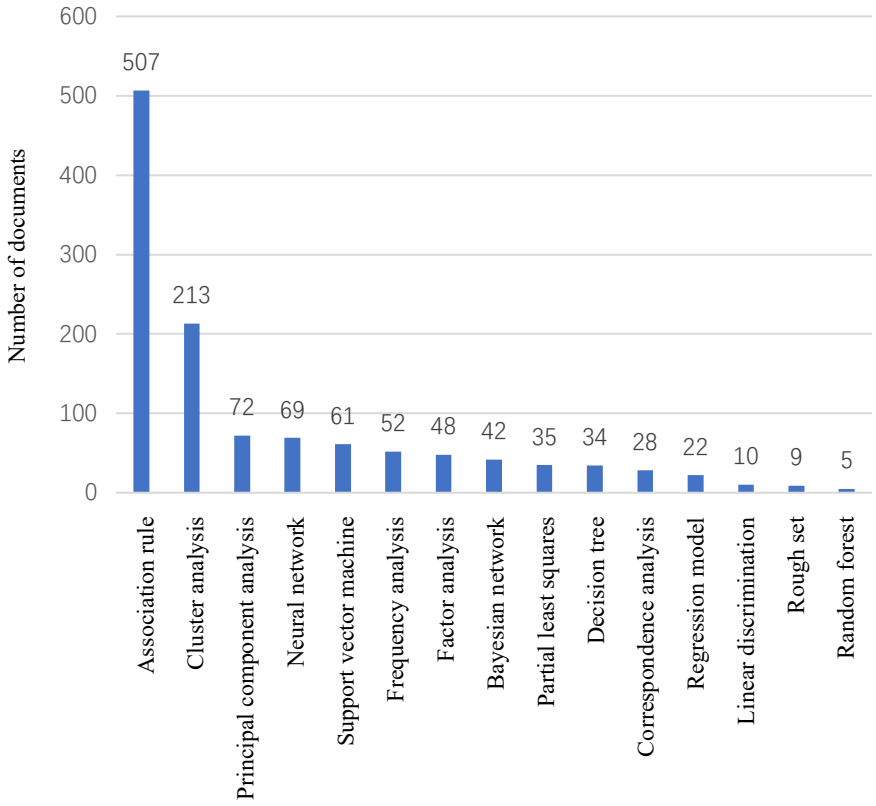


Figure 2-5. The number of literatures related to each data mining method.

We have done the statistical analysis of the dissertation in the literature database. The statistical results are shown in Table 2-8. The number of dissertations retrieved is 98. Among them, 14 is from Beijing University of Traditional Chinese Medicine, accounting for 14.29%; 13 is from Shandong University of Traditional Chinese Medicine, accounting for 13.27%; 10 is from China Academy of Traditional Chinese Medicine, accounting for 10.20%; 6 is from Shanghai University of Traditional Chinese Medicine or Nanjing University of Traditional Chinese Medicine, each accounting for 6.12%; 5 is from Chengdu University of Traditional Chinese, accounting for 5.10%; 4 is from Shandong University, accounting for 4.08%; both Fudan University and Guangzhou University of Traditional Chinese Medicine have 3 dissertations, each accounting for 3.06%; 2 is from each of Chinese Academy of Military Medical Sciences, Shanxi University, Guangxi University of Traditional Chinese Medicine, Fuzhou University, Fujian University of Traditional Chinese Medicine, and Liaoning

University of Traditional Chinese Medicine, each accounting for 2.04%; 1 is from each of 21 universities including Central South University for Nationalities, People's Liberation Army, Academy of Medical Sciences, China Union Medical University, Lanzhou University, Dalian Medical University, Tianjin University, *et al*, each accounting for 1.02%. The universities of Traditional Chinese medicine that publish their dissertations include Beijing University of Traditional Chinese Medicine, Shandong University of Traditional Chinese Medicine, Shanghai University of Traditional Chinese Medicine, Nanjing University of Traditional Chinese Medicine, Chengdu University of Traditional Chinese Medicine, Guangzhou University of Traditional Chinese Medicine, and Guangxi University of Traditional Chinese Medicine, Fujian University of Traditional Chinese Medicine, Liaoning University of Traditional Chinese Medicine, Zhejiang University of Traditional Chinese Medicine, and Hubei University of Traditional Chinese Medicine. The number of these university is 11, which accounts for close to 1/2 of the total number of universities of Traditional Chinese medicine in China (24 in total). The number of dissertations related to data mining method applied to Traditional Chinese Medicine efficacy published by these universities is 55, accounting for 56.12% of the total number of dissertations from these universities in the database. Moreover, these statistical dissertations do not include some undisclosed dissertations. Thus, most universities of Traditional Chinese medicine have research involving data mining in Traditional Chinese Medicine.

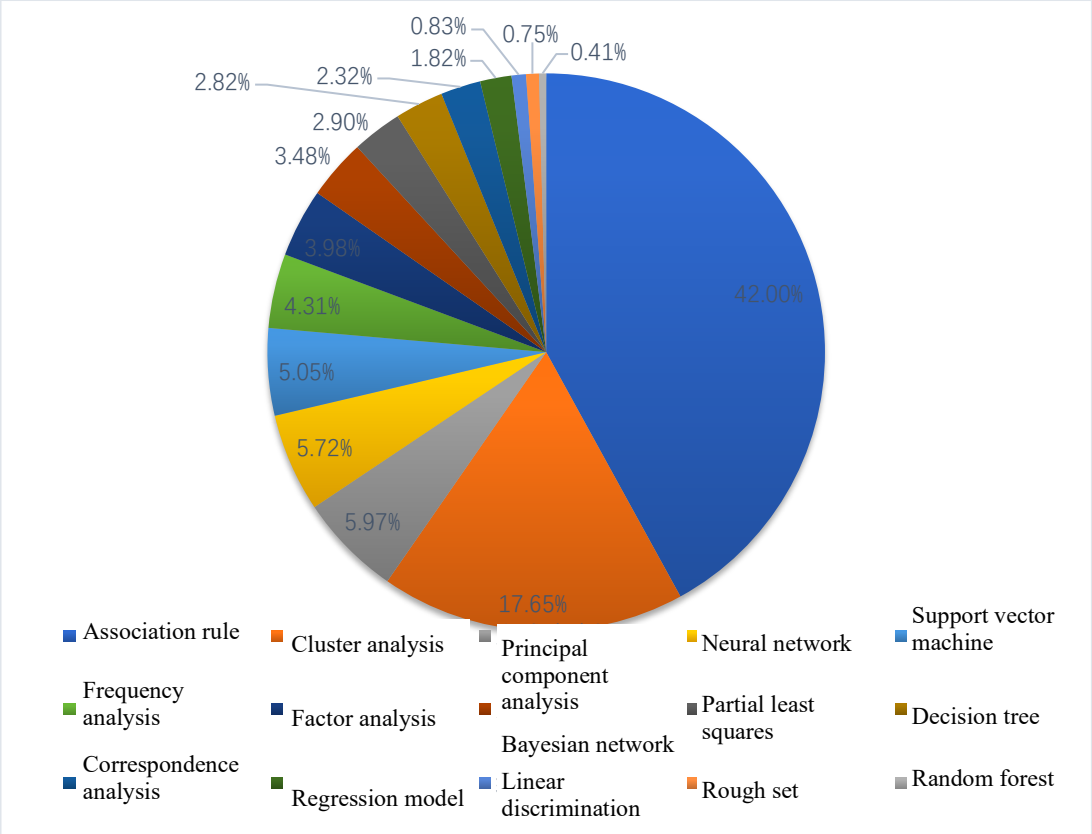


Figure 2-6. The percentage of literatures related to each data mining method.

Table 2-6. The number and percentage of literatures of each type.

Literature type	Number	Percentage (%)
Journal paper	1684	92.86
Dissertation	98	5.41
Conference paper	30	1.66
Newspaper article	1	0.06

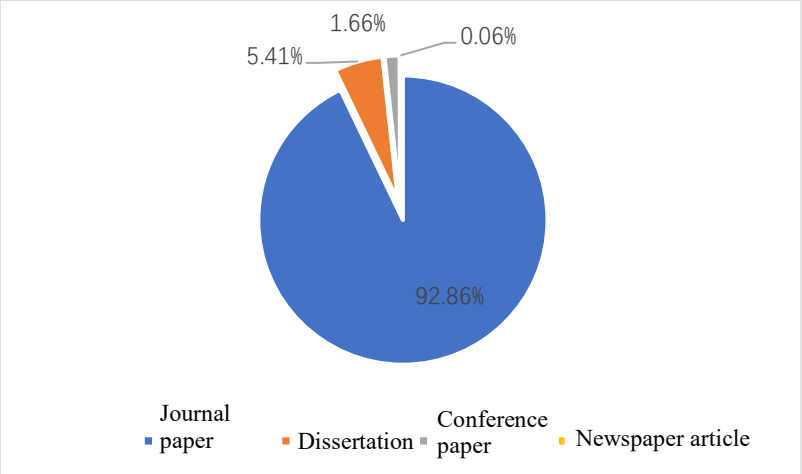


Figure 2-7. The percentage of literatures of each type.

Table 2-7. The number and percentage of published literatures about the application of data mining in the prediction of the TCM efficacy for journals.

No.	Journal	Number	Percentage (%)
1	Chinese Journal of Traditional Chinese Medicine	106	5.85
2	Chinese Journal of Experimental Formulaology	76	4.19
3	World Science and Technology: Modernization of Traditional Chinese Medicine	58	3.20
4	Chinese Journal of Traditional Chinese Medicine	43	2.37
5	Chinese herbal medicine	40	2.21
6	China Journal of Traditional Chinese Medicine and Information	37	2.04
7	Liaoning Journal of Traditional Chinese Medicine	32	1.77
8	Shizhen Chinese Medicine Journal	30	1.66
9	Journal of Shandong University of Traditional Chinese Medicine	24	1.32
10	World medicine	20	1.10
11	Chinese Pharmacovigilance	18	0.99
12	Crop journal	17	0.94
13	Chinese Agricultural Science	16	0.88
14	Journal of Liaoning University of Traditional Chinese Medicine	16	0.88
15	Chinese medicine pharmacology and clinical	15	0.83
16	Chinese Medicine Herald	14	0.77
17	Chinese Medicine Journal	13	0.72
18	Chinese Journal of Traditional Chinese Medicine	13	0.72
19	Chinese clinical rehabilitation	13	0.72
20	Chinese pharmacy	13	0.72
21	Journal of Traditional Chinese Medicine	12	0.66
22	Chinese herbal medicine	12	0.66
23	Chinese modern Chinese medicine	11	0.61
24	Chinese Journal of Basic Medicine in Traditional Chinese Medicine	10	0.55
25	Chinese Rice Science	10	0.55
26	Asia Pacific Traditional Medicine	10	0.55
27	Inner Mongolia Traditional Chinese Medicine	10	0.55
28	International Journal of Traditional Chinese Medicine	10	0.55
29	Trace elements and health research	10	0.55

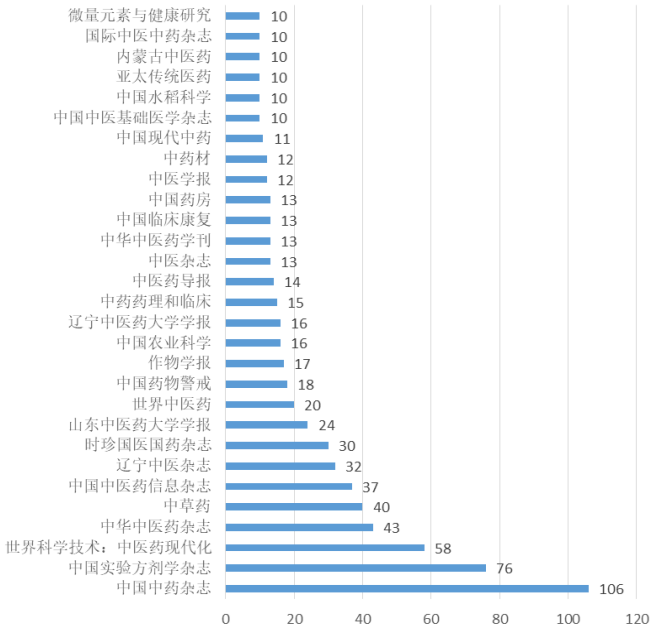


Figure 2-8. The number of published literatures about the application of data mining in the prediction of the TCM efficacy for journals.

(b) Application trend analysis. Through statistical analysis, the number of literatures published from 2000 to 2018 in the application of data mining in the prediction of the TCM efficacy is shown in Figure 2-9. From 2000 to 2004, the number of related literatures is less than 10. In 2014, it reaches 100. In 2017, it reaches 200. It increases from 0 in 2000 to 229 in 2018. The number of literatures related to data mining in the prediction of TCM efficacy is increasing year by year. The number of published literatures after 2015 rapidly increases. Researchers are paying more and more attention to the application of data mining methods. Modern research methods are used to conduct in-depth research on the TCM efficacy.

(c) Application content distribution. According to the inclusion criteria, the literatures are screened and classified, the data mining method and its applied research content are sorted and summarized. Then, the network diagram is drawn by Cytoscape software to visualize the relationship between the data mining method and the mining content efficacy. The application distribution of current mining methods in predicting TCM efficacy is obtained, which is shown in Figure 2-10. The data mining methods commonly used in the prediction of efficacy of single TCM are neural network, principal component analysis, decision tree, least squares method, Bayesian network, support vector machine and the correspondence analysis. Data mining methods commonly used in the prediction of TCM compound efficacy are high-dimensional data reduction, neural network, rough set, support vector machine, space vector. Data mining methods commonly used in discrimination of cold-hot drug properties are artificial neural network, high-dimensional data reduction, linear discrimina analysis, support vector machine, principal component analysis, Fisher discrimina analysis, Bayesian network, discrimina analysis, regression tree, decision tree. In addition, data mining methods for discriminating flatness or non-flatness of TCM are principal component analysis and support vector machine. The data mining method for predicting five-flavors and efficacy of TCM components is

Bayesian network. The data mining method for predicting property and channel tropism of TCM is decision tree. The data mining method for neural immune prediction is Logistic regression.

Table 2-8. The number and percentage of dissertations of universities.

No.	University	Number	Percentage (%)
1	Beijing University of Chinese Medicine	14	14.29
2	Shandong University of Traditional Chinese Medicine	13	13.27
3	China Academy of Chinese Medical Sciences	10	10.20
4	Shanghai University of Traditional Chinese Medicine	6	6.12
5	Nanjing University of Chinese Medicine	6	6.12
6	Chengdu University of Traditional Chinese Medicine	5	5.10
7	Shan Dong University	4	4.08
8	Fudan University	3	3.06
9	Guangzhou University of Chinese Medicine	3	3.06
10	Chinese Academy of Military Medical Sciences	2	2.04
11	Shanxi University	2	2.04
12	Guangxi University of Traditional Chinese Medicine	2	2.04
13	Fuzhou University	2	2.04
14	Fujian University of Traditional Chinese Medicine	2	2.04
15	Liaoning University of Traditional Chinese Medicine	2	2.04
16	South Central University for Nationalities	1	1.02
17	Chinese Academy of Military Medical Sciences	1	1.02
18	Peking Union Medical College	1	1.02
19	Lanzhou University	1	1.02
20	Beijing Jiaotong University	1	1.02
21	Beijing University of Chemical Technology	1	1.02
22	Peking Union Medical College	1	1.02
23	Northeastern Polytechnic University	1	1.02
24	South China University of Technology	1	1.02
25	Nanjing University of Science and Technology	1	1.02
26	Nanchang University	1	1.02
27	Harbin Engineering University	1	1.02
28	Dalian Medical University	1	1.02
29	Tianjin University	1	1.02
30	Guangxi Medical University	1	1.02
31	Ji Nan University	1	1.02
32	Zhejiang University of Traditional Chinese Medicine	1	1.02
33	Hubei University of Traditional Chinese Medicine	1	1.02
34	Northwest University	1	1.02
35	Guizhou University	1	1.02
36	Liaoning Medical College	1	1.02

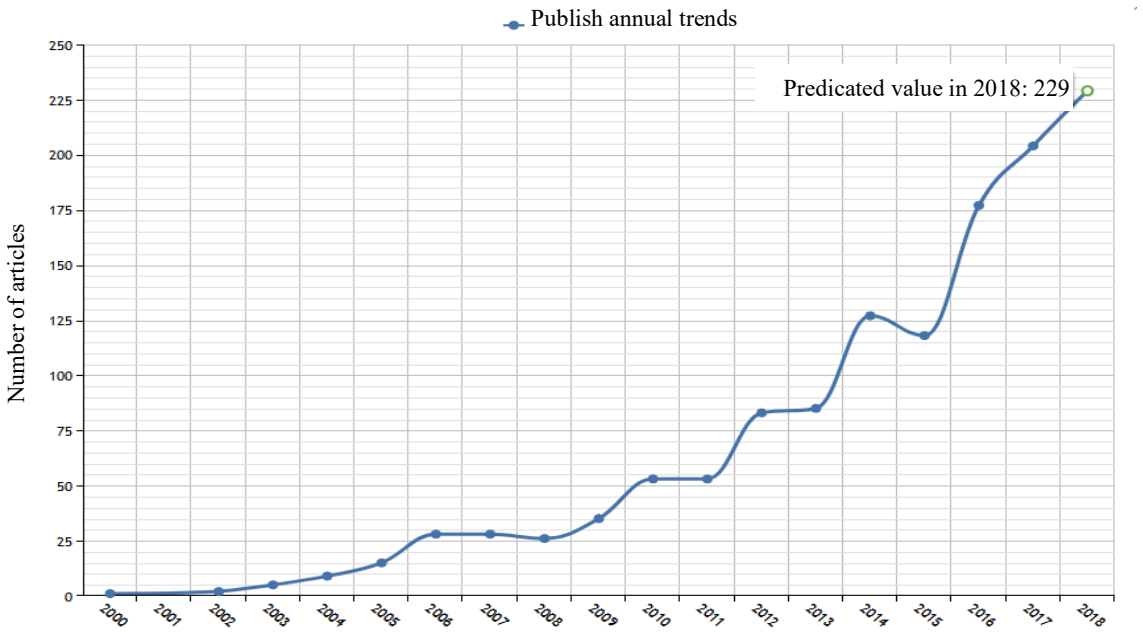


Figure 2-9. The annual trend of number of literatures related to the data mining method on the TCM efficacy.

According to the further analysis of the compiled data, the data mining content of TCM efficacy and the number of its corresponding data mining methods are shown in Table 2-9. The number of data mining methods in TCM efficacy is visualized. The results are shown in Figure 2-10. This can reflect the hot issues of data mining in the study of different aspects of TCM efficacy and the content to be further studied.

Table 2-9. The data mining content of TCM efficacy and the number of its corresponding data mining methods.

Date mining content	Number of corresponding data mining methods
Cold and hot drug property	15
Single TCM efficacy	8
Compound efficacy	5
Efficacy relationship	2
Flat and non-flat drug property	2
Five-flavor prediction	1
Component efficacy prediction	1
Natural taste and channel tropism prediction	1
Component drug prediction	1
Pharmacodynamic index prediction	1
Neuroimmuno prediction	1
Neural effect prediction	1

(d) Distribution of other applications. In the process of collating the literature, the data mining software or platforms applied in the literature are counted. The data mining software and platforms are R software, IBM SPSS Clementine statistical software, MATLAB, SAS, Clementine data mining, PAST statistical software, SpecAlign, Hugin Reseacher, Enterprise

Miner (EM), *etc*. Some data mining application software and its basic functions are shown in Table 2-10. The network relationship diagram between TCM efficacy and data mining method is shown in Figure 2-11.

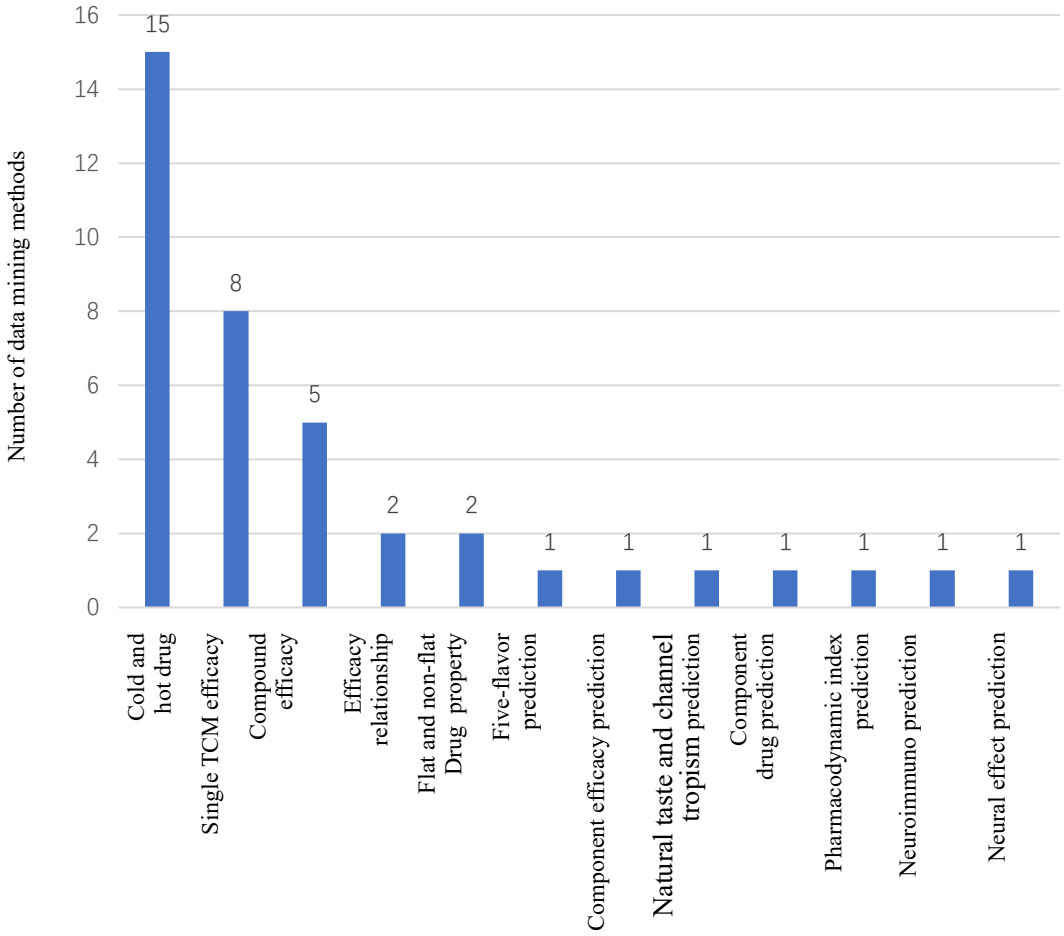


Figure 2-10. The data mining content of TCM efficacy and the number of its corresponding data mining methods.

Table 2-10. The data mining software of TCM efficacy and its basic functions.

Software	Function
R software	Data processing, calculation and mapping, *etc*.
IBM SPSS	Data management, statistical analysis, chart analysis, output management, *etc*.
MATLAB	Numerical analysis, digital image processing technology, algorithm development, data visualization, *etc*.
SAS	Data access, data storage and management, application development, graphics processing, data analysis, prediction, *etc*.
Clementine	Handling all kinds of different types of data. It is the best and most extensive data mining technology available.
PAST	Data management, statistical analysis, mapping and modeling, *etc*.

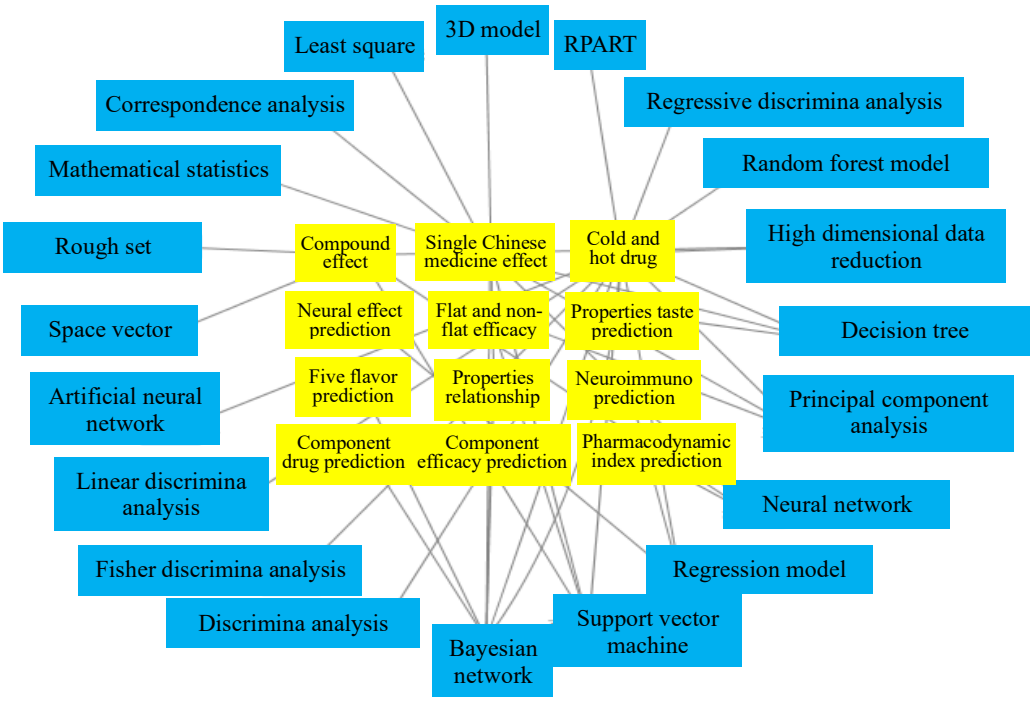


Figure 2-11. The network relationship diagram between TCM efficacy and data mining method.

2.2.3 Systematic evaluation of cold-hot drug nature discrimination for TCM

At present, there are many types of data mining methods for the identification of cold-hot drug properties of TCM, and the study of cold-hot drug discrimination. The cold-hot drug discrimination model is built based on the basic information of TCM, and the model is tested by some known information. Thus, there will be data such as the total quantity of predication, the correct number of predication and the accuracy rate of predication. It is feasible to apply the evidence-based method to the evaluation of the data mining method in TCM efficacy. Then, the experimental research data is further utilized and presented in the form of knowledge.

(1) Literature inclusion criteria

According to the research requirement, the inclusion criteria are as follows:

(a) The language of the incorporated literature is limited to Chinese;

(b) The training set and the test set are randomly grouped. The cold-hot drug nature is clear, and the amount of cold drug is the same as that of the hot drug;

(c) The research object is the TCM efficacy;

(d) The prediction is based on the chemical composition of TCM or other TCM information. This can be used as a model for establishing pharmacological discrimination, such as pharmacological indicators, natural taste, and material basis.

(2) Research object

The research objects are the application of data mining method in the discrimination of cold-hot drug nature of TCM from the database construction of CNKI, VIP and WanfangData to September 9, 2018, and the comparative study of multiple data mining methods in drug nature discrimination.

(3) Type of research

The evidence-based method is used to evaluate the data mining method of cold-hot drug nature of TCM. Thus, the research type is selected as "randomized controlled trial". The research object is the cold-hot drug nature of TCM. The intervention measures are data mining methods. The two data mining methods are used to compare the cold-hot drug natures of TCM. The design pattern of the drug property discrimination model is shown in Figure 2-11.

(4) Outcome indicators of the evaluation

The discrimination model of cold-hot drug property of TCM is built based on the basic information of TCM, such as property and flavor, channel tropism, pharmacological indicators, chemical composition, *etc*. Then, some data information is input as the test set to test the model. The data, such as predication accuracy rate of drug property discrimination and the accurate number of predictions, can be obtained. These data can reflect the application effect of the data mining method. Therefore, we select the discriminative accuracy rate of the test set as the main outcome index to evaluate the data mining method of cold-hot drug property discrimination.

(5) Document exclusion criteria

According to the research requirement, the following exclusion criteria are listed:

(a) Research that is clearly unrelated to the content of the drug discrimination;

(b) Other studies have a neurological effect and whether there is a prediction of immunomodulatory effects or a comparative study;

(c) Literature review articles;

(d) Repetitive publications, incomplete data records, unclear sources of information or inconsistent with actual research;

(e) The literature of the test set sample size is less than or equal to 20;

(f) The number of authors using the same two data mining methods is less than 2, and no comparative research content can be formed.

(6) Literature retrieval strategy

On the basis of the existing research, some data mining methods that are obviously not suitable for TCM efficacy mining are excluded. 33 kinds of data mining methods, such as principal component analysis, factor analysis, Bayesian network, as well as TCM efficacy, drug property of TCM, TCMs effectiveness are used as the search terms for the literature retrieval of in the CNKI, VIP, and WanFangData databases. The retrieval time ranges from the establishment of the database to September 9, 2017, and the literature unrelated to TCM is excluded. The retrieval formula in the three databases and the number of literatures retrieved are shown in Table 2-11.

(7) System evaluation method

(a) Data extraction. According to the research content of the subject, the data extraction form is designed by Excle2016. The extraction contents include publication journal, publication year, first author, mining method, mining type, the total amount of sample, training set vs test set, modeling basis, software, prediction the number of data, accurate predication of the number of data, accuracy rate of predication and the remarks. Before extracting data, we must check the literature, delete the duplicate literatures. Then, we read the abstracts and the full text to exclude the literature that is unrelated to the research requirement. This improves the efficiency of extracting data. The design block diagram of the drug property discrimination model is

shown in Figure 2-12.

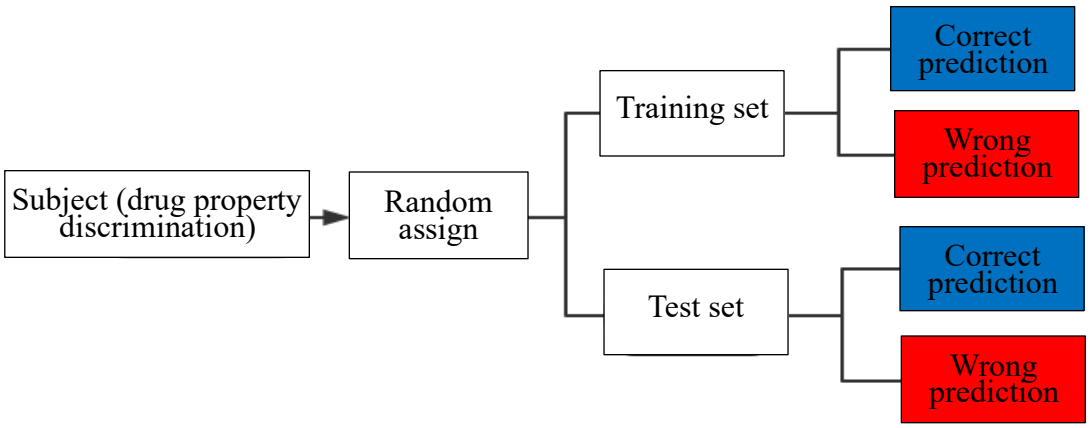


Figure 2-12. The design block diagram of the drug property discrimination model.

(b) Quality evaluation. Good quality evidence is the guarantee for doing a second study. This study mainly evaluates the quality of information for the literature included in the study. The content of the evaluation includes whether the drug property is clear; whether the modeling basis is the same; whether the selection of the outcome indicators is the same; whether the identical data mining method for comparison has been used by two authors.

(c) Statistical methods. Meta-analysis is performed using Stata15.0. The variable uses the odd ratio (OR) value as the effect index. Each effect quantity provides its point estimation Z value and 95% CI confidence interval. Chi-square test is used for heterogeneity test in each study. $P<0.1$ is used as the test level. If $P>0.1$, $I^2\leq 50\%$, the heterogeneity between the studies is small. The fixed effect model is used for meta-analysis. If $P\leq 0.1$ and $I^2>50\%$, the heterogeneity between the studies is large, the random effects model is used for meta-analysis. Meta-analysis is not suitable for heterogeneous model tests. Sub-group analysis or sensitivity analysis is usually used to explore the source of heterogeneity, or only descriptive analysis.

(8) Software optimization

The analysis software used in this study is Stata, which is a powerful tool. It was released in 1985 and the latest version is Stata15.0. After continuous updating and expansion, software analysis functions are becoming more and more perfect. It is characterized by allowing users to modify and add program files according to their own research requirements. Therefore, many programmers have written many advanced statistical modules including Meta analysis. This is not only powerful, but also beautiful. Since Stata15.0 does not have a Meta Analysis Module, it requires to be installed b users. We click on the “Help-Search” in toolbar, select “search all”, enter the “meta dialog” as the keywords in the pop-up dialog box, and click ok. The new interface appears. Then, we click on the page link “pr0012 from http://www.stata-Journal.com/software/sj4-2”, and click “click here to install” in the new interface. The installation is complete with the appearance of “installation is complete”. Then, we create a folder profile.do and save it to the c:ado\plus directory. We enter the “run c:\ado\plus\profile.do” command in the command box to enable it to run automatically when you start the Stata software. The Meta-analysis module, the Stata 15.0 toolbar and the Meta-analysis module

installation interface will appear in the Stata15.0 Toolbar User, as shown in Figure 2-13. Finally, we install Meta-analysis commands for two-category data in Stata15.0:

Ssc install metan;

Ssc install meta;

Ssc install metaan;

Table 2-11. Document retrieval strategy.

Database	Retrieval strategy
CNKI: 225	SU = (data mining + principal component analysis + factor analysis + regression model + multiple regression + logistic regression + naive Bayesian estimation + naive Bayesian classification + Bayesian network + genetic algorithm + regression analysis + clustering analysis + neural network + artificial neural network + decision tree + association analysis + association rule + association law + rough set theory + integrated learning + support vector machine + correlation analysis + prediction + association mining + information entropy theory + frequency analysis + correspondence analysis + shopping basket analysis + multi-scale method + deviation analysis + visualization technology + evolution analysis + complex system entropy clustering) * (Traditional Chinese medicine effectiveness + Traditional Chinese medicine efficacy + drug property of Traditional Chinese medicine)
VIP: 1677	U = data mining OR U = principal component analysis OR U = factor analysis OR U = regression model OR U = multiple regression OR U = logistic regression OR U = naive Bayesian estimation OR U = naive Bayesian classification OR U = Bayesian network OR U = genetic algorithm OR U = regression analysis OR U = cluster analysis OR U = neural network OR U = artificial neural network OR U = decision tree OR U = association analysis OR U = association rule OR U = association law OR U = rough set theory OR U = integrated learning OR U = support vector machine OR U = correlation analysis OR U = prediction OR U = association mining OR U = information entropy theory OR U = frequency analysis OR U = correspondence analysis OR U = shopping basket analysis OR U = multi-scale method OR U = deviation analysis OR U = visualization technology OR U = evolution analysis OR U = complex system entropy clustering) AND (U = Traditional Chinese medicine efficacy OR U = Traditional Chinese medicine effectiveness OR U = drug property of Traditional Chinese medicine)
WanFang: 162	SU: (“data mining” + “principal component analysis” + “factor analysis” + “regression model” + “multiple regression” + “logistic regression” + “simple Bayesian estimation” + “simple Bayesian classification method” + “Bayesian network” + “genetic algorithm” + “regression analysis” + “cluster analysis” + “neural network” + “artificial neural network” + “decision tree” + “association analysis” + “association rule” + “association rule” + “rough set theory” + “integrated learning” + “support vector machine”+ “correlation analysis” + “prediction” + “association mining” + “information entropy theory” + “frequency analysis” + “correspondence analysis” + “shopping basket analysis” + “multi-scale method” + “deviation analysis” + “visualization technology” + “evolution analysis” + “complex system entropy clustering” * Subject: (“Traditional Chinese medicine efficacy” + “Traditional Chinese medicine effectiveness” + “drug property of Traditional Chinese medicine”)

(9) System evaluation results

(a) Literature retrieval results. The number of literatures from CNKI, WanfangData and VIP are 225, 162 and 1677, respectively. The screening process and results of literatures retrieval are shown in Figure 2-13.

(b) Characteristics of literature research. The literature extraction information of the discrimination and classification of TCM cold-hot drug property is statistically compiled in Table 2-12. Because of insufficient document width, information such as TCM sources and mining software is omitted. To avoid excessive exclusion of research-related literatures, the

screening of the literature is carried out independently by two people. When their opinions are different, the discussion is decided by both parties. When the discussion cannot be unified, the third-party judges the result. The basic characteristics of the literature are shown in Table 2-12. The data mining methods used in the discrimination of cold-hot drug property include support vector machine, regression discriminant analysis, Fisher discriminant method, partial least squares discriminant analysis, RPART, Bayesian network, *etc*. It reflects that some data mining methods are not good enough, and the related problems are limited by certain factors. The modeling basis of cold-hot drug property discrimination are chemical composition, material composition, biological effect index, mass-to-charge ratio variable, natural taste, channel tropism, efficacy, indications, *etc*. The establishment and verification of these models prove once again that the above factors are closely related to drug property of TCM.

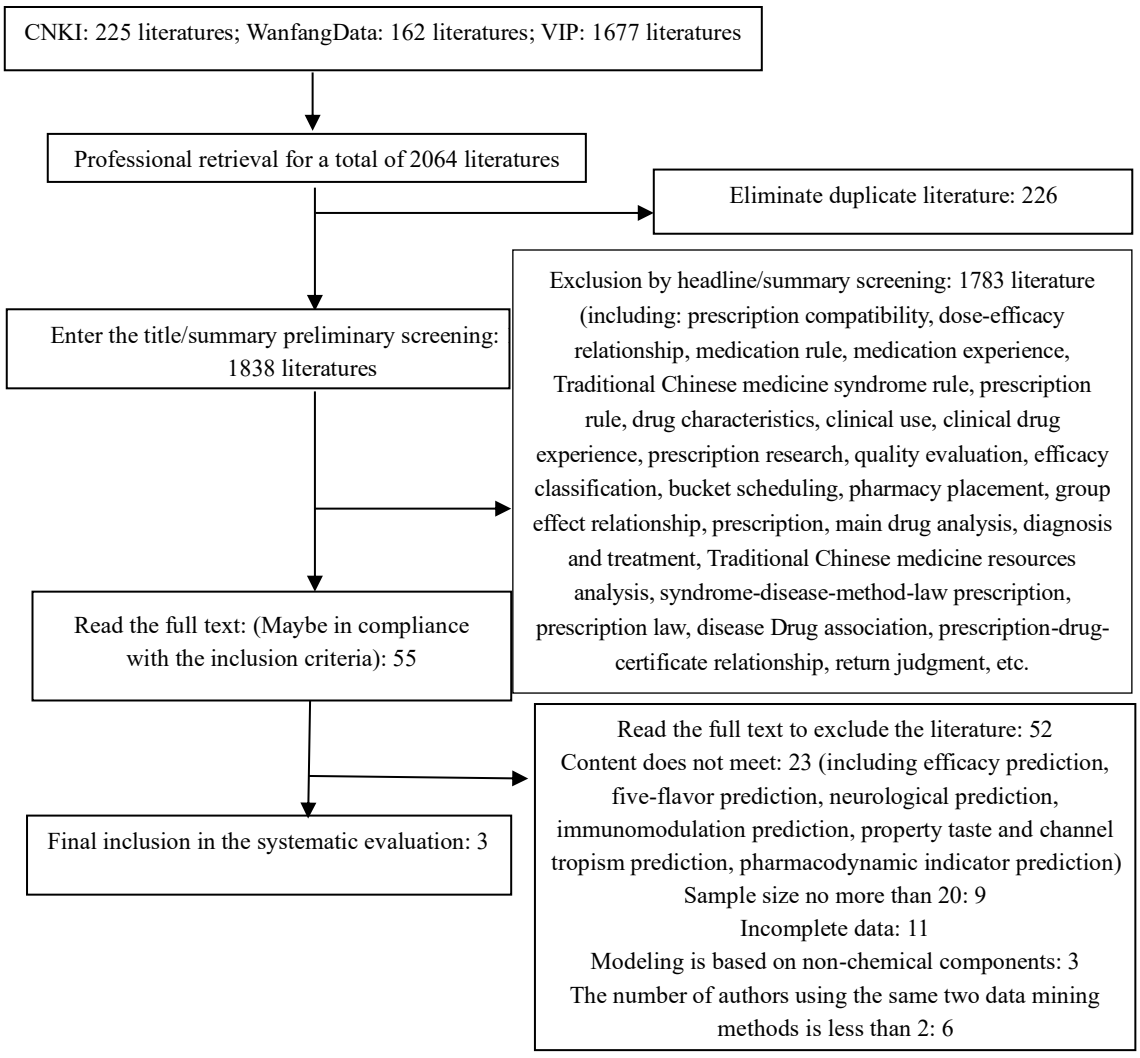


Figure 2-13. The screening process and results of literatures retrieval.

Table 2-12. Literatures characteristics of cold-hot drug property discrimination of TCM.

Year	Author	Mining Content	Method	Type	Amount	Train vs Test	Model basis	Predication amount	Accurate No.	Accuracy rate
2015	Wang Xiaoyan	Drug property discrimination	Support Vector Machines	Single drug	60	48vs12)*4	chemical composition	120	67	58.30%
2015	Wang Xiaoyan	Drug property discrimination	Regressive discriminant analysis	Single drug	60	48vs12)*4	chemical composition	120	48	40.00%
2015	Wang Xiaoyan	Drug property discrimination	Principal component linear discrimination analysis	Single drug	60	48vs12)*4	chemical composition	120	53	44.30%
2015	Wang Xiaoyan	Drug property discrimination	Partial least squares discriminant analysis	Single drug	60	48vs12)*4	chemical composition	120	62	51.80%
2015	Wang Xiaoyan	Drug property discrimination	Random forest model	Single drug	60	48vs12)*4	chemical composition	120	38	31.30%
2014	Wang Peng	Drug property discrimination	Linear discriminant analysis	Single drug	60	48vs12)*4	chemical composition	48	29	60.15%
2014	Wang Peng	Drug property discrimination	Regressive discriminant analysis	Single drug	60	48vs12)*4	chemical composition	48	20	41.68%
2014	Wang Peng	Drug property discrimination	Principal component linear discrimination analysis	Single drug	60	48vs12)*4	chemical composition	48	22	45.83%
2014	Wang Peng	Drug property discrimination	Partial least squares discriminant analysis	Single drug	60	48vs12)*4	chemical composition	48	23	47.90%
2014	Wang Peng	Drug property discrimination	Random forest model	Single drug	60	48vs12)*4	chemical composition	48	25	51.25%
2014	Wang Peng	Drug property discrimination	Support Vector Machines	Single drug	60	48vs12)*4	chemical composition	48	24	50.00%
2011	Long Wei	Drug property discrimination	Linear discriminant analysis	Single drug	284	284vs284	chemical composition	284	206	72.40%
2011	Long Wei	Drug property discrimination	Support Vector Machines	Single drug	284	284vs284	chemical composition	284	230	81.00%
2010	Wang Wei	Drug property discrimination	Fisher discriminant	Single drug	20	None	chemical composition	20	19	95.00%
2012	Wang Xiaoyan	Drug property discrimination	Support Vector Machines	Single drug	20	20vs20	chemical composition	20	20	100.00%
2010	Zhou Zhengli	Drug property discrimination	Fisher discriminant	Single drug	20	20vs20	chemical composition	20	20	100.00%
2010	Li Jingwen	Drug property discrimination	Fisher discriminant	Single drug	20	20vs20	chemical composition	20	20	100.00%
2010	Zhou Zhengli	Drug property discrimination	Fisher discriminant	Single drug	20	20vs20	chemical composition	20	20	100.00%
2011	Qifang	Drug property discrimination	Partial Least Squares Discriminant Analysis/Bayesian Network Model	Single drug	61	49vs12	Material composition	12	12	100.00%

Table 2-12. *Cont.*

Year	Author	Mining Content	Method	Type	Amount	Train vs Test	Model basis	Predication amount	Accurate No.	Accuracy rate
2014	Huang Liping	Drug property discrimination	Decision tree	Single drug	12	None	Biological effect indicator	12	12	97.39%
2011	Li Yu	Drug property discrimination	Logistic regression/BP neural network	Single drug	1728	415vs276	Attribute characteristics	691	494	71.49%
2016	Hu Yanan	Drug property prediction	Decision tree	Single drug	7	None	Pharmacological index	7	5	72.94%
2011	Chen Yongxin	Drug property discrimination	Support Vector Machines	Single drug	20	None	chemical composition	20	20	100.00%
2012	Zhang Xinxin	Drug property discrimination	Principal component-linear discriminant analysis	Single drug	1725	1380vs345	Attribute characteristics	345	311	90.14%
2013	Qi Jieping	Drug property discrimination	Support Vector Machines	Single drug	28	None	chemical composition	280	262	93.60%
2012	Wang Peng	Drug property discrimination	Logistic regression/support vector machine	Single drug	1728	None	Various Chinese medicine indicators	864	636	73.61%
2012	Liu Wenhui	Drug property discrimination	Partial least squares linear discriminant analysis	Single drug	1725	1380vs345	Indications	1725	1633	94.67%
2012	Liu Wenhui	Drug property discrimination	Partial least squares discriminant analysis	Single drug	1725	1380vs345	Indications	1725	1600	92.75%
2011	Qifang	Drug property discrimination	Partial least squares discriminant analysis	Single drug	61	None	Material composition	13	12	92.31%
2008	careful	Drug property prediction	Decision tree	Single drug	507	None	Natural taste	no	no	no
2008	careful	Drug property prediction	Decision tree	Single drug	507	None	Natural taste	no	no	no
2014	Wu Siyuan	Classification prediction	Random forest	Single drug	96	67vs29	Mass-to-charge ratio variable	24	21	87.50%
2014	Wu Siyuan	Classification prediction	RPART	Single drug	96	67vs29	Mass-to-charge ratio variable	24	23	95.20%
2014	Wu Siyuan	Classification prediction	Support Vector Machines	Single drug	96	67vs29	Mass-to-charge ratio variable	24	22	92.50%
2013	Wang Xiaoyan	Drug property discrimination	Principal component linear discrimination analysis	Single drug	20	30vs30	chemical composition	20	20	100.00%
2015	Nie Bin	Drug property discrimination	Random forest	Single drug	109	None	chemical composition	no	no	94.74%

Note: “None” in the table only represents that this item is unrecorded in the literature, and the “Drug property discrimination” defaults to the discrimination of cold-hot drug property.

The literatures are sorted and classified. Finally, three literatures closely related to the research are obtained. The discriminant basis is the chemical composition. The data is completed, and the common authors used the same two data mining methods to meet the requirements of Evidence-based data analysis requirements. There are six data mining methods in these three literatures, including support vector machine (SVM), regression discriminant analysis (Logistic-DA), principal component analysis-linear discriminant analysis (PCA-LDA), and partial least squares discriminant analysis (PLS-DA), random forest model (RF), linear discriminant analysis (LDA). The basic characteristics of the qualified literatures included in the cold-hot drug property study are shown in Table 2-13.

(10) Document retrieval process and quality evaluation

(a) Literature retrieval results. The screening process and results of the literature retrieval are shown in Figure 2-14. The test results of comparative data mining model are collated. The two groups of comparative data mining methods are formed as the control group and the experimental group, respectively. The predicted accurate quantity and the total sample size of the control group are set to te and tt, respectively. The predicted accurate quantity and the total sample size are set to ce and ct, respectively. The results of the comparison data are shown in Tables 2-14 to 2-21.

Table 2-13. Basic characteristics of the literature included in the study.

Author	The amount of sample	Cold/hot (taste)	Train vs Test (taste)	Predication amount	Mining method	Accurate No.	Accuracy rate (%)
Wang Xiaoyan 2015	60	30/30	48vs12	120	SVM	67	58.30
Wang Xiaoyan 2015	60	30/30	48vs12	120	Logistic-DA	48	40.00
Wang Xiaoyan 2015	60	30/30	48vs12	120	PCA-LDA	53	44.30
Wang Xiaoyan 2015	60	30/30	48vs12	120	PLS-DA	62	51.80
Wang Xiaoyan 2015	60	30/30	48vs12	120	RF	38	31.30
Wang Peng et al 2014	60*4	(30/30)*4	(48vs12)*4	48	LDA	29	60.15
Wang Peng et al 2014	60*4	(30/30)*4	(48vs12)*4	48	Logistic-DA	20	41.68
Wang Peng et al 2014	60*4	(30/30)*4	(48vs12)*4	48	PCA-LDA	22	45.83
Wang Peng et al 2014	60*4	(30/30)*4	(48vs12)*4	48	PLS-DA	23	47.90
Wang Peng et al 2014	60*4	(30/30)*4	(48vs12)*4	48	RF	25	51.25
Wang Peng et al 2014	60*4	(30/30)*4	(48vs12)*4	48	SVM	24	50.00
Long Wei et al 2011	284	153/131	284vs284	284	LDA	206	72.40
Long Wei et al 2011	284	153/131	284vs284	284	SVM	230	81.00

Table 2-14. Data collation of SVM and Logistic-DA test set.

Author	SVM		Logistic-DA	
	te	tt	ce	ct
Wang xiaoyan 2015	67	120	48	120
Wang peng 2014	24	48	20	48

Table 2-15. Data collation of SVM and PCA-LDA test set.

Author	SVM		PCA-LDA	
	te	tt	ce	ct
Wang xiaoyan 2015	67	120	53	120
Wang peng 2014	24	48	22	48

Table 2-16. Data collation of SVM and PLS-DA test set.

Author	SVM		PLS-DA	
	te	tt	ce	ct
Wang xiaoyan 2015	67	120	62	120
Wang peng 2014	24	48	23	48

Table 2-17. Data collation of SVM and RF test set.

Author	SVM		RF	
	te	tt	ce	ct
Wang xiaoyan 2015	67	120	38	120
Wang peng 2014	24	48	25	48

Table 2-18. Data collation of SVM and LDA test set.

Author	SVM		LDA	
	te	tt	ce	ct
Long wei 2011	230	284	206	284
Wang peng 2014	24	48	29	48

Table 2-19. Data collation of PCA-LDA and Logistic-DA test set.

Author	PCA-LDA		Logistic-DA	
	te	tt	ce	ct
Wang xiaoyan 2015	53	120	48	120
Wang peng 2014	22	48	20	48

Table 2-20. The discrimination analysis of PLS-DA and Logistic-DA.

Author	PLS-DA		Logistic-DA	
	te	tt	ce	ct
Wang xiaoyan 2015	62	120	48	120
Wang peng 2014	23	48	20	48

Table 2-21. The discrimination analysis of PLS-DA and PCA-LDA.

Author	PLS-DA		PCA-LDA	
	te	tt	ce	ct
Wang xiaoyan 2015	62	120	53	120
Wang peng 2014	23	48	22	48

(b) Literature quality evaluation. The TCM used in the literatures included in this study have clear drug properties. The proportion of cold and hot drug properties is not much different. The discriminating basis is the chemical composition of TCM. The prediction data are selected from the prediction results of the test set. The application of the comparative mining method has the same two authors. The experimental data records are complete, and the included literature records the correct number and total number of predictions. The above features ensure the quality of the literature included in the study.

(11) Meta-analysis of data mining methods in the discrimination of cold-hot drug properties of TCM

We import the data in Table 2-14 to Table 2-21 into Stata15.0, enter gen tn=tt-te, gen cn=ct-ce in the command box to obtain the data tn and cn. The te and ce represent the predicted accurate quantity. tt, ct represent the total sample size. We click on the toolbar "User-Meta-analysis-of Binary and Continuous (metan)" to select the adapted model for regular meta-analysis.

According to the literatures, six data mining methods are formed into eight comparison groups. The data mining methods SVM, Logistic-DA, PCA-LDA, PLS-DA, RF, and LDA are represented by A, B, C, D, E, and F, respectively. The comparison groups formed are AB, AC, AD, AE, AF, BC, BD, CD. The meta-analysis is performed using STATA 15.0. The AE and AF groups are heterogeneous, so the synthesis of them is abandoned. The analysis results of the comparison groups AB, AC, AD, BC, BD, and CD are shown in Figures 2-14 to 2-19.

The comparison results of method A and method B is shown in Figure 2-14. The prediction accuracy rate of A is 54.17% (91/168), and the prediction accuracy rate of B is 40.48% (68/168). The inter-AB heterogeneity test shows that there is no statistical heterogeneity (P = 0.534, I^2 = 0.0%) between them. The forest map uses a fixed effect model. The results show that the combined OR values are 1.74>1 and 95% CI (1.13, 2.68). There is a statistical significance between them (P = 0.012 < 0.05). The prediction accuracy rate of the data mining method A in the discrimination of cold-hot drug property is significantly higher than that of method B.

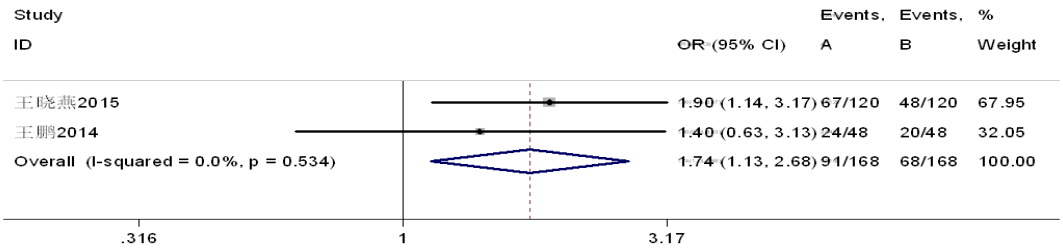


Figure 2-14. Meta-analysis forest map of prediction accuracy rate for drug property discrimination of SVM and Logistic-DA.

The comparison results of method A and method C is shown in Figure 2-15. The prediction accuracy rate of A is 54.17% (91/168), the prediction accuracy rate of C is 44.64 (75/168). The two inter-AC heterogeneity test shows that there is no statistical heterogeneity (P = 0.533, I^2 = 0.0%) between them. The forest map uses a fixed effect model. The results show that the combined OR values are 1.47>1 and 95% CI (0.95, 2.25). There is no statistical significance between them (P = 0.081>0.05). The methods A and C have similar effects in discrimination of drug property.

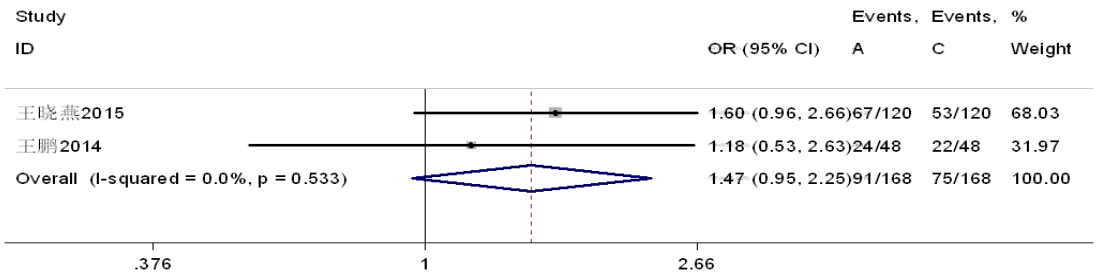


Figure 2-15. Meta-analysis forest map of prediction accuracy rate for drug property discrimination of SVM and PCA-LDA.

The comparison results of method A and method D are shown in Figure 2-16. The prediction accuracy rate of A is 54.17% (91/168), and the prediction accuracy rate of D is 50.60% (85/168). The two inter-AD heterogeneity test shows that there is no statistical heterogeneity (P = 0.865, I^2 = 0.0%) between them. The forest map uses a fixed effect model. The combined OR values are 1.15>1 and 95% CI (0.75, 1.77). There is no statistical significance between them (P = 0.512>0.05). The methods A and D have similar effects in drug discrimination.

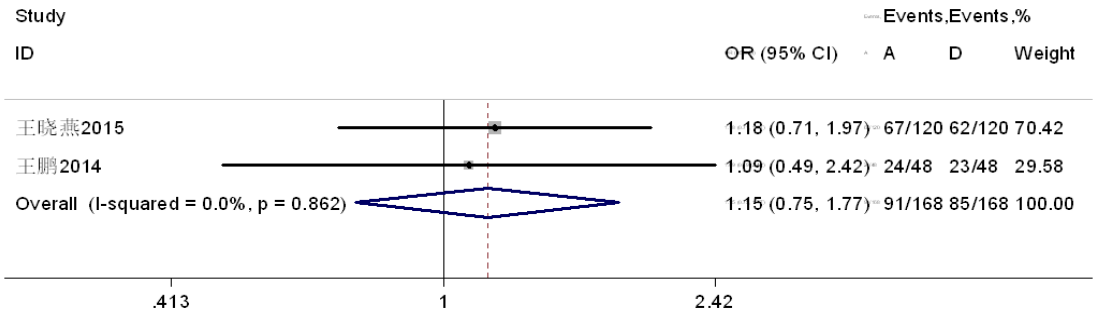


Figure 2-16. Meta-analysis forest map of prediction accuracy rate for drug property discrimination of SVM and PLS-DA.

The comparison results of method B and method C are shown in Figure 2-17. The prediction accuracy rate of B is 40.48% (68/168), and the prediction accuracy rate of C is 44.64% (75/168). The inter-BC heterogeneity test shows that there is no statistical heterogeneity (P = 0.486, I^2 = 0.0%) between them. The forest map uses a fixed effect model. The combined OR values are 1.19>1 and 95% CI (0.77, 1.83). There is no statistical significance between these two (P = 0.440>0.05). The methods B and C have similar effects in drug property discrimination.

The comparison results of method B and method D are shown in Figure 2-18. The prediction accuracy rate of B is 40.48% (68/168), and the prediction accuracy rate of D is 50.60% (85/168). The inter-BD heterogeneity test shows that there is no statistical heterogeneity (P = 0.252, I^2 = 0.0%) between them. The forest map uses a fixed effect model. The combined OR values are 1.51>1 and 95% CI (0.98-2.32). There is no statistical significance between them (P = 0.063>0.05). The methods B and D have similar effects in drug property discrimination.

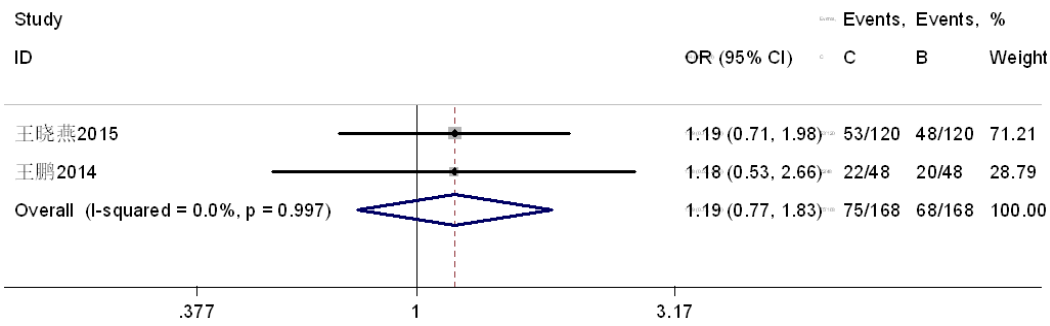


Figure 2-17. Meta-analysis forest map of prediction accuracy rate for drug property discrimination of PCA-LDA and Logistic-DA.

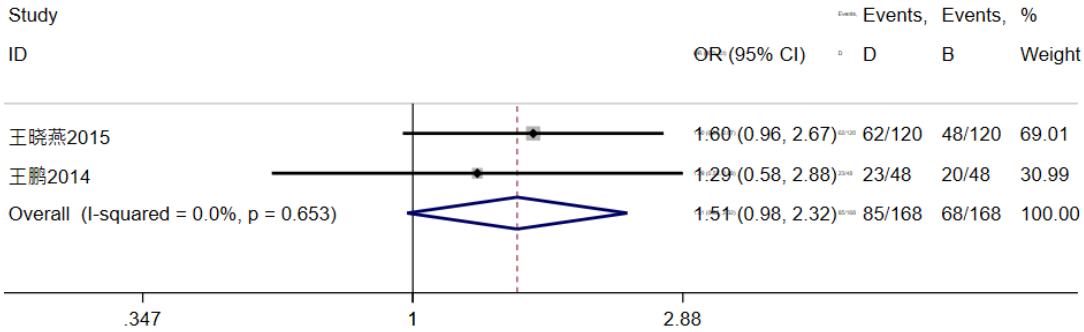


Figure 2-18. Meta-analysis forest map of prediction accuracy rate for drug property discrimination of PLS-DA and Logistic-DA.

The comparison results of method C and method D are shown in Figure 2-19. The prediction accuracy rate of C is 44.64% (75/168), and the prediction accuracy rate of D is 50.60% (85/168). The inter-CD heterogeneity test shows that there is no statistical heterogeneity (P = 0.252, I^2 = 0.0%) between them. The forest map uses a fixed effect model. The combined OR values are 1.27>1 and 95% CI (0.83, 1.95). There is no statistical significance between them (P = 0.275>0.05). The methods C and D have similar effects in drug property discrimination.

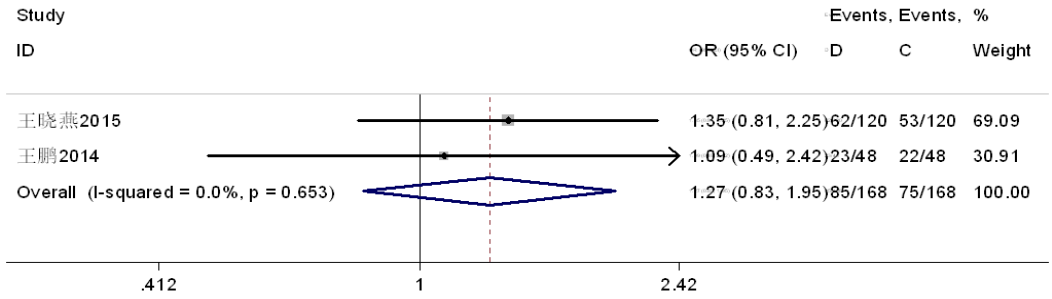


Figure 2-19. Meta-analysis forest map of prediction accuracy rate for drug property discrimination of PLS-DA and PCA-LDA.

(12) Sensitivity analysis

If the combined results from the two models are consistent, the analysis results will be reliable. For the six groups with no significant heterogeneity, the fixed effect model and the

random effect model are used for analysis, and results are shown in Table 2-22. the differences in the OR value, 95% CI, *Z* value and *P* value between the two models in each group is extremely small. Thus, the reliability of the research results is high.

Table 2-22. Sensitivity analysis of comparison results of predication accurate rates of different application methods for drug property discrimination.

No.	Comparison	Model	OR value 95% CI	*Z* value	*P* value
1	SVM *vs* Logistic-DA	Fixed effect model	1.74 (1.13,2.68)	2.50	0.012
2	SVM *vs* Logistic-DA	Random effect model	1.74 (1.13,2.68)	2.50	0.012
3	SVM *vs* PCA-LDA	Fixed effect model	1.47 (0.95,2.25)	1.74	0.081
4	SVM *vs* PCA-LDA	Random effect model	1.47 (0.95,2.25)	1.74	0.082
5	SVM *vs* PLS-DA	Fixed effect model	1.15 (0.75,1.77)	0.66	0.512
6	SVM *vs* PLS-DA	Random effect model	1.15 (0.75,1.77)	0.66	0.512
7	PCA-LDA *vs* Logistic-DA	Fixed effect model	1.19 (0.77,1.83)	0.77	0.440
8	PCA-LDA *vs* Logistic-DA	Random effect model	1.19 (0.77,1.83)	0.77	0.440
9	PLS-DA *vs* Logistic-DA	Fixed effect model	1.50 (0.98,2.32)	1.86	0.063
10	PLS-DA *vs* Logistic-DA	Random effect model	1.51 (0.98,2.32)	1.86	0.063
11	PLS-DA *vs* PCA-LDA	Fixed effect model	1.27 (0.83,1.95)	1.09	0.275
12	PLS-DA *vs* PCA-LDA	Random effect model	1.27 (0.83,1.95)	1.09	0.275

In summary, through the meta-analysis in evidence-based science, the data generated by the application of data mining method in the drug property discrimination of TCM is analyzed. The mining results of each data mining method are compared.

The professional retrieval collects a total of 2064 literatures. Finally, three valid literatures are included, containing six data mining methods and a total of eight groups of comparisons. Among them, the predication accuracy rates of “support vector machine” and “regressive discriminant analysis” predictive effect comparison, the total prediction accuracy rate are OR = 1.74, 95% CI (1.13,2.68). Two groups are heterogeneous. The remaining five groups are not statistical significance (P>0.05). “support vector machine” is better than the "regressive discriminant analysis" in the drug property discrimination of TCM. The five comparison groups of “support vector machine” *vs* “principal component linear discriminant analysis”, “regressive discriminant analysis” *vs* “principal component linear discriminant analysis”, “support vector machine” *vs* “partial least squares discriminant analysis”, “regression discriminant analysis” *vs* “partial least squares discriminant analysis”, “principal component linear discriminant analysis” *vs* “partial least squares discriminant analysis” have similar predicative effects in drug property. In this chapter, we change the model, and use the random effect model and the fixed effect model to analyze the OR, the 95% confidence interval, the *Z* value and the *P* value. Then, the above data are compared, and the difference among them is very small. Thus, the reliability of this study is considered to be high.

2.2.4 Systematic evaluation of data mining methods for predication of TCM efficacy

(1) Literature inclusion criteria

According to the retrieval requirement, the inclusion criteria are as follows:

(a) The language of the incorporated literatures is limited to Chinese;

(b) Literature for establishing models and conducting efficacy prediction through basic information of TCM;

(c) The training set and the test set are randomly grouped, and the efficacy of the TCM or the compound efficacy of the TCM compound is clear;

(d) The research object is the efficacy of TCM or the compound efficacy of TCM;

(e) The modeling bases are four natures, five flavors, channel tropism, toxicity, primary biomass metabolites, secondary biomass metabolites, pharmacological indicators, efficacy of TCM, *etc.* These can be used to establish a prediction model of compound efficacy of TCM.

(2) Research type

The research type of this chapter is "randomized controlled trial". The research object is the prediction of TCM efficacy, and the intervention measure is data mining method. The comparison in efficacy predication of TCM between two data mining methods is performed. The prediction model for the efficacy or compound efficacy of TCM is shown in Figure 2-20.

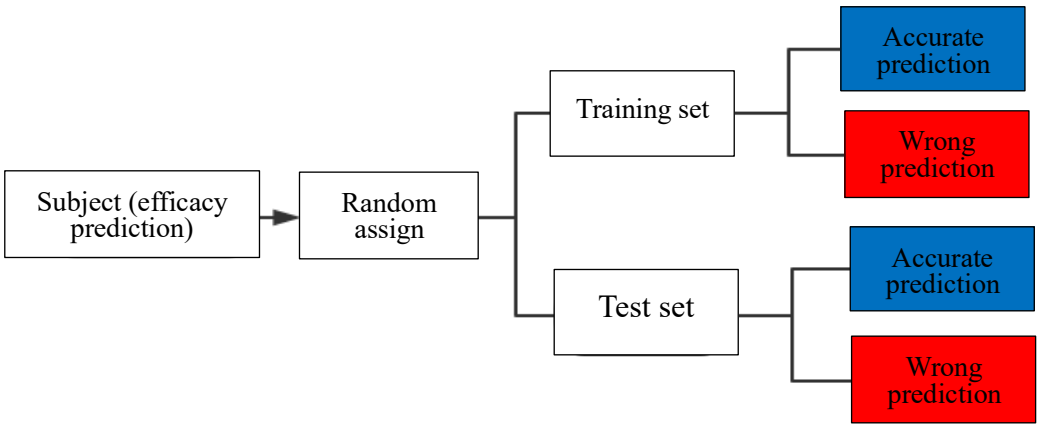


Figure 2-20. The prediction model for the efficacy or compound efficacy of TCM.

(3) Research object

The research object is the literature about the data mining method applied to the prediction of efficacy or compound efficacy of TCM in CNKI, VIP, and WanfangData from the database construction to September 9, 2018.

(4) Literature exclusion criteria

According to the research requirement, the following exclusion criteria are listed:

(a) Research that is significantly unrelated to the prediction of efficacy or compound efficacy of TCM;

(b) Research about the neurological effect or immunomodulatory effect;

(c) Literature review articles;

(d) Literatures that are repeatedly published, whose source is unknown, or that are inconsistent with the actual research.

(5) Outcome indicators of evaluation

The prediction model of TCM efficacy is built according to the basic information of TCM, and then the model is tested by some data information. The prediction accuracy rate and the

accurate prediction quantity of the TCM can reflect the application effect of data mining method in TCM. Thus, in this chapter, we select the prediction accuracy rate of single TCM efficacy and prediction accuracy rate of TCM compound efficacy as the main outcome indicators.

(6) System evaluation results

The literature data of the prediction of single Chinese medicine efficacy and the prediction of TCM compound efficacy are extracted. The extraction results are shown in Table 2-23 to 2-24. The extracted information includes the publication year, author, mining content, mining method, mining type, total size of the sample, and construction model basis, predication of total quantity, predication of accurate quantity, and prediction accuracy rate. Due to the insufficient document width, the information such as mining software and prescription source is omitted.

(a) Evaluation of data mining methods for efficacy prediction of single TCM

As shown in Table 2-23, there are few data mining methods applied to the prediction of the efficacy of single Chinese medicine and of TCM compound. The data mining methods for the prediction of the efficacy of single TCM are artificial neural network, principal component analysis, decision tree, least squares method, support vector machine, Bayesian network, correspondence analysis and mathematical statistics. The mining content is mainly classified into three categories of efficacy classification, efficacy prediction and performance analysis. The number of literatures for both efficacy classification and efficacy prediction is five. 12 groups of data include a total of 7 researchers. Among these groups, four of which use Bayesian networks, three of which use neural networks, and 2 of which use least squares-support vector machines, one of which uses mathematical statistics, one of which uses the corresponding analysis, and one of which uses the principal component analysis. The model of four groups is established based on the natural taste and channel tropism. The model of three groups is established based on pharmacological indicators. The model of two groups is established based on characteristic attributes. The model of two groups is established based on trace elements. The model establishment basis of one group is not clear. The prediction accuracy rate for the above models is greater than 79.2%, of which six groups have prediction accuracy rate greater than 90%, three groups have accuracy rate of 100%. Three groups of data are insufficient, and their model prediction accuracy rates are not obtained. Although the known model prediction rate is high, it can be seen from the table that the total amount of prediction is the number of test sets. There are five studies with sample size ≤20, of which two is about the efficacy classification and three is about efficacy predictions. Except for Yao Meicun and other functional classification studies, the prediction accuracy rates are basically consistent. Chen Zhao *et al.* achieved a prediction accuracy rate of 78.6% by the least squares-support vector machine. The prediction accuracy of the other three groups reached 100%. Except for the above four groups and three groups of incomplete data, the other three studies have a large number of test sets, all of which are ≥30, and the prediction accuracy rates of them are between 79.2% and 96.3%. The increase of the number of training sets and test sets can increase the authenticity of model test. The small number of test sets results in that the accuracy rate of model prediction can prove that the model may be suitable for only single TCM efficacy prediction. Thus, and the reference is low.

.

Table 2-23. The characteristic information of literatures about efficacy predication of single TCM.

Year	Author	Mining content	Mining method	Type	Amount	Train set *vs* test set	The basis of model	Quantity of prediction	Accurate quantity	Accuracy rate
2003	Yao Meicun	Efficacy classification	Artificial neural network	Single drug	426	426 *vs* 5	Natural taste	5	5	corresponding
2009	Gao Jinhong	Efficacy classification	Principal component analysis-artificial neural network	Single drug	20	17 *vs* 3	Trace element	3	3	100.00%
2004	Yao Meicun	Efficacy classification	Decision tree	Single drug	54	54 *vs* 54	Natural taste	54	47	87.00%
2004	Yao Meicun	Efficacy classification	Artificial neural network	Single drug	54	54 *vs* 54	Natural taste	54	52	96.30%
2017	Chen Zhao	Efficacy prediction	Least square-support vector machine	Single drug	74	50 *vs* 24	Feature attribute	38	30	79.20%
2017	Chen Zhao	Efficacy prediction	Least square-support vector machine	Single drug	64	50 *vs* 14	Feature attribute	14	11	78.60%
2008	Liu Ying	Efficacy prediction	Bayesian network	Single drug	55	36 *vs* 19	Pharmacological index	19	19	100.00%
2012	Zhong Nujuan	Efficacy prediction	Bayesian network	Single drug	60	None	Natural efficacy	60	None	None
2015	Sinking	Efficacy prediction	Bayesian network	Single drug	301	201 vs 100	Pharmacological index	12	12	100.00%
2010	Gao Jinhong	Efficacy classification	Correspondence analysis	Single drug	13	None	Trace element	None	None	None
2012	Wang Houwei	Performance analysis	Mathematical statistics	Single drug	212	None	None	None	None	None

Note: “None” in the table only represents unrecorded item in the literature.

Table 2-24. The characteristic information of literatures about compound efficacy prediction of TCM.

Year	Author	Mining content	Mining method	Type	Amount	Train set *vs* test set	The basis of model	Quantity of prediction	Accurate quantity	Mining software	Accuracy rate
2016	Li Weiwei	Compound efficacy	BP neural network	TCM prescription	126	116 *vs* 10	Natural taste	40	37	None	92.50%
2006	Peng Jing	Compound efficacy	Neural network	TCM prescription	700	None	Natural taste	None	None	None	83.87%
2006	Peng Jing	Compound efficacy	High dimensional data reduction	TCM prescription	700	None	Natural taste	None	None	None	92.29%
2011	Wu Xiaodong	Compound efficacy	BP neural network	TCM prescription	None	None	Natural taste	None	None	None	None
2012	Zhang Bo	Drug-effect relationship	Rough set/neural network	TCM prescription	14	None	None	None	None	None	None
2013	Wu Huimin	Compound efficacy	Rough set/support vector machine	TCM prescription	225	225 *vs* 26	Feature attribute	26	21	None	80.00%
2010	Mai Qipeng	Efficacy prediction	BP neural network	TCM prescription	48	None	Natural taste	15	14	None	93.33%

Note: “None” in the table only represents unrecorded item in the literature.

(b) Analysis of evaluation results of data mining methods for compound efficacy prediction of TCM. The methods applied in the prediction of compound efficacy mainly include BP neural network, high-dimensional data reduction, rough set, and support vector machine. The mining contents mainly include the compound efficacy prediction of TCM and drug-effect relationship. Most of the studies focus on the former one, and only one literature studies the drug-effect relationship. Among the seven groups of data included, there are six authors, five groups use the neural network, one group uses the high-dimensional data reduction, two groups use rough sets, and one group use support vector machine. The model of two groups is established based on natural taste and channel tropism. The model of two groups is established based on the dose of natural taste and channel tropism. The model of one group is established based on the characteristic attribute. The basis of model of one group is still unclear. The prediction accuracy rate in the model application is greater than 80%, among which the prediction accuracy rate of the three groups is greater than 90%, and two groups fail to obtain the model prediction accuracy rate. In general, the prediction effects of these models are better. This indicates that the data mining methods have certain guiding significance for the prediction of the efficacy of TCM prescription. This provides reference for the study of compound efficacy

2.2.5 Discussion and prospect of data mining in TCM efficacy

(1) Discussion on data mining in TCM efficacy

We summarize the research on data mining methods in TCM efficacy, and introduce the commonly used data mining methods. The evidence-based methods are first used to evaluate the mining results of data mining methods. The data mining methods included in the search keywords are relatively complete. The retrieval is relatively complete with the time of database construction. The results of the included literatures are detailed, and the quality of the literature is high. The sensitivity analysis ensures the reliability of the research. It provides a reference for the reasonable application of data mining methods in TCM information.

In addition, there are four new discoveries in the research process. First, the application scope of data mining in the prediction of TCM is expanding. Besides the efficacy prediction, it has been applied to the prediction of five-flavors, natural taste and channel tropism, efficacy indicators, whether it has neurological effects and whether it has immunomodulatory effects. However, there are few published literatures in these aspects, and the literature characteristics of other aspects of data mining for TCM efficacy are shown in Table 2-25. Second, the emphases on the prediction of drug properties are different. Some of the predictions are based on chemical composition, and some are based on TCM characteristics, physiological and biochemical indicators, biological effects, *etc.* There are many types of predictions. Similarly, there are such predictions in efficacy prediction. Third, the number of test sets in some literatures is small, results in the limitation. In addition, data of some literatures are incomplete. This book avoids the shortcomings of these three aspects. Thus, some documents cannot be included for comparison. This indicates that the quality of Meta analysis results is not intrinsic, and depends on the quality and quantity of the original research. This deficiency should be taken seriously. Fourth, the six data mining methods studied in this chapter can only form eight groups for comparative meta-analysis. For mining methods that do not satisfy the direct

comparison conditions, they cannot be applied. As the number of basic experimental research increases, the number of data mining methods increases, and the sample size of the research increases, more comparison groups can be formed, and the evidence will be more sufficient. The results of various data mining methods can be compared. The research quantity of the efficacy prediction of TCM is far less than that of discrimination of cold-hot drug property. In addition, the number of data mining methods applied in research is small, the number of training sets and test sets in the existing research is small, and the evidence is insufficient. Thus, the quality of prediction result is not very high, and the prediction of TCM efficacy will become a hot spot in the study of efficacy data mining.

(2) Conclusions of data mining in TCM efficacy

TCM information based on data mining method has become one of the hotspots of research of modern TCM. This chapter systematically evaluates the application effect of data mining method in TCM efficacy based on evidence-based scientific method. This provides reference for researchers.

This chapter statistically summarizes various types of information in the database through mathematical statistics methods, visually reproduces some of the information through the column and pie charts in Excel2016. In addition, we use the Cytoscape software to draw graph about the relationship between data mining method and TCM efficacy. The most popular applications of data mining methods in TCM efficacy mining are association rule, cluster analysis, frequency analysis. Common data mining methods for efficacy prediction of single TCM include neural network, principal component analysis, decision tree, least square method, Bayesian network, support vector machine, correspondence analysis, *etc.* Common data mining methods applied to compound efficacy prediction include high-dimensional data reduction, neural network, rough set, support vector machine, space vector, *etc.* Common data mining methods applied to drug property discrimination include artificial neural network, partial least squares, linear discriminant analysis, support vector machine, principal component analysis, Fisher discriminant analysis, discriminant analysis, regression tree, decision tree, *etc*. From the overall application situation, correlation analysis, correlation analysis, prediction, cluster analysis and frequency analysis are the most widely used data mining methods in TCM efficacy. From the analysis results, the number of articles in this aspect is increasing year by year. The quality of the literature is becoming higher and higher. From the perspective of research institutions, most universities in TCM universities and colleges use data mining to solve scientific research problems. From the research content, the papers that discuss the discrimination of cold-hot drug properties are the most in current statistical applications. There are 15 types of data mining methods for cold-hot drug discrimination. In addition, the study of TCM efficacy involves the efficacy prediction of single TCM, efficacy prediction of TCM compound, prediction of nerve effect, prediction of component efficacy, and drug effect index prediction. The application of data mining methods and their research on the efficacy of TCMs has a many-to-many relationship.

In this chapter, to provide reference for the application of data mining methods in this aspect, save the cost of basic experimental research and improve research efficiency, the meta-analysis method is used to evaluate the mining effect of data mining methods in cold-hot drug property discrimination. For the eight groups of comparison, the predication accuracy rate of SVM is higher than that of Logistic-DA, the prediction effect of SVM-LDA and SVM-RF abandons the synthesis analysis. In the remaining five groups, the predication accuracy rates of

"SVM and PCA-LDA", "SVM and PLS-DA", "Logistic-DA and PCA-LDA", "Logistic-DA and PLS-DA", "PCA-LDA and PLS-DA" are different, and the data mining methods of the two users' test results are consistent. The comparison is not statistically significant. Thus, this shows the application effect of these five data mining methods in drug property discrimination is quite good, and the application effect of SVM in the discrimination of cold-hot drug property of TCM is better. The sensitivity analysis shows that there is no significant difference between these groups, which indicates that the research results in this chapter are reliable.

The research on the efficacy of TCM in this chapter is more complicated than that on the drug property of TCM. The efficacy of single TCM is closely related to the drug property. The compound efficacy is not the sum of efficacies of TCM in prescription. Therefore, in the efficacy study, the high-dimensional and multi-linear and other complex cases often emerge in TCM data. This leads to that many data mining methods cannot be applied to the prediction of the TCM efficacy. This chapter analyzes the application of data mining in the efficacy prediction of single TCM and TCM compound. In addition, we summarize the data mining methods in the efficacy prediction. Compared with the case of discrimination of cold-hot drug property, there are few data mining methods for the research on efficacy prediction. Bayesian network and neural network are the most used methods, and frequently appear in the prediction of efficacy and compound efficacy of single TCM. The prediction accuracy rate is greater than 80%. Thus, they are reliable data mining methods of efficacy prediction of TCM. Table 2-25 shows the literature characteristics of other aspects of efficacy mining of TCM.

(3) Prospects of data mining in the TCM efficacy

With the advent of the Internet and the era of big data, research on the efficacy of TCM through data mining has become important to modernize TCM. The research in this area is relatively mature. The professionals in the relevant fields of TCM and computer related fields are constantly repeating scientific experiments to improve the mining efficiency, and to explore more efficient models and algorithms. However, these researches require a lot of basic experiments. This means that it takes a lot of manpower and material resources to pave the way for data mining in efficacy of TCM. Applying the evidence-based method to the data mining method in the efficacy predication of TCM is beneficial to make full use of the generated data. This can help to optimize the data mining method through the existing data, save the experimental cost and improve the research efficiency. However, data mining is a double-edged sword that can be fully utilized, not overly dependent. The discrimination of cold-hot drug property of TCM has certain subjective knowledge. The records of different ancient books and modern works are not consistent, and there is a certain difference from the understanding of modern clinical experience. Therefore, the data mining results cannot be completely used as the final result of drug property discrimination. Similarly, the application of data mining in efficacy prediction is the same.

Besides to the discrimination of cold-hot drug property, the research on other aspects of natural effect of TCM is relatively scarce. There are few basic experimental studies, which cannot be evaluated by meta-analysis. As the number of studies in this area increase, the basic data accumulates, which provides more evidence for the research. Then, results will be more complete. The evaluation of data mining method applied in natural effect of TCM based on evidence-based method is beneficial to make full use of experimental data. It is better to optimize data mining method, save experiment cost and manpower. It has certain development prospects.

Table 2-25. Literature characteristics of other aspects of efficacy mining of TCM.

Year	Author	Ming content	Mining method	Type	Amount	Train set *vs* Test set	The basis of model	Quantity of prediction	Accurate quantity	Accuracy rate
2012	Zhong Nujuan	Natural relationship	Bayesian network	Single TCM	60	None	Natural efficacy	60	None	None
2010	Deng Jiagang	Flatness and non-flatness	Principal component analysis/support vector machine	Single TCM	20	None	chemical composition	20	17	85.00%
2008	Zhang Pei	Five flavor prediction	Bayesian network	Single TCM	198	None	Pharmacological index	9	9	≫90%
2010	Liu Jin	Flatness and non-flatness	Principal component analysis/support vector machine	Single TCM	76	None	Infrared spectral data	76	63	82.89%
2011	Wang Mei	Component efficacy prediction	Bayesian network	Single TCM	363	None	Pharmacological index	None	None	None
2014	Hu Yanan	Natural taste prediction	Decision tree	Single TCM	no	None	Pharmacological index	None	None	None
2014	He Wenjing	Component drug prediction	Bayesian network	Component	140	None	Pharmacological index	140	107	76.24%
2013	Yang Xuemei	Pharmacological classification	Association rule	Single TCM	8980	None	None	None	None	None
2013	Yang Xuemei	Pharmacological classification	Association rule	Single TCM	8980	None	None	None	None	None
2008	Ye Wei et al	Pharmacodynamic index prediction	BP neural network	Single TCM	47	None	Natural taste	15	14	93.33%
2011	Tang Jinbo	Effectiveness association	Three-dimensional model	Single TCM	285	None	Natural taste	285	122	42.81%
2010	Liu Yongqi	Neuroimmuno prediction	Logistic regression	Single TCM	507	None	Natural taste	507	453	89.30%
2009	Liu Yongqi	Neural effect prediction	Logistic regression	Single TCM	507	None	Natural taste	507	420	82.84%
2010	Zhang Yachao	Flatness and non-flatness	Support Vector Machines	Single TCM	100	None	Material composition	None	None	73.00%

Note: “None” in the table only means no record in the literature

References

1. J.W. Han. Data Mining Concepts and Techniques. Mechanical Industry Press: Beijing, China, 2017.
2. Z.H. He. Research on parallel frequent item set mining algorithm. Ph.D Thesis, Sichuan Normal University, Sichuan, China, 2018.
3. X.J. Fu, P. Wang, Z.G. Wang. Hypothesis on building of chemical constituent element system of cold-heat nature based on study of "nature-structure relationship" of traditional Chinese medicine. *World Science and Technology (Modernization of Traditional Chinese Medicine and Materia Medica)* **2011**, *13*, 919-924.
4. H.J. Chen, Z.K. Chen, G.L. Du. Modern research progress on compatibility rules of traditional Chinese medicine prescription. *Chinese Archives of Traditional Chinese Medicine* **2018**, *36*, 2835-2841.
5. J.Q. Wang, C.X. Wang, J.P. Cui, *et al*. Research about therapeutic utility and formula compatibility regularity of splenic related disease based on data mining. *Chinese Archives of Traditional Chinese Medicine* **2018**, *36*, 2605-2610.
6. L.S. He, P. Song, L.H. Zhao, *et al*. Overview of clinical research on dose-effect relationship. *Journal of Traditional Chinese Medicine* **2019**, *60*, 80-84.
7. T.B. Li, J. Shang, H.Z. Zhang, *et al*. Study on medication rules of dunhuang prescriptions based on data mining. *Chinese Journal of Information on Traditional Chinese Medicine* **2019**, *26*, 102-105.
8. X.X. Song, S.H. Li, X.R. Wang, *et al*. Analysis of prescription formula rules of professor ZHANG Shiqing in the treatment of infantile anorexia. *Journal of Pediatrics of Traditional Chinese Medicine* **2019**, *15*, 4-7.
9. S. Cui, M.W. Shen, L.J. Che. He Liren's medication characteristics for the treatment of palpitation on the basis of data-mining. *Acta Universitatis Traditional Medical Sinensis Pharmacologiaeque Shanghai* **2013**, *27*, 32-35.
10. X.M. Yang, X.M. Lai, M.M. Chen, *et al*. Association relationship between functions and flavors of TCM based on classification association rules. *Chinese Journal of Information on Traditional Chinese Medicine* **2013**, *20*, 31-34.
11. H.P. Wu, H.X. Liu, L.X. Lan, *et al*. Application of association analysis in arrangement of drawer position of traditional Chinese medicine. *China Pharmacy* **2012**, *23*, 4120-4122.
12. H.P. Wu, L.X. Lan, H.X. Liu, *et al*. The design of medicine bucket placement of traditional Chinese medicines based on data mining. *China Licensed Pharmacist* **2013**, *10*, 53-56.
13. H.Y. Ji. Research on the Use of Prescriptions Based on Multi-Level Literature Mining. Ph.D Thesis, Beijing University of Chinese Medicine, Beijing, China, 2012.
14. D.J. Xue, T. Huang, Y.B. Shen, *et al*. Analysis on diagnosis and treatment rules of contemporary TCM cases of brain tumors based on data mining. *China Journal of Traditional Chinese Medicine and Pharmacy* **2016**, *31*, 2846-2849.
15. W. Shen, G. Zheng, J.P. Zhan, *et al*. To investigate the rules of symptom, syndrome, strategy and formula in the treatment of rheumatoid arthritis on date mining. *Rheumatism and Arthritis* **2013**, *2*, 5-9.

16. W. Wang. Study on the genesis of Chinese medicine returning theory. Ph.D Thesis, Liaoning University of Traditional Chinese Medicine, Liaoning, China, 2012.
17. L.P. Liu, X.Y. Zhang, Y.K. Guo, *et al*. Study on pharmacodynamics of TCM meridian and tonifying deficiency drugs based on bp neural network. *Software Guide* **2019**, *3*, 1-4.
18. M. Wang, P. Zhang, W. Wang, *et al*. Study on the prediction of Chinese medicine component function based on pharmacological effects. *World Science and Technology (Modernization of Traditional Chinese Medicine and Materia Medica)* **2011**, *13*, 93-95.
19. L.P. Huang, M.F. Zhu, R.Y. Yu, *et al*. Study on discrimination mode of cold and hot properties of traditional Chinese medicines based on biological effects. *China Journal of Chinese Materia Medica* **2014**, *39*, 3353-3358.
20. Y.C. Zhang, M.L. Li, Y.Z. Guo, et al. Study on the correlation between the active ingredients of the mild traditional Chinese medicine and their properties. *Chemical Research and Application* **2010**, *22*, 67-72.
21. B. Zhang. Pharmacodynamic relationship research on data-mining technology in traditional Chinese prescription compatibility based on rough set and neural network. *Journal of Jilin Institute of Chemical Technology* **2012**, *29*, 35-37.
22. J.B. Tang, Z. Shuai, S.J. Xu, *et al*. Based on the mathematical thinking to establish the "natural-effect connection" model of Chinese herbal property. *World Science and Technology (Modernization of Traditional Chinese Medicine and Materia Medica)* **2011**, *13*, 1099-1102.
23. Y.Q. Liu, X.J. Wu, P. Fan, *et al*. Create mathematical model and analysis of correlation between traditional Chinese medicinal characteristics and neurobehavioral effects. *Lishizhen Medicine and Materia Medica Research* **2010**, *21*, 1508-1510.
24. Z.L. Jin, J.X. Hu, H.W. Jin, *et al*. Analysis of traditional Chinese medicine prescriptions based on support vector machine and analytic hierarchy process. *China Journal of Chinese Materia Medica* **2018**, 43, 2817-2823.
25. X.Y. Wang, F. Li. Study on relationship between cold-hot natures and HPLC fingerprints of polysaccharides of 20 traditional Chinese drugs based on SVM. *Journal of Shandong University of Traditional Chinese Medicine* **2012**, *36*, 439-442.
26. J. Yin, C.M. Lu, G.J. Yang. Combinations of discriminatory analysis and logistic regression for classification. *Journal of Applied Statistics and Management* **2014**, *33*, 256-265.
27. Y.Z. Zhao, D.C. Liu, J. Wu, *et al*. Survivability evaluation of backbone network based on linear discriminant analysis and principal component analysis. *Power System Technology* **2014**, *38*, 388-394.
28. J.W. Chen, C.Y. Hu, W. Ma. Rapid identification of Radix Panacis Quinquefolii by spectral imaging combined with partial least square discriminant analysis. *Chinese Journal of Hospital Pharmacy* **2017**, *37*, 847-850.
29. J. Yan, Q. Zhang, W. Liu, *et al*. Establishment of a quality control method for Lu-Gui Tincture based on ~1HNMR-PLS-DA. *Food Science* **2012**, *33*, 69-72.
30. C.G. Wang, B.B. Han, Y. Wang, *et al*. Evaluation on animal model of deficiency cold syndrome and deficiency heat syndrome by using partial least squares-discriminant analysis. *Liaoning Journal of Traditional Chinese Medicine* **2014**, *41*, 1275-1278.

31. W.H. Liu, Y. Li, Y.J. Ji, *et al*. Partial least squares in the discrimination of traditional Chinese herbal medicine property. *Journal of Shandong University (Health Sciences)* **2012**, *50*, 151-154.
32. Y. Wang. Discriminant Model for the Cold and Hot Property of Traditional Chinese Medicine by HPLC-TOF MS. Ph.D. Thesis, Fudan University, Shanghai, China, 2010.
33. Z.L. Zhou. The identification of Cold-Hot medicine Property (CHMP) and the establishment of Partial Least Squares Path Model (PLS path model) based on primary materials. Ph.D. Thesis, Shandong University of Traditional Chinese Medicine, Shandong, China, 2012.
34. B. Nie, J.Q. Du, G.L. Xu, *et al*. Classification and discrimination for traditional Chinese medicine' nature based on PLS-DA. *2010 International Forum on Information Technology and Applications* **2010**, 362-364.
35. N. Hui. The Application of Near Infrared Spectroscopy for Quality Control of Codonopsis Pilosula and Fu Fang Dan Shen Tablets. Ph.D. Thesis, Lanzhou University, Lanzhou, China, 2011.
36. B. Nie, Z.L. Hao, B. Gui, *et al*. The Research for metabolomics discriminant method for cold and hot property of traditional Chinese medicine based on random forest. *Journal of Jiangxi University of Traditional Chinese Medicine* **2015**, *27*, 82-86.
37. T.T. Zheng, M. Yang, Z.J. Zhai, *et al*. Optimization of effect of traditional Chinese medicine compound jinfukang on lung cancer cell proliferation based on random forest regression model. *Chinese Journal of Experimental Traditional Medical Formulae* **2017**, *23*, 177-182.
38. S.Y. Wu, Y.F. Hu, X.W. Liu, *et al*. Study on the random forest model of cold and hot property classification of traditional Chinese Medicine. *Software Guide* **2014**, *13*, 71-74.
39. X.Y. Ji. Application of Combined Chromatography in Chinese Medicine and Plasma Amino Acids. Ph.D. Thesis, Central South University, Hunan, China, 2010.
40. X.X. Zhang, Y. Li, Y.J. Ji, *et al*. Discrimination of properties of Chinese Traditional Medicine with principal component analysis-linear discriminant analysis. *Journal of Shandong University (Health Sciences)* **2012**, *50*, 143-146.
41. R.G. Zhu, Z.C. Zheng, H.J. Sun. Locality-regularized linear regression classification-based discriminant analysis. *CAAI Transactions on Intelligent Systems* **2019**, *5*, 1-8
42. L. Xue. Research and Design of Facial Feature Extraction and Classification Recognition Algorithm. Ph.D Thesis, Nanjing University of Posts and Telecommunications, 2018.
43. F. Chen, Y.K. Huang, Y.Y. Yuan, *et al*. Nondestructive identification of adulterated honey based on electronic nose. *Journal of Xihua University (Natural Science Edition)* **2018**, *37*, 56-60.
44. K.F. Li, J.Y. Lu, M.M. Huang, et al. Face recognition based on wavelet denoising and linear discriminant analysis. *Bulletin of Science and Technology* **2018**, *34*, 115-118.
45. Q.S. Lu. Optimization and Research of BP Neural Network. Master Thesis, Zhengzhou University, Henan, China, 2011.
46. L. Su, X.D. Song. Implementation and application of probabilistic neural network based on Matlab. *Computer and Modernization* **2011**, *11*, 47-50.
47. W.W. Li, X.Y. Zhang, Y. Yan, *et al*. Effects of traditional Chinese medicine compound based on BP neural network. *Guiding Journal of Traditional Chinese Medicine and Pharmacy* **2016**, *22*, 38-41.

48. H.M. Wu, S.Z. Ye. Analysis for Chinese medicine efficacy of osteoarthritis compound based on rough set and SVM. *Journal of Fuzhou University (Natural Science Edition)* **2013**, *41*, 311-316.
49. J.H. Zhang. Evidence-based Chinese Medicine. Shanghai Science and Technology Press: Shanghai, China, 2018; pp. 17-18.
50. X.Y. Wang, F. Li. Study on statistical pattern recognition model for relationship between cold-heat natures and lipid based on GC-MS of 60 traditional Chinese medicines. *Liaoning Journal of Traditional Chinese Medicine* **2015**, *42*, 1303-1305.
51. P. Wang, H.L. Zhou, F.Z. Xue, *et al*. Analysis of infrared spectra of 60 kinds of plant extract of traditional Chinese medicine and study on the identification and evaluation of characteristics of the regional markers associated with cold and heat nature. *Spectroscopy and Spectral Analysis* **2014**, *34*, 58-63.
52. W. Long. The Application of Computational Chinese Materia Medica to the Studies of Chinese Materia Medica Properties and Traditional Chinese Medicine Prescriptions. Ph.D. Thesis, Peking Union Medical College, Bejing, China, 2011.
53. W. Wang, Z.G. Zhou, J. Li, *et al*. Analysis of correlation between contents of primary materials of 20 kinds of traditional Chinese drugs and the property. *Journal of Shandong University of Traditional Chinese Medicine* **2010**, *34*, 99-102.
54. Z.L. Zhou, F. Li, P. Hu. GC/MS fingerprints and fisher discrimination analysis on water-soluble sugar from 20 Chinese herbs with cold and heat properties. *Chinese Journal of Experimental Traditional Medical Formulae* **2010**, *16*, 41-44.
55. J.W. Li, F. Li, Z.L. Zhou. Relationship between high performance liquid chromatographic fingerprints of water-soluble sugar of 20 kinds of traditional Chinese herbs and both cold and heat properties. *Journal of Shandong University of Traditional Chinese Medicine* **2010**, *34*, 195-199.
56. Z.L. Zhou, F. Li, J.W. Li. Study on relationship between sugar content and cold-heat nature of 20 kinds of herbs by fisher analysis. *World Science and Technology (Modernization of Traditional Chinese Medicine and Materia Medica)* **2010**, *12*, 558-561.
57. F. Qi, R. Rong, F.Z. Xue. Application of the Bayesian network in Chinese herbal medicine property recognition. *Journal of Shandong University (Health Sciences)* **2011**, *49*, 147-152.
58. Y. Li, W. Li, F.Z., *et al*. Discrimination of properties of Chinese traditional medicines based on an artificial neural network. *Journal of Shandong University (Health Sciences)* **2011**, *49*, 57-61.
59. L.P. Huang, M.F. Zhu, R.Y. Yu, *et al*. Study on discrimination mode of cold and hot properties of traditional Chinese medicines based on biological effects. *China Journal of Chinese Materia Medica* **2014**, *39*, 3353-3358.
60. Y.N. Hu, Y.L. Ren, J. Cao, *et al*. Predictive study on properties of traditional Chinese medicine components based on pharmacological effects. *China Journal of Chinese Materia Medica* **2014**, *39*, 2382-2385.
61. Y.X. Chen, F. Li, Z.Y. Sun, *et al*. Support vector machines analysis of free lipid compositions on cold or heat property of traditional Chinese medicines. *Liaoning Journal of Traditional Chinese Medicine* **2011**, *38*, 127-129.

62. J.P. Tan, J. Liu, Y.P. Chen, *et al*. Study on the correlation between the volatile components of TCMs for relieving the exterior syndromes and their medicinal properties. *Computers and Applied Chemistry* **2013**, *30*, 85-88.
63. P. Wang, Z.G. Wang, F.Z. Xue, *et al*. Correlation between traditional Chinese medicine traits and medicinal properties based on support vector machine method. *Jiangxi Traditional Chinese Medicine* **2012**, *43*, 65-68.
64. F. Qi, R. Rong, F.Z. Xue. The PLS statistic pattern recognition model for identifying the CHMP-markers. *Chinese Journal of Health Statistics* **2011**, *28*, 628-631.
65. X.Y. Wang. Study on correlation between HPLC chromatogram of botanical polysaccharides and cold-heat potencies based on principal components-linear discriminant analysis. *Journal of Shandong University of Traditional Chinese Medicine* **2013**, *37*, 156-159.
66. Z. Chen, Y.F. Cao, S.B. He, *et al*. Study on classification model of heat-clearing herbs based on theory of medicinal property. *China Journal of Traditional Chinese Medicine and Pharmacy* **2017**, *32*, 2107-2111.
67. N.J. Zhong, Y.M. Song, G.S. Liu, *et al*. Construction and application of the Bayes network model in traditional Chinese medicine elements. *Journal of Shandong University (Health Sciences)* **2012**, *50*, 157-160.
68. W. Shen, W.P. Chen. Study based on Bayesian networks of the relationship between clinical effect and pharmacological effect of traditional Chinese medicine on activating blood circulation and removing stasis in compendium of Materia medic. *Journal of Nanjing University of Traditional Chinese Medicine* **2015**, *31*, 231-233.
69. J.H. Gao, Q.Y. Wang, S.R. Wang. Research on efficacy of Chinese medicines by correspondence analysis. *Analytical Instrumentation* **2010**, *6*, 56-60.
70. H.W. Wang, Y.L. Dou. Drug nature-efficacy analysis of anti-microbial traditional Chinese drugs based on computer programming. *Liaoning Journal of Traditional Chinese Medicine* **2012**, *39*, 1716-1718.
71. X.D. Wu. Numerical model of traditional Chinese medicine efficacy based on BP neural network and subjective evaluation. *Gansu Science and Technology* **2011**, *27*, 19-21.
72. B. Zhang. Pharmacodynamic relationship research on data-mining technology in traditional Chinese prescription compatibility based on rough set and neural network. *Journal of Jilin Institute of Chemical Technology* **2012**, *29*, 3
73. Q.P. Ma, Y.H. Wu, X.E. Li. BP model for relation between anti-aging and four natures, five flavors and meridian tropism. *Computer Engineering and Applications* **2010**, *46*, 220-223.
74. J.G. Deng, J. Liu, J.P. Zhai, *et al*. Analysis of the properties and relationships between the infrared spectra of 20 kinds of TCM. *Chinese Journal of Spectroscopy Laboratory* **2010**, *27*, 741-744.
75. J. Liu, J.G. Deng, J.P. Yan, *et al*. Study on the property recognition of traditional chinese medicines based on infrared spectrum. *Lishizhen Medicine and Materia Medica Research* **2010**, *21*, 561-563.
76. W.J. He, Y.N. Hu, Y.L. Zhang, *et al*. Study on self-similarity relationship between decoction pieces property and component property. *China Journal of Chinese Materia Medica* **2014**, *39*, 2375-2377.

77. X.M. Yang, D.Y. Lin, X.M. Lai, *et al*. Mining rules on determination of four properties based on traditional Chinese medicine functional combination. *China Journal of Chinese Materia Medica* **2013**, *38*, 1624-1626.
78. X.M. Yang, X.M. Lai, M.M. Chen, *et al*. Determination of rules of five viscera channel tropism based on Chinese herbs functional combination mining. *Journal of Fujian University of Traditional Chinese Medicine* **2013**, *23*, 49-52.

Chapter 3 Application and Examples of Algorithms in Traditional Chinese Medicine

3.1 Research and example of parallel association rule mining algorithm based on spark

Data mining (DM) is also known as knowledge discovery from data (KDD). Its main function is to discover valuable information from massive data. Common data mining functions include characterization and differentiation, frequent pattern, association, and correlation mining, classification and regression, cluster analysis, and outlier analysis. Among them, association rule mining is to discover association relationships hidden in different objects or variables from data, and is the basis for studying other branches of data mining. Mining association rule is divided into two steps. One is to discover frequent itemsets from the existing database. The other one is to generate strong association rules from frequent itemsets. The basic mining methods are multi-candidate generation (Apriori, division, sampling, *etc*.), pattern growth (FP-growth, HMmine, FPMax, Close +, *etc*.), and vertical format (Eclat, CHARM, *etc*.).

As the amount of data grows exponentially every day, the traditional single-machine, serially running algorithms can no longer meet the requirement. Therefore, it is necessary to develop distributed and parallel computing mining algorithms. Hadoop MapReduce is a disk-based distributed parallel computing model. However, there are many shortcomings, such as low abstraction level, only supporting Map and Reduce operations, low processing efficiency, and not suitable for iterative operations. Most data mining algorithms are iterative based on memory. Apache Spark is a parallel computing framework developed by the University of California, Berkeley, AMLab laboratory for fast processing of large data. Compared with th case of Hadoop MapReduce, memory-based programs of Spark run 100 times faster and disk-based operations of Spark run 10 times faster. The core concept in Spark is the elastic distributed data set RDD, which is an immutable distributed object set. Each data set in it is divided into logical partitions and can be calculated in parallel on different nodes in the cluster. Spark overcomes many of the shortcomings of Hadoop MapReduce, has higher levels of abstraction, higher efficiency, and features of iterative processing and memory computing.

3.1.1 An overview of association rule data mining

(1) Association rule

$\mathrm{I}=\{I_1,\ I_2,\ldots,\ I_n\}$ is a set of items with a total of n items. $\mathrm{T}=\{T_1,\ T_2,\ldots,\ T_m\}$ is a task-related database transaction set with a total of m transactions, each transaction is required to be a non-empty item set (the set of item is called an item set, containing k terms), and $T_k \subseteq I\left(1 \le k \le m\right)$.

$\mathrm{A}\left(A \subseteq T\right)$ and $\mathrm{B}\left(\mathrm{B} \subseteq T\right)$ are item sets, and satisfy $\mathrm{A} \subset \mathrm{I}$, $\mathrm{B} \subset \mathrm{I}$, $\mathrm{A} \neq \varnothing$, $B \neq \varnothing$ $\mathrm{A} \cap B = \varnothing$. The association rule is an implication of the form "$A \Rightarrow B$". A is the antecedent of

the implied form, and B is the seccedent of the implied form. Each association rule requires corresponding support and confidence. Support is the number of items in the transaction database. The confidence formula is $confidence(A \Rightarrow B) = \frac{support(A \cup B)}{support(A)}$. In general, if the rule "$A \Rightarrow B$" meets the minimum support threshold ($min\text{-}sup$), the rule "$A \Rightarrow B$" is a frequent pattern. If the rule "$A \Rightarrow B$" meets both the minimum support threshold and the minimum confidence threshold ($min\text{-}conf$) (minimum support threshold and minimum confidence threshold are generally provided by experts or users with professional knowledge), then the rule "$A \Rightarrow B$" is called a strong association rule. A strong association rule is a sense of decision makers Rules of interest. Association rule mining is to discover all frequent patterns from the existing database, and then to discover strong association rules from frequent patterns. It can be known from the confidence formula that the confidence is easily derived from the support degree. Thus, association rule mining is largely affected by the mining efficiency of frequent patterns.

(2) Association rule mining algorithm

The main overhead of association rule mining algorithm is to generate all frequent itemsets. Therefore, the association rule algorithm in the general sense is an algorithm for mining frequent patterns. There are Apriori, FP-growth, Eclat, DHP, Sampling, DIC, Partition, Piner Search, Clique, MaxEclat, dEclat, Charm ++, MaxClique and other algorithms. Apriori, FP-growth, and Eclat are the three most classic algorithms of association rule mining algorithms.

(a) Apriori algorithm. The Apriori algorithm was proposed by Agrawal and R. Srikant in 1994. It uses a layer-by-layer search strategy to constantly explore *k* item sets from frequent *k* + 1 item sets. Steps of Apriori algorithm are as follows:

(I) Discovering all frequent 1-item sets;

(II) Candidate 2 item sets are generated by frequent 1 item set concatenation. The database is scanned multiple times to determine whether the candidate 2 item sets meet the minimum support degree. If it is not satisfied, it will "prune". If it is satisfied, the candidate item set forms a frequent 2 item set.

(III) After frequent 2 item sets are "connected" and "primary pruning", candidate 3 item sets are generated. The database is scanned multiple times to determine whether the candidate 3 item sets meet the minimum support degree. If it is not satisfied, it will "prune". If it is satisfied, the candidate forms a frequent 3 itemset.

(IV) Repeating step III, and continuing to form candidate *k* + 1 item set from frequent *k* item set, until the candidate item set can no longer be formed. Then, the algorithm ends.

The idea of Apriori algorithm is simple and easy to implement. It mainly uses the characteristics of Apriori. However, it has the disadvantages that it needs to scan the database multiple times and generate a large number of frequent itemsets.

(b) FP-growth algorithm. The FP-growth algorithm is an algorithm proposed by Han et al. to mine frequent itemsets from a transaction set without generating candidate sets. The core idea is to transform the database into a FP-tree that stores compressed frequent pattern information, and then mine frequent itemsets in the FP-tree. The steps of the algorithm are to first scan the database, obtain the support count of the item set and then sort all transactions in

descending order of support count. We construct the FP-tree according to the transaction order. Finally, we construct the conditional pattern base of the items from low to high, and then obtain the corresponding conditional pattern tree. Frequent itemset mining is performed recursively on the tree. However, when the data set is large, the constructed FP tree will occupy much memory, resulting in limited operations.

(c) Eclat algorithm. The Eclat algorithm is different from the FP-growth and Apriori algorithm. The Eclat algorithm first needs to convert the data into a vertical format, which is {item:TID_set}, where “item” is the name of the item, and “TID_set" is the set containing the transaction identifier of the item. The idea of the algorithm is to discover the intersection of frequent k + 1 itemset to generate candidate k + 1 itemset. The candidate k + 1 itemset is cropped to generate frequent k + 1 itemset, and then the intersection set is used to generate the candidate k + 2 itemset. Iterating is operated until the itemsets are normalized. The Eclat algorithm only needs to scan the database once, and the calculation efficiency has been greatly improved. However, it has a large dependence on memory. When the data set is large, it is difficult to achieve.

3.1.2 Parallel association rule mining algorithm based on Spark

(1) Spark parallel computing framework

Spark is a fast and general-purpose computing engine designed for large-scale data processing. It is a Hadoop MapReduce-like open sourced universal parallel framework created by UC Berkeley AMP lab.

Compared with the MapReduce computing framework, Spark has the advantages of faster processing of data, support for data query functions, streaming computing modes, a variety of data sources, multiple languages, and strong availability.

Figure 3-1 shows the schematic diagram of parallel tasks of Spark. The cluster resources are managed by the cluster manager. Each worker node contains a job, and multiple tasks are concurrently calculated in the job.

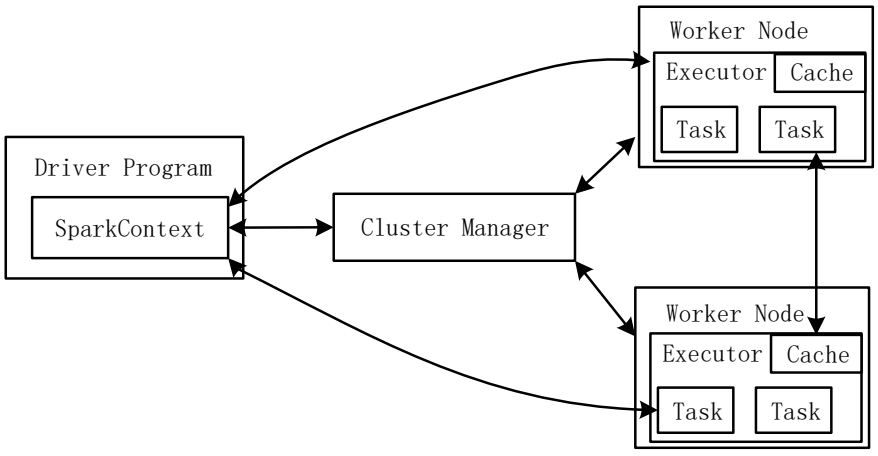


Figure 3-1. Spark parallel principle.

The Resilient Distributed Dataset (RDD) is the core of Spark. When RDD performs iterative operations, it stores intermediate results in distributed memory. Therefore, it can be

directly extracted data from memory in the next data processing, reducing many I/O operations. Thus, iterative algorithm operation efficiency has been greatly improved.

(2) Parallel Apriori algorithm

Qiu *et al*. proposed a classic parallel Apriori algorithm based on Spark RDD computing model in 2014: YAFIM. This algorithm is the first to port the Apriori algorithm to the Spark platform for parallel computing. The main work of the algorithm is divided into two stages. The first stage loads the data set from HDFS to Spark RDD and finds all frequent 1-item sets. The second stage of iteration uses frequent k itemsets to generate frequent (k + 1) itemsets. In this process, a hash tree is used to store the dataset. The operations performed on the RDD in these two phases are shown in Figure 3-2. YAFIM algorithm design is simple and easy to implement. However, there are two disadvantages. (a) The map and reduce tasks need to be started multiple times for counting, which will cause a lot of I/O consumption. (b) Many candidate sets will be generated, occupying storage space. When the data set is large, it will affect the efficiency of the algorithm.

Shaosong Yang and others proposed the SIAP algorithm in 2015. By avoiding repeated scanning of the database, they improved the data structure used by the Apriori algorithm. The new pre-pruning concept was proposed in the second stage. The number of frequent itemsets is reduced, thereby improving the efficiency of the algorithm. The improved data structure of this algorithm is shown in Figure 3-3. However, when the dimensionality of the transaction data is too high, the data storage matrix will be a sparse matrix. This will cause space wastages.

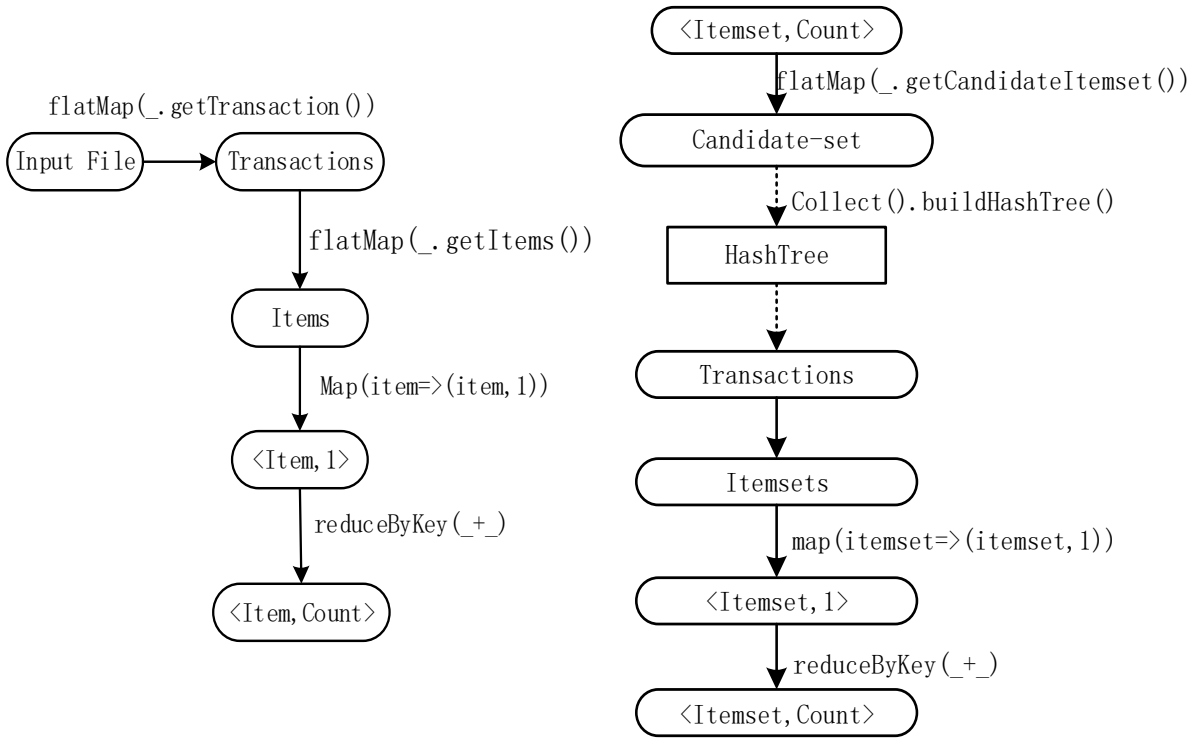


Figure 3-2. The RDD pedigree diagram of YAFIM algorithm.

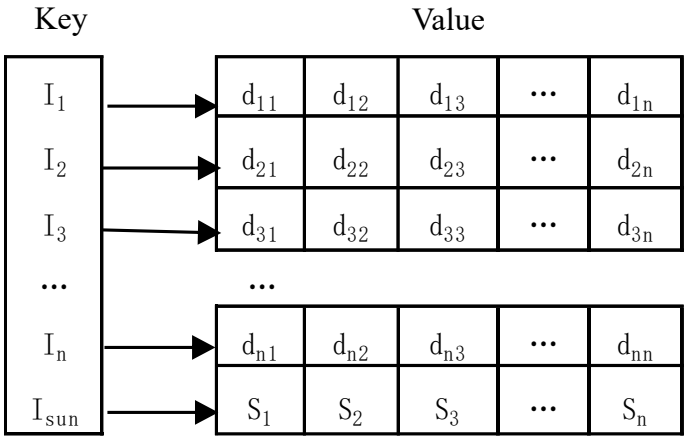


Figure 3-3. Data structure of SIAP algorithm.

Sanjay Rathee *et al*. proposed a new and improved algorithm R-Apriori, which is based on the YAFIM algorithm. This greatly reduces the time and space complexity of generating candidate sets in the second stage. The R-Apriori algorithm is divided into three stages. The first stage acquires frequent 1-item sets. The following improvements are made in the second phase: (I) Removing the step of generating candidate sets; (II) Using Bloom Filter instead of hash tree. The Bloom Filter data structure is used to store and retrieve single item sets, which has a faster rate than a hash tree. Therefore, when the number of single item sets is large, the R-Apriori algorithm is better than YAFIM in terms of time and space complexity. R-Apriori is more scalable as a core function. (III) The third stage is an iterative process of frequent k itemsets generating frequent (k + 1) itemsets. The disadvantage is that the error rate of the bloom filter increases with the increase of the data set. This will affect the efficiency of the algorithm. The improved Spark-based I-Apriori algorithm is similar to the R-Apriori algorithm. It is performed by improving the Bloom filter, eliminating candidate set generation, and reducing the number of database scans. However, the algorithm still has additional I/O disadvantages of the load.

Yonghong Luo et al. proposed a sparse Boolean matrix distributed frequent mining algorithm FISM in 2016. They used transactions as the columns of the matrix and the itemsets as the rows of the matrix to obtain a sparse Boolean matrix. It is easy to obtain the support of the item set from the matrix, and then simply perform an "AND" operation on all itemsets to generate frequent (k + 1) itemsets through frequent k itemsets. Although the FISM algorithm reduces I/O consumption, we only need to scan the database once. When the number of itemsets is huge, the amount of AND operations between the itemsets also becomes huge. The AMRDD algorithm is introduced by Niu Hailing to improve the algorithm. The concept of matrix is used to reduce the number of times the transaction database is scanned. In addition, the algorithm also uses local pruning and global pruning methods to reduce the number of candidate frequent itemsets quantity to improve algorithm efficiency. An improved Apriori algorithm (Spark + IApriori) based on Spark was proposed by Lu *et al.* They optimized the data storage structure to further improve the efficiency of the algorithm. The algorithm flowchart is shown in Figure 3-4. Both the DFIMA algorithm in the literature and the improved algorithm in the literature improve efficiency by transforming the storage structure and eliminating the generation of

candidate sets. However, they have the problems of sparse matrices and large memory consumption.

Krishan Kumar Sethi *et al*. proposed an algorithm HFIM that uses a vertical data set format to reduce the time to scan the database. Then, it generated the number of candidates sets. The format of vertical data is an item/itemset list followed by its transaction ID. The advantage of vertical data sets is that there is no need to scan the entire data set in each iteration. Vertical data sets carry enough information to generate possible candidate sets and calculate their support counts. In each record of the vertical data, the TIDs are sorted in ascending order, which makes the support calculation easier. The support for the k candidate set can be calculated by simply crossing (k - 1) the TID of the item set. However, it will cause problems at the same time, such as excessive calculation and excessive memory consumption.

Wang Qing *et al*. proposed a Spark-based optimization algorithm SP-Apriori in 2016. In the mining process, the frequency is used to indicate the degree of support. The combined strategy is used to obtain the total rule category to obtain the various sets K_{key}. Then, they used the prior nature of the Apriori algorithm to remove the extra item set K_{key}. This compresses the search space to improve the efficiency of the algorithm. The number of nodes required by the algorithm increases as the data set increases. This makes the hardware requirements of the operating environment too harsh.

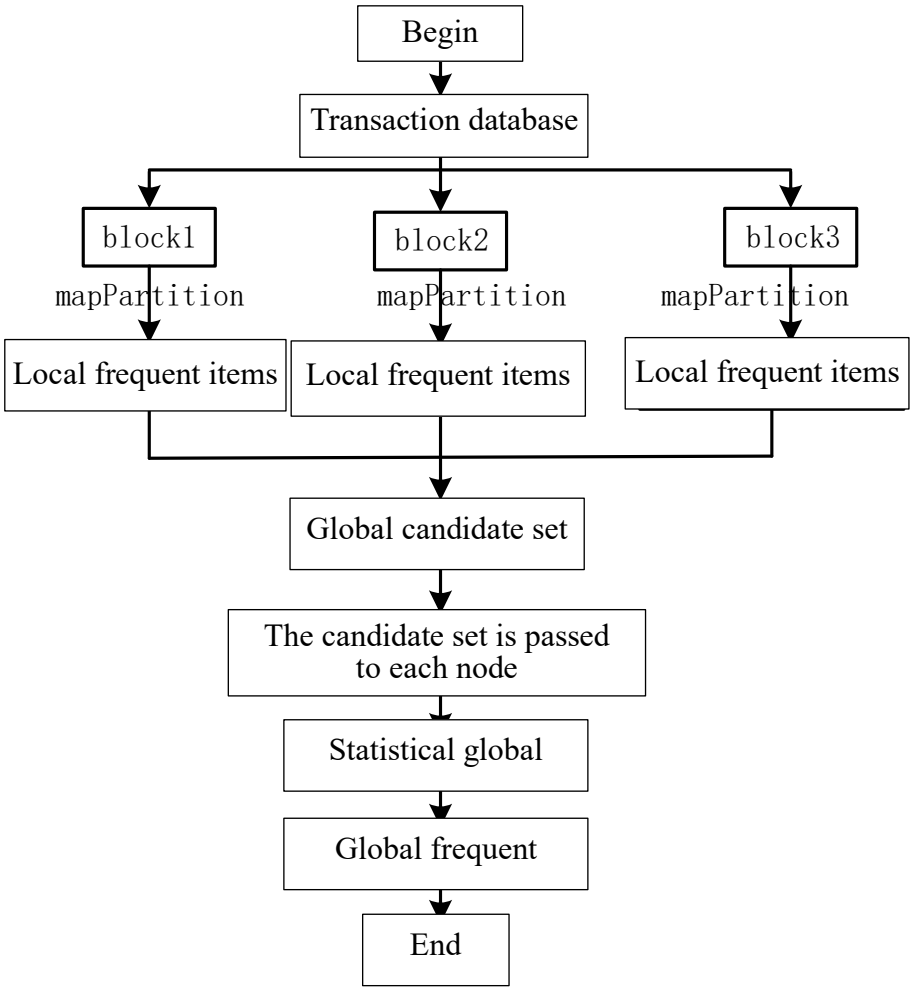


Figure 3-4. The flowchart of Spark + IApriori algorithm.

(3) Parallel FP-growth algorithm

The FP-growth algorithm only needs to scan the database twice without generating candidate sets, store the frequent itemsets of all transactions on the FP-tree, and then traverse the tree to obtain frequent itemsets. Xiang FANG *et al*. proposed an improved Spark-based FP-growth parallel algorithm IPFP-growth in 2016, which realizes the transplantation of the improved PFP-growth algorithm to the Spark computing model for distributed computing. The algorithm is mainly divided into four steps. The first step is to obtain the original information from HDFS and convert it into RDD. Then, a series of operators based on Spark are used to obtain the descending order of frequent 1-item sets. The second step is determined according to the size of the obtained file. The number of groups is divided into several groups according to the rules of balanced grouping. The third step is to perform frequent item set mining for each group separately. The fourth step is to combine the results obtained by each group to obtain the result. Figure 3-5 shows the implementation steps of the IPFP-growth algorithm. However, the data storage structure is not changed. Thus, many conditional FP-Trees are stored in memory during the iteration process. The memory overhead will increase as the data increases.

Zhang Wen *et al*. proposed a load-balancing parallel frequent pattern growth algorithm CWBPFP based on the Unicom weight matrix between items in a transaction. This algorithm is applied to the FP of each working node of Spark by fusing the Unicom weight matrix and constrained subtree. The tree mining process improves the performance of the FP-growth algorithm. However, the algorithm does not fully consider the problem of balanced grouping. This is a waste of space and time.

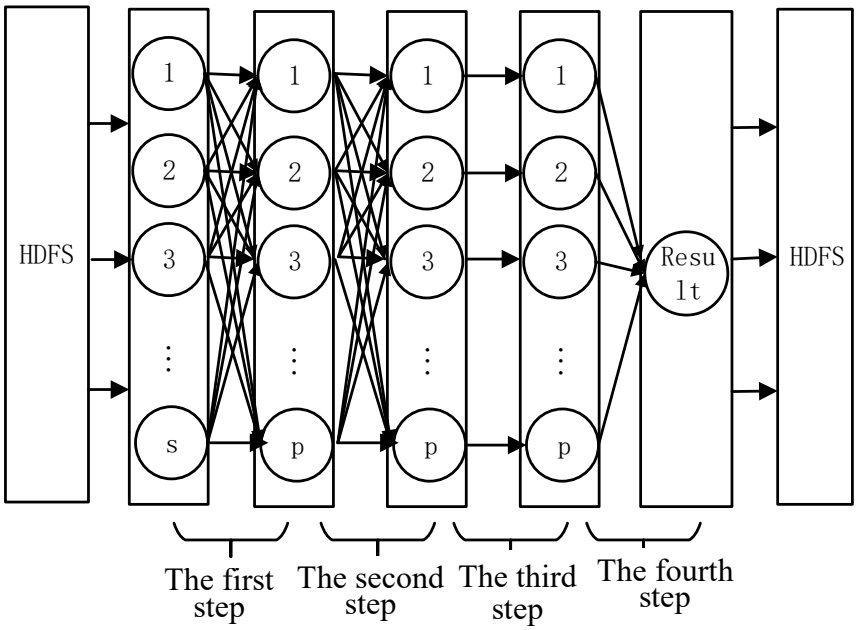


Figure 3-5. The step of IPFP-growth algorithm.

Jiao Runhai *et al*. proposed an improved frequent itemset mining algorithm SMFI for high-dimensional data samples. This algorithm combines the two-way search strategy of the DMFIA algorithm and the parallelized idea of the Spark distributed framework. We have improved deep path search and length-first inspection to improve the efficiency of high-dimensional data

mining. The SMFI algorithm mainly solves three problems in the DMFIA algorithm. The first is to use partition projection to form a parallel subset candidate subset, when the amount of data is large. This projection can avoid data communication between different nodes. In addition, this reduces the repetition rate of neutron records in the partition and memory consumption. The second is that for high-dimensional data to form the maximum frequent candidate set, a frequent item tree consisting of a frequent 1 item set is generated. A deep path recursive search is performed on the frequent item tree for each partition to obtain the maximum frequent partition Candidate set. The third is that the superset test in the loop is directly operated on the most frequent candidate set. We use the method of prefix number and length, sorted in ascending order, and then perform the superset test from top to bottom, short and long. Finally, the remaining candidate sets are the maximum frequent itemset. Due to the need for repeated tests, the algorithm has higher time complexity and redundancy.

Shao Liang *et al*. proposed a FP-growth vertical frequent itemset mining algorithm, referred to as FP-VFIM algorithm. They used the concept of vertical layout of the data set to reduce the time required to scan the data set during each iteration. In addition, the vertical data can be used to calculate the support of the k candidate set by crossing (k - 1) the TIDs of the itemsets. The algorithm is mainly divided into two steps. During the first step, they mainly used flatmap and map functions of Spark to generate frequent 1-item sets on vertical data formats. During the second step, they generated iterative frequent k-item sets. This process uses Spark's broadcast variables and execution. It can reduce the I/O consumption and disk space. Figure 3-6 shows the basic flowchart of the FP-VFIM algorithm.

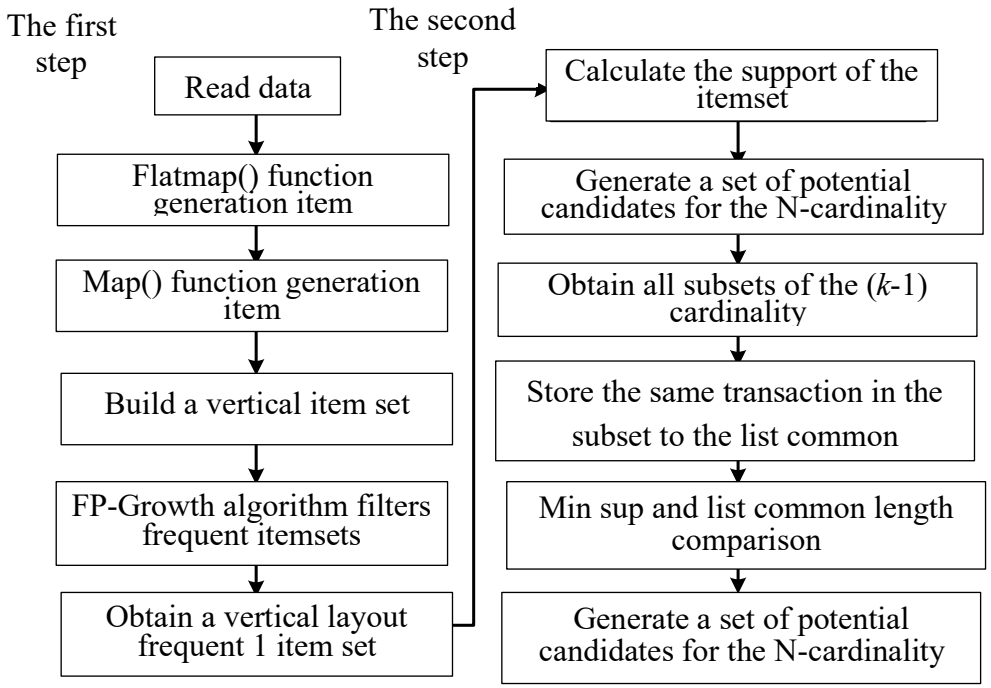


Figure 3-6. The flowchart of the FP-VFIM algorithm.

Gu Junhua *et al*. proposed a new parallel FP-Growth algorithm BFPG in 2018. This algorithm is based on the Spark platform. It improves the F-List grouping strategy for frequent pattern tree

sizes and partition calculations to reduce the load. The sum is evenly divided into each partition. They optimized the data set partitioning strategy by creating a list P-List. This reduces the number of times that the database is scanned to reduce time complexity. The parallel mining time of frequent itemsets depends on the completion time of the last partition. Thus, we try to make each completion time of partition equal during grouping. The BFPG algorithm uses two measurement formulas of $\text{Calculation} = \text{Lg}\left(\text{L}\left(\text{item}, \text{F}-\text{List}\right)\right)$ and $\text{Size} = \text{item\_sup} \times \left(\text{item}_{\text{loc}} + 1\right)/2$ for the optimization strategy of grouping. Considering the time and space complexity from the horizontal and vertical dimensions of FP-Tree, Figure 3-7 is a schematic diagram of the optimization of the grouping strategy.

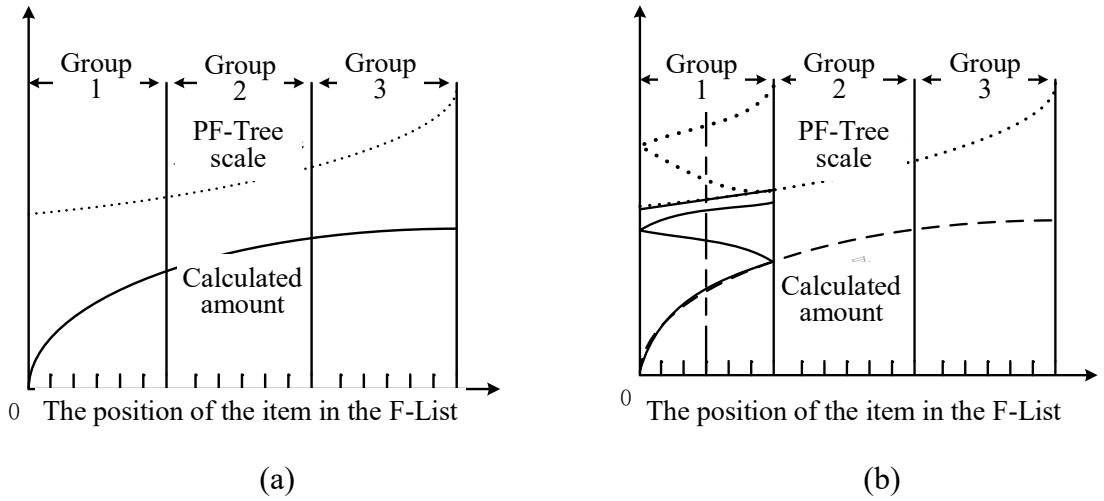


Figure 3-7. Schematic diagram of group optimization strategy. (a) Unoptimized grouping strategy; (b) Optimized grouping strategy.

Shi Lukui *et al*. proposed an improvement of the RPFP algorithm similar to the BFPG algorithm. This improved the algorithm in equalizing group and reducing time complexity. The difference is that the RPFP algorithm is based on the PFP-Growth algorithm. It adds a balanced group optimization algorithm, which optimizes the linked list structure algorithm. Then, it adds a hash expression to the fast access element address, thereby reducing time complexity. However, due to the addition of a linked list to the frequent pattern tree, the memory generated by iterative mining in the FP-tree increases. This is a problem that requires to be studied in the next step. The S-FPG algorithm proposed by Gassama *et al*. implements parallel computing on the Spark platform by changing the data structure form and reducing the number of iteratively generated frequent item sets. However, it still does not reduce the memory requirements. The algorithm DFIMA was proposed by Zhang F *et al.* They used a matrix-based pruning method to reduce the generation of candidate sets, and implement parallel operations in conjunction with the Spark platform. However, the algorithm does not conduct in-depth research on data storage formats. Lu Ke *et al.* improved the mining efficiency by optimizing the support count and grouping of the FP-growth algorithm. Since the results of the iterative calculations are stored in memory, they cannot meet the requirements of massive data.

(4) Parallel Eclat algorithm

The transaction data form of Apriori algorithm and FP-growth algorithm process is the

horizontal format. However, the transaction data form of Eclat algorithm is different, which is the vertical format. Feng Xingjie *et al*. proposed a Spark-based parallel Eclat algorithm called SPEclat. By changing the data storage method, the data is divided into different computing nodes by prefixes. The search space of the data is compressed to achieve parallel computing. The algorithm is mainly divided into three stages to complete the parallel calculation. In the first stage, data is read and saved to obtain a frequent 1-item set. The traditional Eclat algorithm stores transactions in the form of (ItemSet, TidSet). However, when the data set is large, the intersection operation of TidSet will take a lot of time. Thus, the algorithm stores the data in the way of (ItemSet, BitSet). The count becomes a BitSet AND operation, avoiding the calculation of a large amount of data. In the second stage, the obtained frequent itemsets are divided iteratively by prefixes and distributed to different computing nodes. In the third stage, itemsets with the same prefix are iterated in each computing node in parallel from bottom to top, generating frequent (k + 1) itemsets with frequent k itemsets. Finally, the frequent (k + 1) itemsets are re-divided, and the SPEclat algorithm is iteratively called to mine frequent itemsets until frequent items are no longer generated. However, the algorithm is not evenly divided when it is divided into different computing nodes. This will cause memory occupation.

(5) Comparison of efficiency analysis among parallel algorithms

(a) Time and space complexity analysis of parallel algorithms. The parallel computing model is mainly to solve the problem of the time complexity and space complexity of the iterative process of the association mining algorithm. Analyzing and comparing the efficiency of the above typical representative algorithms provides a reference for subsequent algorithm improvement. The execution time of the input and output in "Reduce" in the Spark computing framework is the same as the input and output time of traditional association mining. In the mapper, analysis is required. We assume that X is the number of transactions, M is the number of map tasks, and g is the average number of each transaction, f is the number of frequent 1-item sets, t is the time to search for an element in the hash tree, and l is the filter time to search for an element in the Bloom. For the Apriori algorithm, the total time of the second iteration is the sum of the time to generate the candidate set, the time to store the candidate set, and the time used by the mapper. The time to generate the candidate set is the sum of the time of connection and the time of pruning. For the YAFIM algorithm, the time to generate the candidate set is $\frac{f(f-1)}{2}+\frac{2f(f-1)}{2}=\frac{3f(f-1)}{2}$, the time to store the candidate set in the hash tree is $\frac{2f(f-1)}{2}\approx O(f^2)$, and the time used by the mapper is $\frac{X}{M}\times\frac{tg(g-1)}{2}\approx O\left(\frac{X}{M}.g^2\right)$. Thus, the time complexity of YAFIM is $\mathrm{O}\left(f^2+\frac{X(\mathrm{t}-l)}{M}.g^2\right)$. For the R-Apriori algorithm, the total time = time to store frequently 1 item set to Bloom filter + time to pruning + time to generate item set pairs, *i.e.*, $f+\frac{X}{M}\times\frac{lg(g-1)}{2}\approx O\left(f+\frac{X}{M}.l.g^2\right)$. The time complexity of the YAFIM algorithm minus the time complexity of the R-Apriori algorithm is equal to $\mathrm{O}\left(f^2+\frac{\mathrm{X}}{\mathrm{M}}.(\mathrm{t}-l).g^2\right)$. The time complexity of the R-Apriori algorithm is lower than that of the YAFIM algorithm. We compare the spatial complexities between these two algorithms. The

same “Reduce” has the same input and output for the algorithms, and the space occupied by the mapper are different. The spatial complexity of the Apriori algorithm is mainly the sum of the space occupied by storage of the candidate set and the data structure of the candidate set. Let *b* and *c* be the space occupied by a pair of candidate sets and a single item set, respectively. The space for storing candidate sets in the YAFIMA algorithm is $\frac{bf(f-1)}{2} \approx O(bf^2)$. The space complexity for creating a hash tree is $\frac{bf(f-1)}{2} + \frac{cf(f-1)}{2} \approx O((b+c)f^2)$. The sum of the space complexity of the R-Apriori algorithm is the space occupied by storing frequent single-items in the Bloom filter $\frac{f(f-1)(2b+c)}{2} \approx O((b+c)f^2)$. The space complexity of R-Apriori algorithm is much lower than that of YAFIM algorithm. The same analysis of the IABS algorithm results in a total time complexity of $O(f + \frac{T}{M} \times \frac{lg(g-1)}{2})$ and a space complexity of *cf*. The time and space complexity of the IABS algorithm is lower than that of YAFIM, respectively. The total time complexity of the I-Apriori algorithm is $f + \frac{X}{M}.\frac{g(g-1)}{2} \approx O(f + X/Mlg^2)$, and the total space complexity is $O(cf)$.

For the parallel FP-Growth algorithm, the number of itemsets contained is *k*, and the number of items contained in the largest transaction is *m*. The time complexity of FP-VFIM algorithm step 1 is the time required to scan the transaction database once and access each item to build a vertical layout. All *k* items will go through the FP-Growth step. Thus, the process is complicated in the worst case. The degree is *k*, and step 2 is an iterative process. During the $O(Xm+k)$ iteration, all transactions in the frequent item set are scanned and an item list is created. A potential candidate set is generated from the item list, and each candidate is further accessed. We set a subset list of $(i-1)$. The maximum time complexity required is $O(F \times m^{\min(i,m-i)} \times i \times v \times c)$, where *F* is the number of transactions for frequent itemsets, *v* is the number of itemsets in share_data, and *c* is the largest item ID in share_data counting. The total time complexity of the FP-VFIM algorithm is the sum of the time complexity of these two steps. Its space complexity is $O(F \times m + 2^k + i + v + v \times c)$. For the SPEclat algorithm, although the data format is a vertical format, it is still the same method to perform the analysis of time complexity and space complexity. The time complexity of the SPEclat algorithm is the sum of the time complexity of the three stages of $O\left(f^2 + \frac{X(t-l)}{M}.g^2\right)$.

(b) Overall analysis and comparison of algorithms. The improvement of the association rule mining algorithm is mainly to optimize the original data set format, the data structure used by the algorithm, the number of candidate sets generated during the mining process, and the number of times the database scanned during the mining process. The comparison among the above types of parallel algorithms is shown in Table 3-1. Based on the analysis of the time complexity and space complexity of the various parallel algorithms in the previous section, in the parallel Apriori algorithm, except that the R-Apriori algorithm does not generate a candidate

set during the mining process, other algorithms will generate a candidate set. The time complexity and space complexity of R-Apriori algorithm are better than that of other parallel Apriori algorithms. SIAP algorithm and HFIM algorithm use both horizontal and vertical data formats, reducing the time and space complexity of the algorithm during the generation of frequent 1-item sets. The number of candidate sets is not reduced, when association rules are generated. Therefore, the overall complexity of the algorithm is higher than that of the R-Apriori algorithm, lower than that of the YAFIM algorithm. In terms of the data structure used, the R-Apriori algorithm uses bloom filters, SIAP, FISM, AMRDD, Spark + IApriori, SP-Apriori, DPBM and other algorithms use matrices, YAFIM uses hash trees, and HFIM uses the original data structure. Under the same circumstances, algorithms using special data structures have obvious advantages in terms of time complexity and space complexity in generating frequent *k*-itemsets over algorithms without special data structures.

Table 3-1. Comparison of various parallel algorithms.

Variant type	Algorithm name	Data format	Used Data structure	Whether to generate candidate sets	Scanning database times
Parallel Apriori algorithm	YAFIM	Horizontal	Hash tree	Yes	*k*
	SIAP	Horizontal and vertical	Boolean matrix	Yes	1
	R-Apriori	Horizontal	Bloom filter	No	1
	FISM	Horizontal	Boolean matrix	Yes	1
	AMRDD	Horizontal	Boolean matrix	Yes	≪k
	Spark+IApriori	Horizontal	Boolean matrix	Yes	≪k
	HFIM	Horizontal and vertical	no	Yes	≫3
	SP-Apriori	Horizontal	Distribution matrix	Yes	2
	DPBM	Horizontal	Boolean matrix	Yes	1
Parallel FP-Growth algorithm	IPFP-growth	Horizontal	Hash tree	No	2
	CWBPFP	Horizontal	FP-tree	No	2
	SMFI	Horizontal	Suffix tree	Yes	2
	FP-VFIM	Vertical, horizontal	FP-tree	Yes	≪k
	BFPG	Horizontal	FP-tree	Yes	2
	RPFP	Horizontal	Hash table, FP-tree	Yes	2
	S-FPG	Horizontal	FP-tree	Yes	2
Parallel Eclat algorithm	SPEclat	Vertical	Distribution matrix	Yes	≫3

Among algorithms that use special data structures, the Bloom filter used by the R-Apriori algorithm has advantages over the Boolean matrices used by other algorithms. Because Bloom filters can represent the complete set, not any other data structure, it has more advantages in

time and space. The algorithms using special data structures are less than *k* times in terms of scanning the database, and even only need to scan the database once. This significantly reduces the memory footprint and time of scanning the database. In the parallel FP-Growth algorithm, since the FP-Growth algorithm itself only needs to scan the database twice, the algorithm efficiency is better than Apriori algorithm. The IPFP-growth and CWBPFP algorithms do not generate candidate sets and are superior to other similar algorithms in iterative calculations. The SMFI and RPFP algorithms use special data structures to make the algorithm more efficient in the iterative mining of frequent patterns. The FP-VFIM algorithm is the only algorithm that uses the vertical data format. It has an advantage over similar algorithms in the process of discovering frequent itemsets. The implementation of the *k* parallel association rule algorithm on the Spark platform uses a key-value-based storage system. Its query speed is high, the amount of stored data is large, and high concurrency is supported. However, complex conditional queries cannot be performed. Therefore, the influence of data structure storage used by association rule mining algorithms on mining efficiency is very important. As the number of transactions becomes greater and greater, the vertical data format can show advantages. The number of candidate sets is mainly occupied by memory, which will affect the algorithm space complexity. The number of database scans will affect the algorithm time complexity. Therefore, when the algorithm is improved, it is necessary to consider these factors, and combine the characteristics of the actual data structure to find the optimal solution.

3.1.3 Research prospect of parallel association rule mining algorithm based on spark

As the amount of data increases, traditional mining algorithms for association rules can no longer meet the requirements. The association rule mining algorithms that can perform parallel and distributed operations are necessary to solve this problem. Spark is specially designed to solve the calculation problems of big data. It is suitable for iterative operations and is widely used in the field of association rule mining. The frequent pattern mining algorithm studied in this chapter is mainly improved by the following methods. (I) Changing the data structure to speed up the search of frequent itemsets; (II) Reducing the number of candidate item sets; (III) Optimizing balanced grouping to achieve computing load balancing; (IV) Reducing I/O traffic and storage footprint.

The algorithms studied in this chapter are all classic association rule algorithms migrated to the Spark platform. However, as the amount of data expands, the number of association rules generated will increase. Therefore, the focus of future research work may be mainly on the following aspects.

(1) The Spark platform is a memory-based operation and does not have a large amount of I/O consumption. However, because of the increasing data set in reality, its growth rate far exceeds the development speed of memory capacity. The association rules are mined. The process is not a one-step calculation to obtain the results. It is necessary to continuously search and calculate the association rules in the original data set. The memory required during the mining process is greater than the original data. When the maximum memory is reached, parallel computing will cause data loss and repeated search problems. Therefore, solving the

relationship between data size and computer memory size is one of the future research directions. Compressing the original data can effectively solve it. Thus, the data compression technology is a hot research direction in the future.

(2) The new data in the information age is increasing every moment. Thus, the results obtained by the association rule mining algorithm should be updated. Therefore, the method of the algorithm must also be improved as the data is updated. The mining and calculation of parallel association rules are mainly performed in memory. If new data is added and the calculation needs to be restarted, it will cause a large memory loss. Therefore, designing high-performance new algorithms to deal with changes between new and old data is one of the difficulties that need to be overcome in the future.

(3) The design of the algorithm is still the most important part of association rule mining. How to achieve the optimal running algorithm needs to consider the operating conditions of each node. Regardless of the time or space complexity, how to change the algorithm to the optimal or local optimal algorithm has always been the research direction. In addition to the basic pattern mining algorithm research, Multi-layer and multi-dimensional patterns mining algorithms also need to be implemented on parallel platforms. The next step is to investigate association pattern mining of data types such as structural patterns, spatial patterns, images, videos, multimedia, and network patterns.

3.2 Study and example of the use of WD-Get rules algorithm

Facing various massive and complex data systems of TCM information, data mining is a powerful tool. We can analyze and sort huge data to realize the reasonable use of effective information. This technology can be applied to almost all aspects of TCM research, especially in the research of prescription compatibility law, which is a hot issue of TCM data mining. Data mining can provide a reliable method for the scientific and reasonable analysis of the compatibility laws contained in the target prescription data from multiple aspects, such as the frequency of prescription medication and efficacy combination. The results are significance for clinical use and new drug development. TCM defines kidney deficiency as deficiency of kidney essence and qi, which can be divided into kidney yang deficiency, kidney yin deficiency, kidney essence and kidney qi deficiency. In TCM theory, the kidney is the root of the viscera, the root of the twelve classics, and the congenital root lies in the kidney.

In TCM, the kidney-zang elephantology considers that the kidney has the essence, the main water, the main bone, the main qi, and the second stool. When the human kidney is deficient, these functions will be affected. The internal organs will be imbalanced, and health will be threatened. Modern scientific research confirms that kidney deficiency can cause disorders in the body's neuroendocrine function, reproductive function, hearing and bone metabolism. This can reduce the body's immunity. Nowadays, because of the influence of unhealthy living rules and inappropriate environmental factors, kidney deficiency is extremely common. TCM has a history of thousands of years for treating kidney deficiency. It has rich experiences and remarkable results. This chapter uses data mining techniques to explore the TCM prescriptions for treating kidney deficiency in medical books of different generations, and summarizes the law of prescriptions. This can provide reference for clinical treatment of kidney deficiency.

3.2.1 Research data and methods of WD-Get rules algorithm

(1) Prescription collection

Using the three databases of "China TCM Search System", "China How Net" and "Medical Intelligence Data" to collect prescriptions for treating kidney deficiency, excluding prescriptions containing only a single drug, a total of 192 effective prescriptions are collected. We standardize the TCM names in accordance with the "Pharmacopoeia of the People's Republic of China" (2015 Edition). The ingredients that are not in the scope of the Pharmacopoeia remain the original names, such as "pig kidney" and "green salt". Then, a database of prescriptions for treating kidney deficiency is established, and then simulated by Matlab.

(2) Improved association rule algorithm

(a) Association rule mining. Association rule mining is to find all frequent itemsets in the database that are greater than or equal to the minimum support specified by the user. We use frequent itemsets to generate the required association rules, and filter out the strong association rules according to the minimum credibility set. As an important branch of data mining algorithms, association rule has appeared in literature research in the field of TCM for the first time since 2002, and have more than ten years of application experiences. However, TCM information data is "massive, non-linear, unstructured", and traditional data mining techniques may be difficult to solve these problems. Further optimization of the algorithm is required to improve the efficiency and interest of the classic association rule mining algorithm to improve the algorithm's processing speed and ability of TCM information data. This chapter designs an improved algorithm (WD-Get Rules) that combines depth and width search. This algorithm has higher mining efficiency than traditional algorithms in processing large-scale data sets. Therefore, this algorithm is used to analyze association rules of 192 prescriptions for kidney deficiency.

(b) Improved WD-Get Rules algorithm that combines depth and width search. The idea of the WD-Get Rules algorithm is to first use a width-first strategy to find the set of followers H that can make strong association rules. Each set is a 1-item set, with H as the benchmark and a set enumeration. The idea of the association rule generation of set-enumeration tree algorithm (GRSET) is to perform the deep search. The difference is that when performing a deep search, the consequent of the rule is only derived from the elements contained in H. Thus, the elements that cannot be consequent of the rule are skipped directly. If the itemset is a frequent k ($k > 2$) itemset and the number of elements in H is greater than 1. Then, a deep search is performed, otherwise the next frequent itemset is operated. Assuming frequent itemsets $L = L_1 \cup L_2 \cup \cdots \cup L_k$, where L_k represents a set of frequent k itemsets, and each frequent itemset is ordered. The specific flow of the WD-Get Rules algorithm is shown in Table 3-2(a).

Table 3-2(a). WD-Get rules algorithm flow.

Algorithm: WD-Get rules(*L*,min_conf,TIDcount,supData)

Input: frequent item set L, minimum confidence threshold min_conf, total transaction number TIDcount, support count SupData
Output: All association rules F generated by L

F is initially set as {"rule", "support", "confidence"}
Take out the number of rows of L as m
For i = 2:m, starts with a frequent 2 item set
Take out all the frequent i items set as L_1
Obtain the number of rows of L_1 as m_1
For j = 1:m_1
    Take out the jth subset in frequent L_1 as freqSet
    % do a width search first
    Initialization can be done after the rule set H is empty
    Splice all the items in the freqSet into a string S
    For r = 1:i
        Take the rth element in freqSet as a
        Save the string form of a with behind
        Save the difference between freqSet and a to C
        Splicing the elements in C into front
        Find the rule confidence conf = sup(S)/sup(front)
        If conf ≥ min_conf
            Obtain the rule f_1 = front ≥ behind
            Rule support sup = supp(S)/TIDcount
            Append f_1, sup, conf to F with F = [F;f1,sup,conf]
            Append element a to H
        End
    End
    % completes the width search
    Take out the number of elements of H num
    If i > 2 && num > 1
        % deep search
        For jj = 1:num
            Take out the jjth element in H
            S = jj+1;
            If S ≤ num
            F = getall_new(*h*,*s*,freqSet,min_conf,TIDcount,supData,*i*,F,*S*,H)
            End
        End
    Else
        Continue
    End
    End
    End

The getall_new algorithm is shown in Table 3-2(b).

Table 3-2(b). Getall_new algorithm.

Algorithm: getall_new(h,*s*,freqSet,min_conf,TIDcount,supData,*k*,F,S)

Input: the latter set *h*, the starting position *s* of the item to be connected in freqSet (the position of the last item in $h + 1 \le s \le k$), freqSet is the frequent item set, the minimum confidence threshold min_conf, the total number of transactions TIDcount, support count SupData, *k* is the number of items contained in freqSet, rule set F, all items in freqSet are spliced into string S

Output: All association rules generated by freqSet, the latter

Obtain the number of element *h* with *n*

If $n < k - 1$

  For $j = s:k$

    Obtain the *j*th element of freqSet with *a*

    Obtain the rule predecessor C = freqSe − *h* − *a* (where "−" stands for difference set operation)

    Splice all items in C into a regular front-end string front

    Find the rule confidence conf = supData(S)/supData(front)

    If conf ≥ min_conf

      Splice *h* + *a* into a regular post-string followed (where "+" stands for union)

      Obtain the rule f_1 = front => behind

      Rule support sup=supp(S)/TIDcount

      Append f_1, sup, conf to F with F=[F;f_1,sup,conf]

      If $j < k$

        Getall(*h*+*a*,*j*+1,freqSet,min_conf,TIDcount,supData,*k*,F,S)

      End

    End

  End

End

3.2.2 Research results and analysis of WD-Get Rules algorithm

(1) Source analysis of prescriptions

The 192 prescriptions for treating kidney deficiency collected in this chapter are from 88 medical books. 18 books that provide more than three prescriptions are selected as important sources (Table 3-3). Among them, both "General Collection for Holy Relief" and "Prescriptions for Universal Relief" provide 10 prescriptions for treating kidney deficiency. "General Collection for Holy Relief" is the earliest comprehensive medical work compiled by the government organization in China, and has profound historical value. "Prescriptions for Universal Relief" is the largest book of TCM prescriptions in the history of Chinese formula editing, and has been listed as one of the key ancient books for medical treatment. Using data mining technology to mine the treasure trove of TCM culture, and to study ancient medical prescriptions will be significance for clinical treatment.

(2) Frequency analysis of TCM in the prescription library

The frequencies of a total of 343 TCMs in 192 prescriptions are counted. The type of TCMs used no less than 3 times is 123, as shown in Table 3-4. The result shows that the most commonly used TCMs in the treatment of kidney deficiency are Poria, Shu Di Huang, Yam, Eucommia, Aconite, *etc*. Poria and yam have a flat taste and sweetness, which can replenish yin and yang, and can nourish qi and refine essence. Guijing shows that they can act on multiple organs of the lung, spleen and kidney. Therefore, these two TCMs can be compatible with other

ones to treat various types of kidney deficiency. In addition, they can strengthen the spleen, kidney yang, and enhance the efficacy of kidney tonic. Rehmannia glutinosa has been a good medicine for nourishing yin and kidney since ancient times. Because the metaplasia of kidney yang and kidney qi are based on Yin xue jin jing, the syndrome of deficiency of kidney qi and kidney yang is often combined with rehmannia glutinosa to nourish yin and qi. Thus, a variety of prescriptions appear cooked rehmannia. Eucommia has mild temperature and aconite has fever temperature. Both can warm the kidneys and boost the yang. In addition, some ancient recipes describe the dietary supplements, such as the application of "sheep kidney" and "pig kidney". The diet therapy based on the form-filling form embodies the dialectical thinking of analogy in TCM, and there are similar discussions in some ancient books. For example, in the "Tiaoji Diet Argument" written by Zhang Mu in the Qing Dynasty, there is a theory of "heart tonic" and "kidney tonic". In the Tang Dynasty, Sun Simiao had the academic view of "Replenish the internal organs with internal organs". Although the theory of "Complement form with form" is indeed clinically confirmed, it still needs to be treated with syndrome differentiation and scientific treatment. Modern medical research shows that it is not suitable to use animal viscera with high cholesterol content in patients with severe kidney deficiency. From the frequency of use of "sheep kidney" and "pig kidney", this therapy has gradually withdrawn from the kidney deficiency treatment system. The analogy of "Tonifying kidney with kidney" should be applied clinically on the premise of dialectical treatment.

Table 3-3. Important sources for prescriptions of treating kidney deficiency.

The source of prescriptions	The number of prescriptions
"General Collection for Holy Relief"	10
"Prescriptions for Universal Relief"	10
"The Six Books of Medicine"	9
"Sheng Hui"	8
"National Prescription Collection of Traditional Chinese Medicine"	8
"Shou Shi Bao Yuan"	6
"Complete works of Jingyue"	6
"Elementary Medicine"	5
"Medical Records"	4
"Taiping Huimin Heji Jufang"	4
"Three Causes"	4
"Collecting Record of Differentiation of Symptoms and Signs"	4
"Principle of Correct Diet"	3
"Wai Tai"	3
"Prescriptions Assigned to the Three Categories of Pathogenic Factors of Diseases"	3
"Folk Prescription"	3
"Ji Feng"	3
"Ancient and Modern Medical Guide"	3

Table 3-4. TCM with the frequency of medication no less than three times.

TCM	Frequency	TCM	Frequency
Poria	68	Bupleurum	6
Rehmannia	61	Fenugreek	6
Yam	52	Peach kernel	6
Eucommia	51	Fragrant	6
Achyranthes	49	Puzzle	6
aconite	48	Chrysanthemum	6
Angelica	48	Papaya	6
Ginseng	48	Chrysoidine	6
Licorice	46	Acorus calamus	5
Cuscuta	45	Pu Shizhi	5
Schisandra	42	Fructus Toosendan	5
Atractylodes macrocephala	41	Cong bai	5
Astragalus	37	Codonopsis	5
Medical dogwood	37	Rehmannia	5
Psoralen	37	Pubescent angelica	5
Cistanche	30	Scutellaria	5
Cinnamon	29	Campanulaceae	5
Morinda	26	Sulfur	5
Wolfberry	25	Mastic	5
Alisma	24	Amomum	5
Dried ginger	21	Ginger	5
Yuanzhi	21	Evodia rutaecarpa	5
Woody	21	Pinellia	5
Rehmannia	21	Heshouwu	5
Teasel	20	Atractylodes	5
Windproof	19	Cubeb	4
White Peony Root	18	Star anise	4
Velvet	18	Clove	4
Cumin	18	Woodwardia Japonica	4
Liriope	17	Patchouli	4
Dendrobium	17	Lotus seed	4
Cork	15	Antler cream	4
Peony skin	15	Myrrh	4
Chuanxiong	14	Mulberry	4
Magnet	14	Gastrodia	4
Notopterygium	12	Sheep kidney	4
Pepper	12	Bijie	4
Eaglewood	11	Antler gum	4
Keel	11	Fine salt	4
Tangerine	10	White salt	3
Common anemarrhena	10	Azurite	3
Nutmeg	10	Tiger bone	3
Akasaka	10	Amber	3
Asparagus	9	Talc	3
Oyster	9	Nepeta	3
Seed of oriental arborvitae	9	Salvia	3
Raspberry	8	Gentiana	3
Magnolia officinalis	8	Green scorpion	3
Caltrop	8	Scorpion	3
Grassleaved sweetflag rhizome	8	Mulberry parasitic	3
Asarum	8	Astragali complanali	3
Suanzaoren	8	Cnidium	3
Psyllium	7	Musk	3
Turtle board	7	Combined spicebush	3
Scorpion	7	Poppy shell	3
Green salt	7	Cinnabar	3
Cimicifuga	7	Purpura	3
Clam shell	7	Cassia	3
Tamping	7	Acacia	3
Ejiao	6	Pig kidney	3
Angelica	6	Tiger bone	3
Areca	6		

(3) Analysis of natural flavor and channel tropism of TCM in the prescription library

In the 2015 edition of the “Pharmacopoeia of the People's Republic of China”, the

medicinal distribution of each type of TCM is summarized. Due to the numerous medicines, the distribution of 20 TCMs with high frequency of occurrence is listed, as shown in Table 3-5. In addition, the 85 TCMs with the highest frequency of occurrence (≥ 5) in the prescription library are selected. They are c the types of them are calculated according to the four Qi (drug property), five flavors, and channel tropism distribution. The number of frequency and the proportion of the results are shown in Table 3-6, Table 3-7 and Table 3-8.

Table 3-5. Distribution of drug property of commonly used TCMs in the treatment of kidney deficiency.

TCM	Frequency of occurrence	Drug property	Flavor	Channel tropism
Poria	68	Flat	Sweet	Heart, lung, spleen, kidney meridian
Rehmannia	61	Lukewarm	Sweet	Liver and kidney meridian
Yam	52	Flat	Sweet	Lung, spleen, kidney meridian
Eucommia	51	Warm	Sweet, Little spicy	Liver and kidney meridian
Achyranthes	49	Flat	Bitter, sweet, sour	Liver and kidney meridian
Ginseng	48	Warm and flat	Sweet, bitter	Spleen, lung, heart meridian
Angelica	48	Warm	Sweet, spicy	Liver, heart, spleen meridian
Licorice	46	Warm	Sweet	Heart, lung, spleen, stomach meridian
Dodder	45	Flat	Spicy, sweet	Liver, kidney, spleen meridian
Aconite	45	Hot	Spicy, sweet	Heart, kidney, spleen meridian
Schisandra	42	Warm	Sour, sweet	Lung, heart, kidney meridian
Atractylodes macrocephala	41	Warm	Bitter, sweet	Spleen, stomach meridian
Astragalus	37	Warm	Sweet	Lung, spleen, liver, kidney meridian
Hawthorn	37	Lukewarm	Sour	Liver and kidney meridian
Psoralen	37	Warm	Bitter, spicy, warm	Kidney, spleen meridian
Cistanche	30	Warm	Sweet, sour, salty	Kidney, large intestine meridian
Cinnamon	29	Hot	Spicy, sweet	Kidney, spleen, heart, liver meridian
Morinda	26	Lukewarm	Sweet, spicy	Kidney, liver meridian
Wolfberry	25	Flat	Sweet	Liver and kidney meridian
Alisma	24	cold	Sweet	Kidney, bladder meridian

Table 3-6. The proportion of drug property in the prescription database.

Drug property	The type of TCM	Frequency of occurrence	Proportion
Warm	44	796	54.48%
Cold	20	222	15.20%
Flat	17	315	21.56%
Hot	5	112	7.67%
Cool	1	9	0.62%

Note: Each TCM may involve multiple drug properties. Thus, the type and frequency are counted repeatly.

Table 3-7. The proportion of drug flavor in the prescription database.

Drug flavor	The type of TCM s	Frequency of occurrence	Proportion
Sweet	44	1011	69.20%
Spicy	42	613	41.96%
Bitter	36	486	33.26%
Sour	11	217	14.85%
Salty	5	78	5.34%

Note: Each TCM may involve multiple drug flavors. Thus, the type and frequency are counted repeatly.

Table 3-8. Proportion of channel tropism of drugs in the prescription library.

Channel tropism	The number of TCM	Frequency of occurrence	Proportion
kidney	56	1316	78.20%
spleen	39	992	59%
liver	35	816	48.50%
heart	28	651	38.70%
lung	22	573	34.10%
stomach	13	167	9.90%
bladder	8	62	3.70%
the large intestine	7	59	3.50%

Note: Each TCM may involve multiple channel tropism. Thus, the type and frequency are counted repeatly.

According to the analysis of the natural flavor and channel tropism of TCM with high frequency of occurrence in the prescription library, we find that the drug properties for treating kidney deficiency are mainly “warm”, “flat” and “cold”, the drug flavors are mainly “sweet” and “spicy”, and channel tropism is attributed to the “kidney meridian” and “spleen meridian”. According to the relationship between the drug property and efficacy, the TCMs with the property of “warm-spicy-kidney” have the effect of nourishing the kidney and boosting the yang, as well as strengthen the muscle and bone. The TCMs with the property of “flat-sweet-kidney” have the effect of reinforcing kidney and spleen. The TCMs with the property of “cold-sweety-kidney” have the effect taking yin and moistening zao. The TCMs with the property of “warm-sweety-spleen” have the effect of nourishing the spleen and qi. The TCMs with the property of “flat-sweet-spleen” have the effect of reinforcing spleen and nourishing qi. Therefore, according to the drug property, most of the TCMs for treating kidney deficiency directly affect the kidney or both the spleen and kidney. According to the characteristics of drug flavor, the prescriptions for treating kidney deficiency are mostly reinforcing the kidney with “spicy” flavor, nourishing qi and kidney with “sweet” flavor, and the use of both “spicy” and “sweet” flavors can achieve the effect of “xin gan hua yang”. We use “warm” drugs to improve the kidney’s pushing, warming, gasification function, “cold” drugs to nourish the kidneys, clear the yin and fire, promote the body to nourish and restrict the heat. The application is not restricted by the type of kidney deficiency syndrome.

(4) Analysis of the rule of the formula for treating kidney deficiency

(a) Analysis of core pairs of TCM. In the database of treatment of kidney deficiency prescription, a total of 28 groups of core drug pairs with a frequency of ≥20 times are excavated. The efficacy of each group of drugs on the kidney is summarized according to the query of the

Chinese Pharmacopoeia, as shown in Table 3-9.

Table 3-9. The core pair of drug, frequency and efficacy of treating kidney deficiency.

Core pair of drugs	Frequency	Efficacy
Yam, Poria cocos	34	Nourishing kidney-qi
Rehmannia, Poria cocos	32	Tonifying kidney-yin
Ginseng, Poria cocos	27	Nourishing kidney-qi
Rehmannia, Yam	27	Tonifying kidney-yin, Replenishing kidney-essence, Nourishing kidney-qi
Licorice, Ginseng	26	Strengthening spleen-qi
Achyranthes, Poria cocos	25	Tonifying kidney-yin, Clearing deficiency fire
Rehmannia, Schisandra	25	Tonifying kidney-yin, Nourishing kidney-qi
Achyranthes, Rehmannia	24	Tonifying kidney-yin
Yam, Hawthorn	24	Replenishing kidney-essence, Nourishing kidney-qi, Warming kidney-yang
Hawthorn, Rehmannia	23	Replenishing kidney-essence, Nourishing yin and supporting yang
Angelica, Ginseng	23	Invigorating qi and promoting blood circulation
Angelica, Achyranthes	23	Nourishing kidney-yin, Nourishing blood and promoting blood circulation
Hawthorn, Poria cocos	22	Replenishing kidney-essence, Warming the kidney-yang
Schisandra, Poria cocos	22	Nourishing kidney-qi
Eucommia, Rehmannia	22	Nourishing yin and supporting yang
Licorice, Poria cocos	22	Replenishing qi to invigorate the spleen
Atractylodes, Poria cocos	21	Replenishing qi to invigorate the spleen and swelling
Ginseng, Rehmannia	21	Tonifying kidney-yin, Nourishing kidney-qi
Eucommia, Achyranthes	21	Nourishing kidney-qi, Clearing deficiency fire, Strengthening muscles and bones
Angelica, Rehmannia	21	Tonifying kidney-yin, Nourishing blood and promoting blood circulation
Angelica, Licorice	21	Invigorating spleen to replenish qi, Nourishing blood and promoting blood circulation
Licorice, Rehmannia	21	Tonifying kidney-yin, Nourishing kidney-qi
Psoralen, Eucommia	20	Warming kidney-yang
Astragalus, Ginseng	20	Warming kidney-yang, Nourishing kidney-qi
Atractylodes, Licorice	20	Replenishing qi to invigorate the spleen
Schisandra, Yam	20	Replenishing kidney-essence, Nourishing kidney-qi
Dodder, Poria cocos	20	Warming kidney-yang, Replenishing kidney-essence
Achyranthes, Yam	20	Tonifying kidney-yin, Nourishing kidney-qi

From the perspective of efficacy, the core drug pairs in the prescription library are mainly treated from several aspects such as nourishing kidney yin, warming kidney yang, nourishing kidney qi, nourishing kidney essence, strengthening spleen and kidney, nourishing yin and yang, nourishing qi and activating blood kidney deficiency. The first four effects correspond to the treatment of patients with kidney yin deficiency, kidney yang deficiency, kidney qi deficiency, and kidney essence deficiency, respectively. The “spleen and kidney strengthening” efficacy is based on the “spleen and kidney related theory of TCM”. The spleen and kidney are closely related to each other than other organs. In TCM theory, the kidney is the innate foundation, the main possession of essence. The spleen is the acquired foundation, the source of qi and blood

biochemistry. The predecessors also have "inborn results in growing, and growing supplements inborn" to explain the relationship between the spleen and kidney physiologically and mutual pathologically. Therefore, clinically, the spleen is supplemented first to enhance the effect of nourishing the kidney, or the same treatment of the spleen and kidney can achieve the "equal attention to inborn and growing".

This method of nourishing yin and helping yang is a specific application of the theory of yin and yang. Zhang Jingyue proposed that "yin roots in yang, and yang roots in yin". Yang Qi and Yin essence are one and the same source, and cannot be separated. Yang qi is the driving force of human life activities. The yang qi of human must be based on true yin and blood. For the treatment of kidney deficiency, if only sweet-cold drug is used to nourish the yin and fill the essence, it will often hinder the spleen's movement. This will make it difficult to transform into the essence of the kidney, and the evil of the cold will easily damage the vitality. Similarly, if only warming the kidney yang with warm drugs, it can obtain a temporary effect. However, it will inevitably consume the essence and blood in the long run, make kidney gasification and passive. Therefore, there are multiple prescriptions in the database "reinforcing the yang and not forgetting to nourish the yin". This makes the yang drugs warm and not dry, and the yin drugs nourishing and not tired.

The treatment principle of "invigorating qi and promoting blood circulation" for patients with kidney deficiency also coincides with the "reinforcing kidney and activating blood" proposed by Professor Zhang Daning. Professor Zhang believes that kidney deficiency and blood stasis do not exist in isolation. Kidney deficiency must have blood stasis, and blood stasis increases the kidney deficiency, resulting in a vicious circle. There are multiple core drug pairs in the prescription library that can improve blood circulation and stasis, and achieve the effect of replenishing and relieving the evil.

(b) Analysis of association rules. Using the WD-Get Rules algorithm, we set the support degree ≥0.1, the confidence degree ≥0.5, and analyze the association rules of the prescription database for treating kidney deficiency. The mining results are shown in Table 3-10. The visualize networks are shown in Figure 3-8.

The analysis results show that Poria cocos is in the most central position in the prescription library for treating kidney deficiency. Poria cocos itself does not have the effect of replenishing kidney. However, it is soggy and damp, sweet and invigorating, flat and can be combined with yin and yang. Thus, it can be compatible with many drugs to form a core kidney-drug pair, such as yam, Rehmannia glutinous, dogwood, Schisandra, Achyranthes, ginseng, *etc*. Poria cocos is compatible with yam, which can strengthen the spleen, nourish the stomach and kidney, and is suitable for treating patients with spleen-kidney Qi deficiency. The combination of Poria and Dogwood has the effect of warming the liver and kidney, and putting restraint on sperm and to solidify collaps, and is suitable for treating patients with kidney essence deficiency and kidney yang deficiency. Achyranthes root has the effect of nourishing liver and kidney, strengthening muscle and bone, promoting diuresis and relieving stranguria. It is compatible with Poria cocos for treating soreness and weakness of waist and knees caused by kidney deficiency and deficient fire. Rehmannia glutinosa has the effect of nourishing yin and enriching the blood. This is compatible with Poria cocos, which is used to treat kidney yin deficiency. The combination of Poria cocos and Atractylodes can strengthen the spleen and replenish qi. It is suitable for treating

patients with spleen and kidney qi deficiency. The combination of Poria cocos and Schisandra has the effect of nourishing the kidney, replenishing heart, nourishing qi and vitality. The compatibility of Poria cocos and Ginseng has the effect of nourishing kidney and reinforcing qi, especially suitable for treating patients with kidney qi deficiency. In addition, Eucommia and psoralen alone form a frequent drug pair. According to the records of "Plain Questions·On the Essence of Pulse", kidney essence deficiency results in soreness-tired of waist and knee. Patients with kidney deficiency often show weakness in the waist and knees. The compatibility of Eucommia ulmoides and psoralen can both warm kidney yang and strengthen bones.

Table 3-10. Analysis results of TCM association rules.

Rules	Support	Confidence
Hawthorn => Yam	0.125	0.64865
Hawthorn => Rehmannia	0.11979	0.62162
Yam => Poria cocos	0.17708	0.61818
Schisandra => Rehmannia	0.13021	0.59524
Hawthorn => Poria cocos	0.11458	0.59459
Ginseng => Poria cocos	0.14063	0.5625
Rehmannia glutinosa => Poria cocos	0.16667	0.54237
Ginseng => Licorice	0.13542	0.54167
Psoralen => Eucommia	0.10417	0.54054
Astragalus => Ginseng	0.10417	0.52632
Schisandra => Poria cocos	0.11458	0.52381
Licorice => Ginseng	0.13542	0.52
Atractylodes => Poria cocos	0.10938	0.5122
Fuling =>Yam	0.17708	0.50746
Achyranthes => Poria cocos	0.13021	0.5

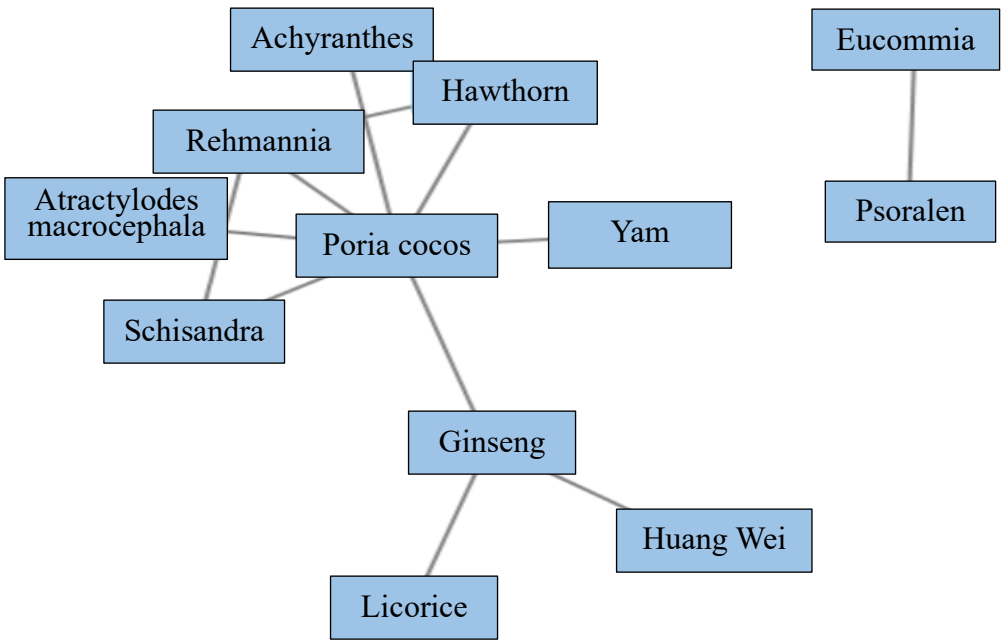


Figure 3-8. Analysis results of TCM association rules.

3.2.3 Discussion of the data mining on the prescription law

In this chapter, we examine the commonly used drugs for clinical treatment of kidney deficiency from the frequency of medication, discuss the compatibility principles of prescriptions from the aspects of drug property, drug flavor and channel tropism, core drug pairs, and degree of association, and analyzed several classical treatment methods of kidney deficiency with mining results. However, the study only focuses on the types of drugs. Because of different historical weights and doses, the dose is not included in the analysis. In further research, the dose will be used as the research object to further dig and interpret the prescription composition rules for treating kidney deficiency.

The WD-Get Rules algorithm used in this research is a new algorithm designed for the prescription database. Compared with the width-first algorithm and the depth-first algorithm, WD-Get Rules is more efficient in mining prescription association rules. The algorithm can also be applied to data mining of other diseases prescriptions. Relevant researchers should focus on the optimization of the algorithm, which will provide more powerful technical support for the interactive fusion of data mining and TCM research.

3.3 Research and examples of k-means clustering algorithm based on initial cluster center optimization

Clustering is an unsupervised learning method that classifies the data based on the different characteristics of the data. Its purpose is to make the density between individuals belonging to the same category as high as possible, and the density between individuals in different categories is as low as possible. Traditional clustering algorithms can be classified as partition-based clustering, density-based clustering, hierarchical-based clustering, grid-based clustering, *etc*. K-means algorithm belongs to one of the partition-based clustering algorithms. This also belongs to a classic distributed clustering algorithm. As an indirect clustering method based on the similarity measure between samples, the K-means algorithm takes the number of cluster classes k as a parameter and divides n data objects into k clusters according to rules. Thus, the similarity within the cluster is high, and the similarity between clusters is low. K-means algorithm has the advantages of simplicity, high efficiency, and easy understanding. However, it still has some shortcomings. It can generally only process spherical or near-spherical data sets, the initial cluster center is unstable, and it is easy to fall into a local optimal solution. Many scholars have made many improvements to overcome these shortcomings. However, the effect on data sets with large differences in density is not good.

For the problem of the sensitivity of the initial clustering center of the K-means algorithm, this chapter proposes an improved initial clustering center selection algorithm. This algorithm introduces the idea of high-density first clustering to improve clustering effect for datasets with large density differences, and enhances the stability of the algorithm. The experiments show that the improved algorithm in this chapter is more stable, and the clustering effect is better. This indicates that the improved algorithm in this chapter is feasible, reasonable and effective.

3.3.1 Basic idea of improved k-means clustering algorithm

To understand the basic idea of the improved algorithm, the basic idea of the K-means algorithm should be understood. First, we randomly select *k* data objects on the data set as the initial clustering center, and then calculate the Euclidean distance between each data object and the center point. The Euclidean distance is divided into the center point with the smallest distance to form *k* cluster classes. The updated cluster center is recalculated. The above steps are repeated until the cluster center no longer changes or the difference between the sum of squared errors in two adjacent clusters less than the threshold.

Suppose the sample data collection $D=\{x_1,x_2,...,x_n\}$, K cluster classes $C=\{C_1,C_2,...C_k\}$, m sets $M=\{M_1,M_2,...M_m\}$

Definition 3-1 Euclidean distance between two data objects is

$$d(x_i,x_j)=\sqrt{\sum_{l=1}^{n}\left(x_{il}-x_{jl}\right)^2} \tag{3-1}$$

where x_i,x_j is the data object, x_{il} is the l th feature attribute of x_i, and x_{jl} is the l th feature attribute of x_j.

Definition 3-2 The center of the cluster $Center_k$ is

$$Center_k=\frac{1}{|C_k|}\sum_{x_i\in C_k}x_i \tag{3-2}$$

where $Center_k$ is the center of the k th cluster class, C_k is the k th cluster class, $x_i\in C_k$ represents a data object belonging to the cluster class C_k.

Definition 3-3 The distance between the point and the sample set is the distance between the point and the mean of the data objects in the set

$$dist=\sum_{x_i\in D'}d\left(x_i,center_{M_m}\right) \tag{3-3}$$

where M represents the set of two points with the shortest distance, D' represents the sample data set after deleting data in the set M, $center_{M_m}$ represents the mean of m th of the set M_m.

Definition 3-4 The objective function, *i.e.*, the sum of squared errors, is

$$E=\sum_{k=1}^{K}\sum_{x_i\in C_k}d(x_i,Center_k) \tag{3-4}$$

Definition 3-5 The sum of the distances between a data object and all data objects in other cluster classes is

$$sum=\sum_{x_j\in C_k}d\left(x_i,x_j\right)\quad\left(x_i\in D\right) \tag{3-5}$$

For the K-means algorithm, the similarity between data objects is calculated using Euclidean distance. The smaller the distance, the higher the similarity. For the denser datasets, the higher the density, the easier it is to get together. If k initial cluster centers can be found, they respectively represent a relatively similar data set. Then, the convergence of the objective

function will be more favorable.

According to the above principle (which can be called the selection principle of initial clustering center), we can find that k points in different densities are used as the initial clustering centers on the spatial distribution of the data. The specific steps are as follows.

(1) Calculate the Euclidean distance $d(x_i, x_j)\ (i, j = 1, 2, ..., n)$ between two data objects according to the Equation (3-1), find the two data objects with the shortest distance to form a sample set $M_m\ (0 \le m \le k)$, and delete them from the total dataset D;

(2) Calculate the mean of all data objects in the sample set M_m;

(3) Calculate the distance between each object in the data set D and the sample set M_m according to the Equation (3-3), find the closest point and add it to the set M_m, and delete it from the data set D;

(4) Calculate the mean of all data objects in the sample set M_m;

(5) Repeat steps 3 and 4 until the data objects in the sample set M_m are greater than or equal to $\alpha \cdot (n/k)$, $(0 < \alpha \le 1)$;

(6) If $m < k$, the number of sets is less than the number of cluster classes, repeat steps 1, 2, 3, 4, 5 until $m \ge k$. When the number of sets equals the number of cluster classes, the process of finding the initial cluster center ends.

For example, there is a two-dimensional data set D, the data size is 14, and its data distribution is shown in Figure 3-9.

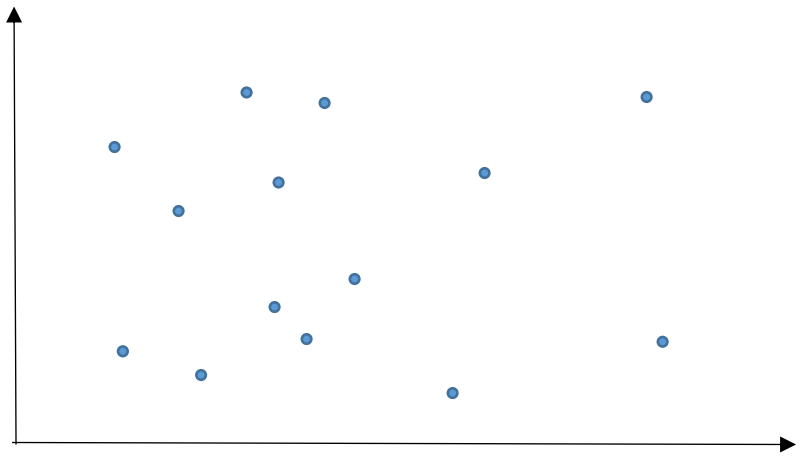


Figure 3-9. Data distribution diagram.

We suppose they need to be classified into two categories. The initial clustering center is found according to the above idea. As shown in the figure, if the distance between a and b is the shortest, a and b will be selected to form a set M_1, and they will be deleted from the data set D. We calculate the distance between the object point in data set D and the set M_1 according to Equation (3-3), find the shortest adjacent point c. Then, we add c to the set M_1 and delete it from D. If the maximum number of data objects in each sample set is 5, we find

d and e in the set M_1 and delete them from D through the thought of the previous step. Then, we find the two closest points i, m in M_2 to form the set D and delete them from D. The closest point in D to M_2 is j. We add j to the set M_2 and delete it from D. In addition, i and f will join the set M_2. We delete them from D. Finally, we calculate the arithmetic mean of the sets M_1 and M_2 as the initial cluster centers of the two cluster classes, respectively.

3.3.2 Specific description of improved k-means clustering algorithm

If the data set $N=\{x_1,x_2,...,x_n\}$ contains n data objects, each data object has s dimensions, the detailed description of the improved algorithm is as follows ($P(0<P\le 1)$ refers to the ratio of the number of data objects to the total data set within the radius R, and the threshold refers to the difference between the sum of the squares of the two adjacent errors):

Input: data set $N=\{x_1,x_2,...,x_n\}$, number of cluster classes K, ratio $P(0<P\le 1)$, threshold β.

Output: clustering results.

(1) Obtain *K* cluster centers as initial centers according to the above-mentioned selection principle of initial cluster center;

(2) Calculate the distance between each data object and the center according to Equation (4-1), and classify it into the nearest cluster center to obtain *K* cluster class;

(3) Recalculate the center of each cluster class according to Equation (4-2);

(4) Re-classify the cluster and update the center;

(5) Until the cluster center no longer changes or the difference between two consecutive values less than the threshold;

(6) The algorithm ends.

3.3.3 Example analysis of improved k-means clustering algorithm

(1) Experimental environment

The processor is Intel (R) Core (TM) i5-3470 CPU @ 3.20GHz with memory of 8.00 GMB, Microsoft Windows10 64-bit operating system, x64 processor. Algorithm writing and compilation are in python3 .5 environment.

(2) Experimental data

To verify the effectiveness of the improved algorithm for clustering in this chapter, the Wine, Hayes-Roth, Iris, Tae, Heart-stalog, Ionosphere, Haberman datasets in the UCI database are selected for experimental analysis. The dimensions of these data sets range from several dimensions to more than ten dimensions. The size of the data set ranges from hundreds to thousands. This reflects the clustering effect of the improved algorithm on data with different dimensions and sizes. This makes the experimental results more persuasive. Details of the data set are shown in Table 3-11.

(3) Experimental results

To verify the effectiveness of the improved algorithm, we compare the traditional K-means algorithm as well as literature algorithm and improved algorithm. In the experiment, we use the

precision, recall, F1 value, and silhouette coefficient (SC) to evaluate the clustering results of the algorithm. Precision refers to the proportion of accurate predictions that are positive for all predictions that are positive. Its value is generally in the [0,1] interval. The greater the value, the more the correctly classified data. The recall is the ratio of accurate predictions that are positive to all predictions that are positive. The ratio is generally in the [0,1] interval. The F1 value is the harmonic mean of precision and recall, and its value is also generally in the [0,1] interval. The silhouette coefficient is an evaluation method for the quality of clustering. The value of silhouette coefficient is in the $[-1,1]$ interval. The greater the value, the closer the clustering result to the real situation. Among them, the silhouette coefficient of the data object can be obtained by the following equation.

$$S(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad (i=1,2,3,...,n)$$

where $a(i)$ represents the average distance between the ith data object and all other points in the cluster to which it belongs. $b(i)$ represents the minimum value of the average distance of ith data object to all points not belonging to its clusters. $S(i)$ represents the silhouette coefficient of any data object.

Table 3-11. Basic information of the dataset.

Datasets	Instance	Dimension	Class
Wine	178	13	3
Hayes-Roth	132	5	3
Iris	150	4	3
Tae	151	5	3
Heart-stalog	270	13	2
Ionosphere	351	33	2
Haberman	306	3	2

Note: Instance represents the dataset size, Dimension represents the data dimension, and Class represents the data category.

In addition, the equation for calculating the precision, recall, and F1 value is

$$P=\frac{Tp}{Tp+Fp};\quad R=\frac{Tp}{Tp+Fn};\quad \mathrm{F1}=\frac{2P\cdot R}{P+R}.$$

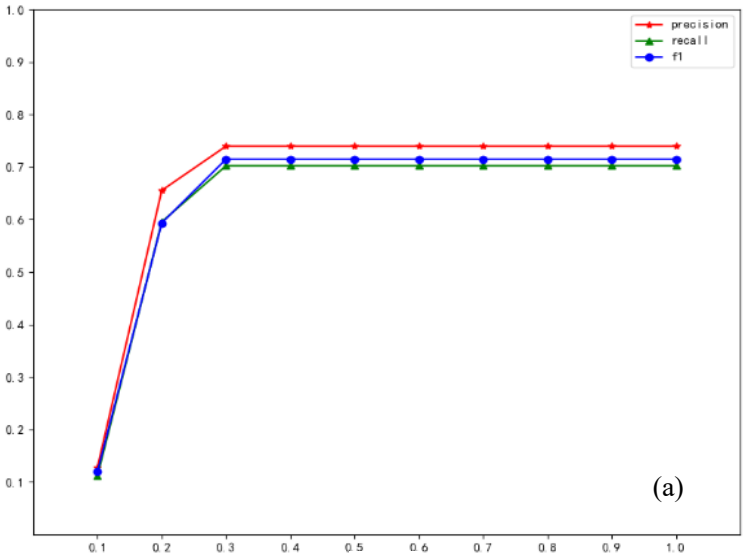
where P is the precision, R is the recall rate, T_P represents that if the sample is positive, the prediction is positive. F_P represents that if the sample is negative, the prediction is negative. T_n represents that if the sample is negative, the prediction is positive. F_n represents that if the sample is positive, the prediction is negative. To better adjust the parameters, the effect of the parameters on the results should be researched and test their performances should be tested. The test results are shown in Figures 3-10(a-g). The red line represents the precision, the green line represents the recall, and the blue line represents the F1.

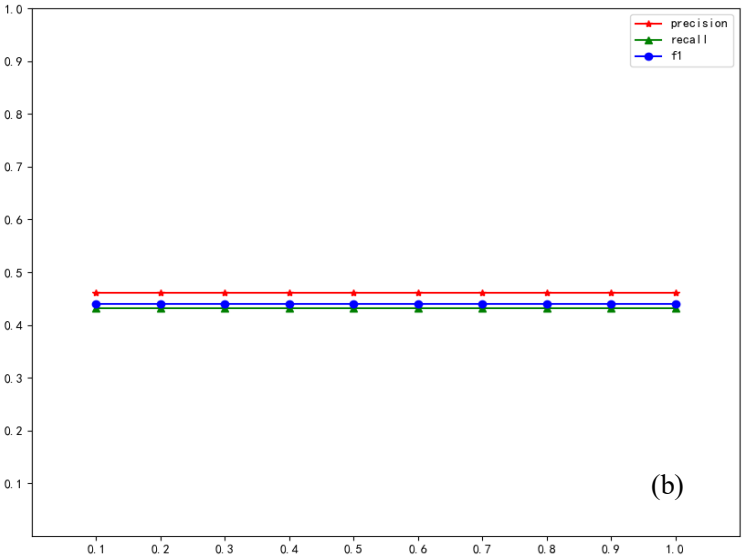
To verify the clustering effect of the improved algorithm, the datasets of Wine, Hayes-Roth, Iris, Tae, Heart-stalog, Ionosphere, Haberman are used as experimental data. Each dataset is tested 5 times. The comparisons among traditional K-means algorithm, K-means ++, literature algorithm, and the improved algorithm in this chapter are shown in Table 3-12, Table 3-13, Table 3-14 and Table 3-15.

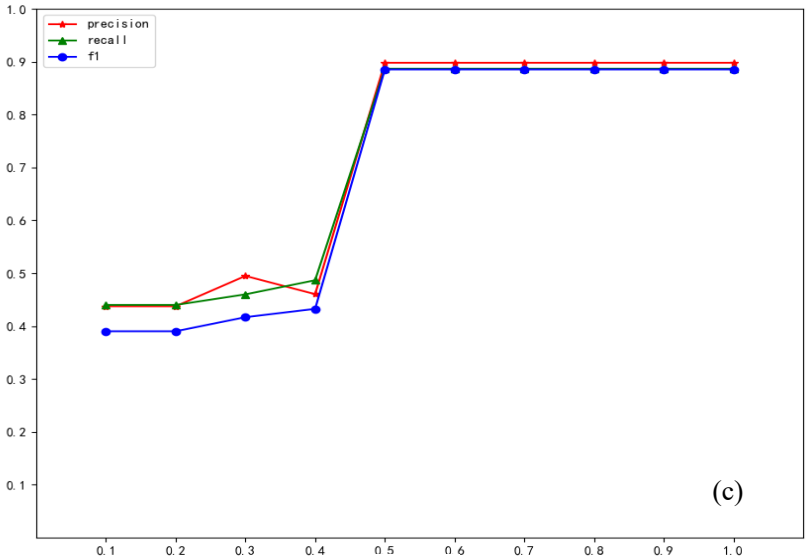
(4) Analysis of research results

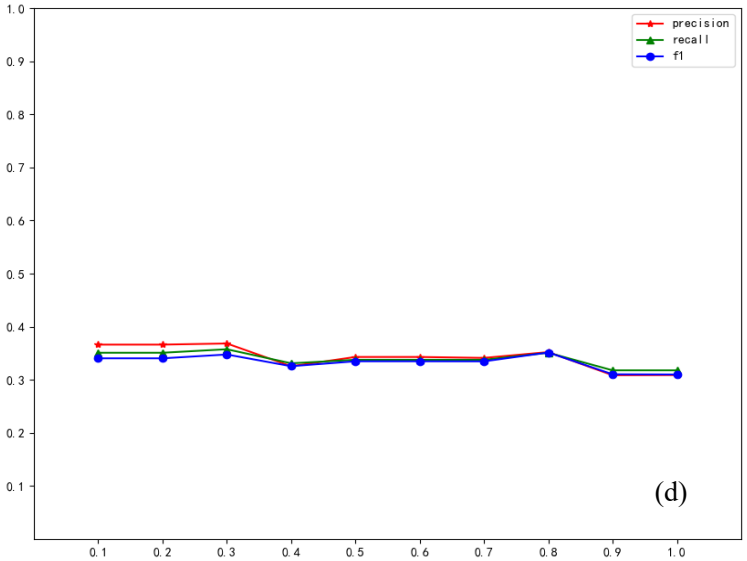
From Tables 3-12 and 3-13, the stability and accuracy of the clustering results of the improved algorithm in the Wine, Hayes-Roth, Iris, Tae datasets are significantly better than those of K-means++, Literature algorithm, traditional K-means algorithm. In addition, in the Heart-stalog, Ionosphere, and Haberman datasets, the stability and accuracy of the clustering results of the improved algorithm in this chapter are better than those of K-means ++ and traditional K-means algorithms, and worse than those of literature algorithms. This is because the algorithm in this chapter performs better when the density of the dataset is high, and the algorithm of the literature performs better when the density of the dataset is low. As shown in Table 3-14, with the same data set, the silhouette coefficient of the clustering results of the improved algorithm in this chapter is greater than or equal to that of the traditional K-means algorithm, K-means ++, and literature algorithm. The F1 of the clustering results of the improved algorithm in this chapter is obviously greater than that of the other three algorithms. Finally, from the data in Table 3-15, the time complexity of the algorithm in this chapter is high. The algorithm runs for a long time, which is not conducive to improving the efficiency of the algorithm. Based on the analysis of Tables 3-12, 3-13, and 3-14, the improved algorithm in this chapter is feasible, reasonable, and effective.

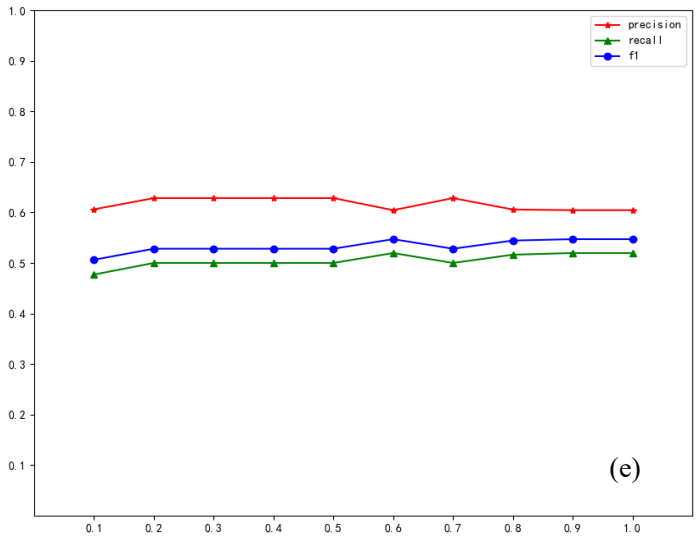
The traditional K-means clustering algorithm is widely used in the field of data mining. With the advent of the era of big data information, this algorithm cannot meet the requirement of data mining. To improve the clustering effect of the algorithm, we propose an improved K - means clustering algorithm in this chapter. The experiment shows that the initial clustering center of the improved algorithm is stable. The sensitivity of the initial clustering center is eliminated. The clustering effect of the improved algorithm is better by comparing the values of the clustering evaluation index. In this chapter, we only focus on the impact of the initial cluster center and the density difference of the sample data. There are still many aspects to be studied. Thus, the next step is to investigate the spatial complexity and time complexity of the improved algorithm. Then, we will explore a better way to improve the efficiency of the algorithm.

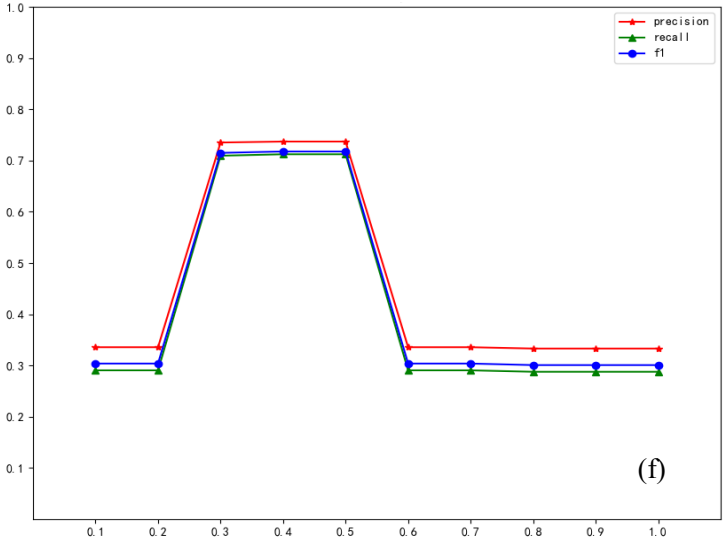

precision
recall
f1
(a)


precision
recall
f1
(b)


precision
recall
f1
(c)


precision
recall
f1
(d)





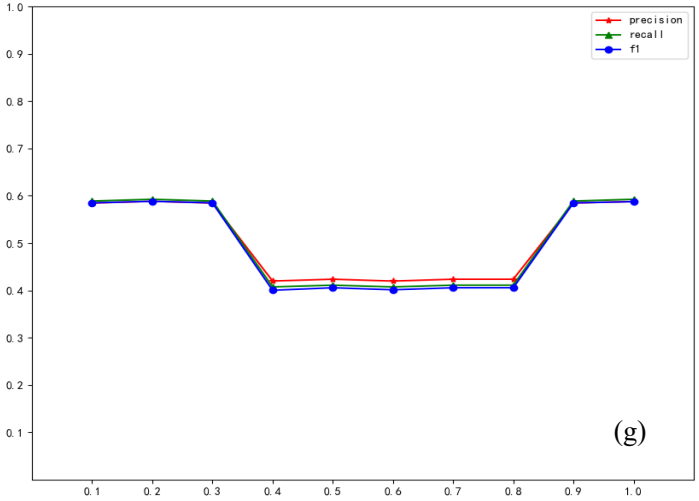


Figure 3-10. Performance test results for (a) The Wine dataset; (b) the Hayes-Roth dataset; (c) the Iris dataset; (d) the Tae data set; (e) the Heart-stalog data set; (f) the Ionosphere data set; (g) the Haberman data set. The horizontal axis represents the parameter q, and the vertical axis represents the parameter performance.

Table 3-12. The comparison of stability and accuracy of clustering results among different algorithms (1).

Algorithm	Experiment	Wine		Hayes-Roth		Iris		Tae	
		Original center	**Accuracy**	**Original center**	**Accuracy**	**Original center**	**Accuracy**	**Original center**	**Accuracy**
Traditional K-means algorithm	1	(74, 86, 80)	0.478	(122, 69, 117)	0.273	(69, 44, 78)	0.240	(102, 44, 125)	0.318
	2	(118, 17, 103)	0.169	(6, 72, 95)	0.386	(18, 146, 5)	0.473	(94, 91, 102)	0.371
	3	(42, 80, 8)	0.702	(29, 51, 105)	0.280	(107, 120, 83)	0.320	(116, 29, 2)	0.298
	4	(139, 141, 9)	0.112	(115, 77, 80)	0.424	(26, 146, 22)	0.500	(48, 69, 58)	0.358
	5	(175, 97, 47)	0.354	(130, 115, 94)	0.295	(30, 100, 145)	0.440	(49, 89, 116)	0.338
	Average	-	0.363	-	0.332	-	0.395	-	0.337
K-means++ algorithm	1	(127, 177, 89)	0.112	(77, 131, 66)	0.280	(144, 149, 16)	0.320	(3, 150, 92)	0.351
	2	(176, 177, 16)	0.500	(96, 131, 21)	0.333	(39, 149, 60)	0.447	(61, 150, 141)	0.298
	3	(27, 177, 103)	0.478	(122, 131, 48)	0.394	(92, 149, 68)	0.440	(47, 150, 148)	0.298
	4	(98, 177, 16)	0.354	(88, 131, 69)	0.379	(106, 149, 11)	0.02	(48, 150, 41)	0.338
	5	(3, 177, 82)	0.702	(120, 131, 7)	0.280	(4, 149, 133)	0.887	(38, 150, 124)	0.305
	Average	-	0.429	-	0.333	-	0.423	-	0.318
Literature algorithm	Five times	(24, 26, 92)	0.478	(33, 34, 98)	0.386	(79 15 136)	0.887	(93, 146, 127)	0.245
Improve algorithm	Five times	-	0.702	-	0.4318	-	0.887	-	0.358

Note: The values of the K-means and K-means++ algorithms are taken from frequent items. The test data removes the non-numeric attributes and tag data attributes in the dataset. "-" indicates that the matching object cannot be found in the dataset.

Table 3-13. The Comparison of stability and accuracy of clustering results among different algorithms (2).

Algorithm	Experiment	Heart-stalog		Ionosphere		Haberman	
		Initial center	Accuracy	Initial center	Accuracy	Initial center	Accuracy
Traditional K-means algorithm	1	(194, 77)	0.589	(269, 184)	0.644	(39, 120)	0.500
	2	(174, 2)	0.411	(289, 216)	0.291	(18, 296)	0.516
	3	(15, 266)	0.411	(10, 239)	0.709	(177, 113)	0.242
	4	(267, 34)	0.589	(44, 46)	0.288	(127, 43)	0.480
	5	(18, 111)	0.411	(250, 264)	0.712	(10, 100)	0.500
	Average	-	0.500	-	0.529	-	0.448
K-means++ algorithm	1	(73, 269)	0.411	(69, 350)	0.709	(39, 305)	0.520
	2	(221, 269)	0.407	(47, 350)	0.288	(136, 305)	0.520
	3	(4, 269)	0.407	(152, 350)	0.291	(72, 305)	0.520
	4	(108, 269)	0.411	(196, 350)	0.288	(192, 305)	0.520
	5	(113, 269)	0.589	(221, 350)	0.288	(268, 305)	0.520
	average	-	0.469	-	0.429	-	0.520
Literature algorithm	Five times	(183, 58)	0.589	(51, 23)	0.712	(181, 215)	0.758
Improve algorithm	Five times	-	0.589	-	0.709	-	0.500

Note: The values of the K-means and K-means++ algorithms are taken from frequent items. The test data removes the non-numeric attributes and tag data attributes in the dataset. "-" indicates that the matching object cannot be found in the dataset.

Table 3-14. The comparison of SC and F1 among different algorithms in the UCI dataset.

UCI data set	Traditional k-means algorithm		k-means++		Literature algorithm		Improve algorithm	
	SC	F1	SC	F1	SC	F1	SC	F1
Iris	0.524	0.357	0.551	0.295	0.541	0.885	0.549	0.885
Wine	0.571	0.370	0.563	0.357	0.571	0.495	0.571	0.715
Hayes-Roth	0.571	0.346	0.571	0.354	0.571	0.395	0.571	0.441
Heart-stalog	0.377	0.495	0.379	0.403	0.377	0.585	0.377	0.585
Ionosphere	0.286	0.457	0.294	0.405	0.295	0.718	0.293	0.715
Tae	0.316	0.330	0.325	0.312	0.335	0.226	0.328	0.348
Haberman	0.387	0.443	0.399	0.547	0.356	0.733	0.393	0.528

Note: The data values in the traditional k-means algorithm and k-means++ are obtained under the condition of averaging. The test data removes the non-numeric attributes and tag data attributes in the dataset.

Table 3-15. The comparison of clustering time among different algorithms in UCI dataset.

UCI dataset	Traditional k-means algorithm	k-means++	Literature algorithm	Improve algorithm
Iris	0.220	0.245	3.450	5.539
Wine	0.312	0.207	4.713	12.270
Hayes-Roth	0.193	0.360	2.470	4.390
Heart-stalog	0.419	0.398	10.821	26.284
Ionosphere	0.333	0.460	18.058	86.439
Tae	0.291	0.358	3.510	6.438
Haberman	0.418	0.527	13.579	25.314

Note: The data values in the traditional k-means algorithm and k-means++ are obtained under the condition of averaging. The test data removes the non-numeric attributes and tag data attributes in the dataset.

References

1. I.H Witten. Data mining. China Machine Press: Beijing, China, 2012.
2. H. Toivonen. Apriori algorithm. *Encyclopedia of Machine Learning* 2011, 39-40.
3. X.Z. Niu, K. She. Mining maximal frequent item sets with improved algorithm of FPMAX. *Computer Science* **2013**, *40*, 223-228.
4. X. Ye, F. Wei, F. Jiang, *et al*. An optimization to CHARM algorithm for mining frequent closed itemsets. *IEEE International Conference on Computer and Information Technology* **2015**, 226-235.
5. R.C. Taylor. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. *BMC Bioinformatics* **2010**, *11*, 1.
6. C. Meng, Z.Y. Cao, L. Wang, *et al*. Charm ++ RTS based fault tolerance mechanism of heterogeneous computing. *Computer Engineering and Applications* **2016**, *52*, 1-7.
7. X.D. Wu, S.Z. Ruan. Comparative study on MapReduce and spark for big data analytics. *Journal of Software* **2018**, *29*, 1770-1791.
8. R. Joy, K.K. Sherly. Parallel frequent itemset mining with spark RDD framework for disease prediction. International Conference on Circuit, 2016.
9. H. Qiu, R. Gu, C. Yuan, *et al*. YAFIM: a parallel frequent itemset mining algorithm with spark. Parallel & Distributed Processing Symposium Workshops, 2014.
10. S. Yang, G. Xu, Z. Wang, *et al*. The parallel improved Apriori algorithm research based on spark. Ninth International Conference on Frontier of Computer Science & Technology, 2015.
11. S. Rathee, M. Kaul, A.R. Kashyap. Apriori: an efficient Apriori based algorithm on spark. Workshop on PHD Workshop in Information & Knowledge Management, 2015.
12. Q.P. Li, L.G. Zhang, X.Y. Geng. Apriori: an improved Apriori algorithm based on spark platform. *Science Technology & Engineering* **2017**, *27*.
13. Y. Luo, Z. Yang, H. Shi, *et al*. A distributed frequent itemsets mining algorithm using sparse Boolean matrix on spark. Asia-Pacific Web Conference, 2016.
14. H. Niu, Lu H, Liu Z. The improvement and research of Apriori algorithm based on spark. *Journal of Northeast Normal University* 2016, *1*.
15. Y. Lu, Z.S. Dong. An improved algorithm of association rules based on the spark. *Application of Electronic Technique* **2017**, *6*.

16. F. Feng, J. Cho, W. Pedrycz, *et al*. Soft set-based association rule mining. *Knowledge-based Systems* **2016**, *111*, 268-282.
17. W. Wang, W. Zhao, C.P. Li, *et al*. Feature extension and category research for short text based on spark platform. *Journal of Frontiers of Computer Science and Technology* **2017**, 11, 732-741.
18. R. Dharavath, S. Raj. Quantitative analysis of frequent itemsets using Apriori algorithm on apache spark framework. *Proceedings of the International Conference on Computational Intelligence in Data Mining* **2017**, 261-272.
19. B. Cao, J.C. Ni, L.L. Li, *et al*. Parallel frequent pattern mining algorithm based on Spark. *Computer Engineering and Applications* **2016**, *52*, 86-91.
20. Rathee S, Kashyap A. Exploiting apache Flink's iteration capabilities for distributed Apriori: community detection problem as an example. International Conference on Advances in Computing, 2016.
21. K.S. Krishan, R. Dharavath. HFIM: A Spark-based hybrid frequent itemset mining algorithm for big data processing. *Journal of Supercomputing* **2017**, *73*, 1-17.
22. Q. Wang, L. Tan, X.H. Yang. Optimization of Apriori parallel algorithm based on spark. *Journal of Zhengzhou University (Natural Science Edition)* **2016**, *48*, 60-64.
23. X. Fang, G.X. Zhang. Optimization of parallel FP-Growth algorithm based on spark. *Modern Electronics Technique* **2016**, *8*.
24. W. Zhang, K. Luo. A parallel FP-Growth mining algorithm based on Spark framework. *Computer Engineering & Science* **2017**, *8*.
25. R.H. Jiao, Q. Zhang, C. Chen. Improved algorithm for mining maximum frequent itemsets based on Spark. *Computer Engineering and Design* **2017**, *38*, 1839-1843.
26. L. Shao, X.Z. He, J.N. Shang. Frequent item sets mining algorithm for big data based on FP-Growth and Spark framework. *Application Research of Computers* **2018**, *35*, 2932-2935.
27. J.H. Gu, J.Y. Wu, X.Y. Xu, *et al*. Optimization and implementation of parallel FP-Growth algorithm based on Spark. *Journal of Computer Applications* **2018**, *38*, 3069-3074.
28. L.K. Shi, X. Zhang, S.L. Shi. Parallelization and optimization of FP_Growth algorithm based on Spark. *Computer Engineering and Applications* **2018**, *54*, 52-58.
29. A.D.D. Gassama, F. Camara, S. Ndiaye. S-FPG: a parallel version of FP-Growth algorithm under Apache spark. IEEE International Conference on Cloud Computing & Big Data Analysis, 2017.
30. F. Zhang, M. Liu, F. Gui, *et al*. A distributed frequent itemset mining algorithm using spark for big data analytics. *Cluster Computing* **2015**, *18*, 1493-1501.
31. K. Lu, W. Gui, Y.Y. Jiang, *et al*. Optimization and implementation of parallel FP-growth algorithm based on spark. *Computer Applications and Software* **2017**, *34*, 273-278.
32. X.J. Feng, X. Pan. Eclat algorithm based on Spark. *Application Research of Computers* **2019**, *1*.
33. X. Yu, Q.Y. Xie, Q.G. Meng. Scientific value of TCM integrative data analysis in big data era. *Chinese Journal of Information on Traditional Chinese Medicine* **2015**, *22*, 1-3.

34. B. Zhang. Research on data-mining technology applied traditional Chinese prescription compatibility based on association rules. *Journal of Gansu Lianhe University (Natural Science Edition)* **2011**, *25*, 82-86.
35. F.S. Chen. Accurate understanding of kidney deficiency science to protect the kidneys. *China Consumer News* **2010**, 3-8.
36. A.F. Zhou. Theory of state of viscera in kidney and its clinical application. *Journal of Tianjin University of Traditional Chinese Medicine* **2014**, *33*, 1-5.
37. W.H. Jia, K. Micheline, P. Jian. Data Mining: Concepts and Techniques (Third edition). Mechanical Industry Press: Beijing, China, 2017.
38. J. Yin, L. Nian, J.Y. Zhang. summary of "Sheng Ji Zong Lu". *Liaoning Journal of Traditional Chinese Medicine* **2015**, *10*, 2024-2026.
39. Y. Zeng, J. Zhang. Application of data mining technology in traditional Chinese medicine. *Chinese Journal of Information on Traditional Chinese Medicine* **2012**, 19, 99-100.
40. B.Z. Yan. A Summary of research on prescription editor's masterpiece "Puji Formula" Since 1980s. *Journal of Pingdingshan University* **2012**, *27*, 48-55.
41. W.M. Li, W.H. Li. Study on compatible law of Shudihuang (Rhizoma Rehmannia Praeparatae) in formula. *China Journal of Traditional Chinese Medicine and Pharmacy* **2011**, *12*, 2810-2812.
42. Y. Lin, Y.C. Wang. Is it feasible to supplement with shape? *TCM Healthy Life-Nurturing* **2017**, *12*, 31-34.
43. Y.M. Li. Is supplement with shape equal to eating to make up? *China Drug Store* **2014**, *12*, 80-81.
44. B. Xiao, W. Wang, W.J. Guo, *et al*. Relationship between Chinese herbal nature combination and functions. *World Science and Technology (Modernization of Traditional Chinese Medicine and Materia Medica)* **2010**, *12*, 902-908.
45. Z. Li. Poria cocos, a classic traditional Chinese Medicine. *Cancer Frontier* **2011**, *3*, 52-53.
46. W. Ma. Overview and application of clustering algorithm. *China New Telecommunications* **2018**, **20**, 225-226.
47. C. Zhao. Research on Clustering Algorithm for Mixed Attributes and Application. Ph.D. Thesis, Yanshan University, Hebei, China, 2017.
48. G. Yuan, P.H. Sun, J. Zhao, D.X. Li, C.W. Wang. A review of moving object trajectory clustering algorithms. *Artificial Intelligence Review* **2017**, *47*.
49. W. Zhao. Survey of hierarchical clustering community. *Wireless Internet Technology* **2015**, *19*, 115-116.
50. K. Song. Review of clustering algorithms. *Journal of Henan Science and Technology* 2015, *22*, 254.
51. M. Hai. Survey of clustering algorithms for big data. *Computer Science* **2016**, *43*, 380-383.
52. W. Jia. Survey on partitional clustering algorithms. *Electronic Design Engineering* **2014**, *22*, 38-41.
53. W. Liu. Survey of clustering algorithms in data mining. *Jiangsu Commercial Forum* **2018**, *7*, 120-125.
54. X.Q. Chen, L.J. Zhou, Y.Z. Liu. Review on clustering algorithms. *Journal of Integration Technology* **2017**, *6*, 41-49.

55. T. Zhou, H.L. Lu. Clustering algorithm research advances on data mining. *Computer Engineering and Applications* **2012**, *48*.
56. Q. Wang, C. Wang, Z.Y. Feng, J.F. Ye. Review of K-means clustering algorithm. *Electronic Design Engineering* **2012**, *20*, 21-24.
57. J.Y. Xie, Y.Z. Wang. K-means algorithm based on minimum deviation initialized clustering centers. *Computer Engineering* **2014**, *40*, 205-211.
58. C.Z. Xing, H. Gu. K-means algorithm based on average density optimizing initial cluster centre. *Computer Engineering and Applications* **2014**, *50*, 135-138.
59. C.L. Wang, J.X. Zhang. Improved K-means algorithm based on latent Dirichlet allocation for text clustering. *Journal of Computer Applications* **2014**, *34*, 249-254.
60. R.W. Zhou, Z.Y. Li, S.Q. Chen, *et al*. Parallel optimization sampling clustering K-means algorithm for big data processing. *Journal of Computer Applications* **2016**, *36*, 311-315.
61. C.X. Yin, H.J. Zhang, R. Zhang, *et al*. An improved K-Means clustering algorithm. *Computer Technology and Development* **2014**, *24*, 30-33.
62. X.X. Chen, Y.Q. Zhai, M. Ren, *et al*. Weighted K-means clustering algorithm based on firefly algorithm. *Application Research of Computers* **2018**, *35*, 466-470.
63. M.L. Liu, M.X. Huang, W.D. Tang. A k-means algorithm for optimized initial clustering center based on discrete quantity. *Computer Engineering & Science* **2017**, *39*, 1164-1170.
64. S.Y. Qian, H.H. Liu, D.Y. Li. Research and application of improved K-means algorithm in text clustering. *DEStech Transactions on Computer Science and Engineering*, 2018.
65. S.K. Majhi, S. Biswal. Optimal cluster analysis using hybrid K-Means and Ant Lion Optimizer. *Karbala International Journal of Modern Science* **2018**.
66. J.E.Z Gbadoubissa, A.A. Ari, A.M. Gueroui. Efficient K-means based clustering scheme for mobile networks cell sites management. *Journal of King Saud University Computer and Information Sciences* **2018**.

Chapter 4 Application and Examples of Neural Network Algorithm in The Field of Traditional Chinese Medicine

4.1 Research on the prediction system of efficacy of TCM prescription based on neural network

TCM prescriptions are usually composed of four parts, such as Jun, Chen, Zuo, and Shi. The relationship between property of TCM and the TCM prescription efficacy is not linear. Thus, the prediction of the TCM prescription efficacy has always been a hot topic and research difficulty in the field of TCM. In this chapter, through the "quantization" treatment of TCM property and TCM prescription information, a "quantum" database of standardized TCM prescriptions is established. The mathematical model in the efficacy prediction of TCM prescription is established by using the principle of neural network. Using MATLAB software, a system for predicting the efficacy of TCM prescription is developed. The satisfactory experimental results are obtained through the simulation operation. The development of this system cannot only predict and analyze the efficacy of TCM prescription, but also provide a reference for the research of TCM prescription compatibility.

4.1.1 The TCM prescription efficacy and neural network

(1) Prescription efficacy

The prescription efficacy is derived from the drug efficacy. The meaning of the term used is similar to that of the TCM, and it is different from the efficacy of the single drug. The four properties, five flavors, and channel tropism as the main content of TCM can explain the efficacy of TCM from different aspects. Therefore, all aspects must be considered to comprehensively study the prescription efficacy. The prescription efficacy is not a simple sum of the efficacies of TCMs in the prescription, and is related to the dose, compatibility, dose form and decoction methods of the various medicines. For example, the cassia twig is the dried sprigs of plant cinnamon, which can reinforce yang and relievie exterior syndrome, warm and dredge collaterals, activate yang to promote diuresis, warm the chest yang, warm middle-jiao to dispel cold. Cinnamon notopterygium soup is made up of Cinnamon soup without peony, ginger, jujube, and adding notopterygium, radix saposhnikoviae. The notopterygium, radix saposhnikoviae are used to expel wind and clear away cold. This has the function of dispelling pathogenic wind from muscles. Ramuli cinnamomi and ginseng decoction is made up of Cinnamon soup without peony, dried ginger, jujube, adding ginseng and atractylodes. Ginseng and Atractylodes are used for replenishing qi to invigorate the spleen. They have the function of warming and nourishing, simultaneous treatment of the interior and exterior.

(2) Neural network

A neural network is a common computing model consisting of a large number of nodes (or

neurons) connected to each other. Each node represents a specific activation function. There is a weighting value through the connection between two nodes, which is simply called weight. The neural network simulates human memory. We apply it to various fields such as pattern recognition, image processing, text classification, and disease diagnosis. The neural network is mainly composed of three parts of the input layer, the hidden layer and the output layer. The three-layer topology is shown in Figure 4-1.

(3) Relationship between TCM prescription and neural network

Through the "quantization" processing of the TCM prescription data, the basic characteristic attributes and effects of the prescription are taken as the input and output of the neural network, respectively. The neural network is used to establish a mathematical model. The system is developed to predict the efficacy of the TCM prescription.

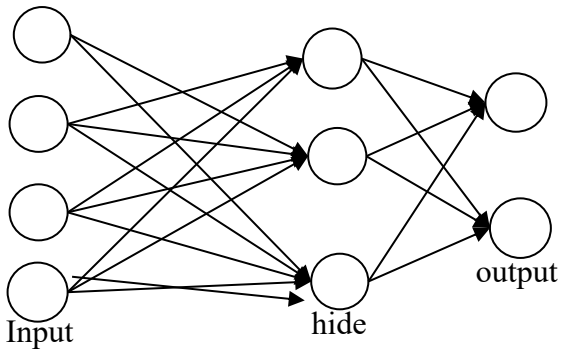


Figure 4-1. Three-layer topology of the neural network.

4.1.2 Establishment of the mathematical model for prediction of TCM efficacy

(1) Quantization of information of TCM prescription

The textual expression of prescription information of TCM is complicated. There are the phenomena of "polysemy" and "homonym words". Thus, to make the computer better handle the information of TCM prescription, we introduce the concept of "quantization" taking prescriptions of heat-clearing TCMs as samples. For example, "yes" and "no" of cold, hot, warm, cool and flat of drug properties of TCM are expressed as "1" and "0", respectively. Based on the binary quantification of single drug, the numerical quantification is adopted according to the compatibility of TCM prescription. The weights adopted by "Jun", "Chen" and "zuoshi" are 0.5, 0.3, and 0.2 respectively. According to the formula $\frac{WI}{Q}$ (where W is the weight, I is the initial value of the flavor effect attribute, Q is the sum of all the WI), we calculate the attribute values of different TCM efficacy and prescription efficacy. 34 kinds of heat-clearing TCM prescriptions are quantified by the method of "quantization" numerical processing. Some results are shown in Table 4-1.

Table 4-1. Partial results of quantification of TCM prescription information.

Prescription name	Cold	Hot	Warm	Freeze	Flat	Acid	Bitter	Sweet	Spicy	Salty	Purging fire	Dampness	Detoxification	Cooling blood
Mahuang decoction	0.8	0	0.04	0.04	0.12	0	0.33	0.3	0.33	0.04	1	0	1	0
Guizhi decoction	0.5	0	0.42	0	0.08	0.16	0.44	0.2	0.2	0	0	0	0	1
Notoptreygium decoction of nine ingredients	0.33	0	0.37	0.3	0	0	0.42	0.38	0.2	0	0	0	0	1
Gegen decoction	0.33	0	0.2	0.19	0.28	0	0.07	0.5	0.43	0	0	0	0	1
Mahuang Fuzi Gancao decoction	1	0	0	0	0	0	0	0.5	0.5	0	1	0	0	0
Huoxiang Zhengqi powder	0.37	0	0.33	0	0.3	0	0.14	0.08	0.78	0	0	0	0	1
Fragrant	0.6	0	0.4	0	0	0	0.1	0.2	0.7	0	0	1	1	0
Pueraria and Pinellia deoction	0.88	0	0.04	0.04	0.04	0.08	0.34	0.08	0.5	0	1	0	1	0
Yinqiao powder	0.1	0	0.5	0.4	0	0.04	0.8	0.08	0.04	0.04	1	0	0	0
.....														
Potent purgative decoction	0.5	0	0.1	0.4	0	0	0.5	0.5	0	0	0	0	0	1
Mill purgative decoction	0.5	0	0.2	0.3	0	0	0.42	0.34	0.25	0	1	1	0	0
Bone clearing powder	0.88	0	0.04	0.04	0.04	0	0.4	0.4	0.2	0	1	0	0	0
Qingying decoction	0.85	0	0.05	0.1	0	0	0.37	0.3	0.33	0	1	0	0	0
Lianmei decoction	0.47	0	0	0.43	0.1	0	0.8	0.1	0.1	01	1	0	0	0

(2) Establishment of the prediction model for efficacy of TCM prescription

(a) Determination of the number of neurons of input, output and hidden layers of the model

The input layer of this model is the basic attribute characteristic value of TCM, including 10 parameters of the four properties and five flavors of TCM, which is 10 neurons with the input range of 0-1. The output layer is an attribute value indicating the efficacy of the heat-clearing prescription, such as clearing heat and purging fire, clearing away heat and dampness, clearing away heat and detoxifying, and clearing heat and cooling blood, which is 4 neurons with the output range of 0-1. The number of hidden layers is one. A large number of practices have shown that a neural network with one hidden layer can display any continuous function with arbitrary precision. The number of neurons in the hidden layer is obtained according to the empirical formula Kolmogorov's law. After continuous testing, the number of neurons in the hidden layer of this model is finally determined to be 25.

(b) model establishment, testing and training

The established BP neural network model is a three-layer topology with 10 input neurons, 25 hidden neurons and 4 output neuron nodes. The transfer function between the layers is a sigmoid function with non-linear features and differentiable characteristics, which can better reflect the input and output characteristics of artificial neurons. In this chapter, 30 of the 34 prescription data samples are included as training samples, and 4 sets of data are used as test samples (Table 4-2). Then, the established model is performed network training.

(c) Prediction results of the model

In this chapter, 34 samples of heat-clearing prescriptions are used as samples, of which 1-30 are training samples, and 31-34 are test samples. The characteristic properties of prescription drugs—four qi and five flavors are the input layer data. The four efficiencies of the prescription, such as heat-clearing and purging fire, clearing away heat and dampness, clearing away heat and detoxifying, and clearing heat and cooling blood, are taken as the output layer data. The established prediction model of prescription efficacy is performed network training. After 200 iterations, the expected error value achieves 0.0002. The network training is completed. The training error curve obtained by MATLAB is shown in Figure 4-2. According to the actual output of the last four sets of prediction data, the prediction effect of the model is good, and the accuracy rate is as high as 87.5%. The comparison between the expected output and the actual output of the predicted sample of the heat-clearing prescriptions is shown in Table 4-3 and Table 4-4.

Table 4-2. Test data.

Prescription	Cold	Hot	Warm	Freeze	Flat	Acid	Bitter	Sweet	Spicy	Salty
Mill purgative decoction	0.5	0	0.2	0.3	0	0	0.42	0.34	0.25	0
Bone clearing powder	0.88	0	0.04	0.04	0.04	0	0.4	0.4	0.2	0
Qingying decoction	0.85	0	0.05	0	0	0	0.37	0.3	0.33	0
Lianmei decoction	0.47	0	0	0.43	0.1	0	0.8	0.1	0.1	0

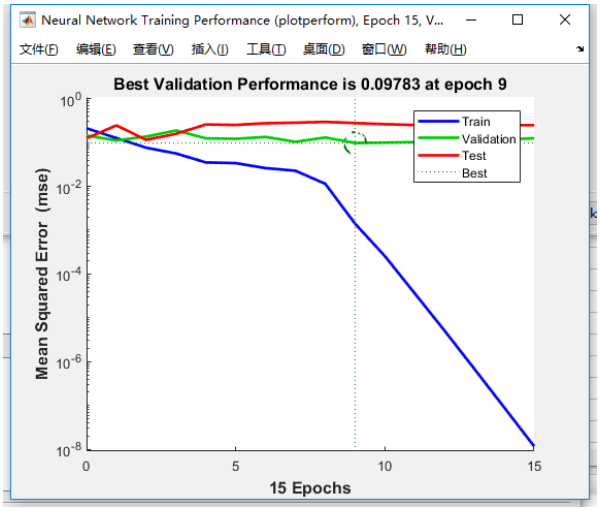


Figure 4-2. Training error curve.

Table 4-3. Expected and actual output of the predicted sample of the heat-clearing prescription.

Prescription name	Actual output				Expected output			
	Purging fire	Dampness	Detoxification	Cooling blood	Purging fire	Dampness	Detoxification	Cooling blood
Mill purgative decoction	0.9972	0.1798	0.2064	0.254	1	1	0	0
Bone clearing powder	0.9973	0.1936	0.3318	0.2264	1	0	0	0
Qingying decoction	0.9972	0.2724	0.4117	0.5048	1	0	0	0
Lianmei decoction	0.9972	0.2552	0.0688	0.0701	1	0	0	0

Table 4-4. The comparison of expected output and actual output of the predicted sample of heat-clearing prescription.

Prescription name	Purging fire	Dampness	Detoxification	Cooling blood
Mill purgative decoction	correct	error	correct	correct
Bone clearing powder	correct	correct	correct	correct
Qingying decoction	correct	correct	correct	error
Lianmei decoction	correct	correct	correct	correct

4.1.3 Design and implementation of the efficacy prediction system for TCM prescription

(1) The requirement analysis of system

In the field of TCM prescription, most researches focus on the dose-efficacy relationship of TCM prescriptions. However, there are few researches about the efficacy prediction. The development of efficacy prediction system of TCM prescription provides a scientific and effective method for predicting the efficacy of TCM prescriptions. The main tasks to be completed by the prediction system include inputting data and setting parameters through the

system, importing the datasheet of TCM prescription, predicting the efficacy of the TCM prescription, and obtaining the accuracy rate of the prediction. In view of the advantages of neural networks, we use neural network models to analyze the property and flavor data of heat-clearing prescriptions and predict the efficacy of prescriptions.

(2) System function design

The main functions of the prediction system of TCM prescription efficacy are three modules of data input, parameter setting and prediction result. The functional module design of the system is shown in Figure 4-3.

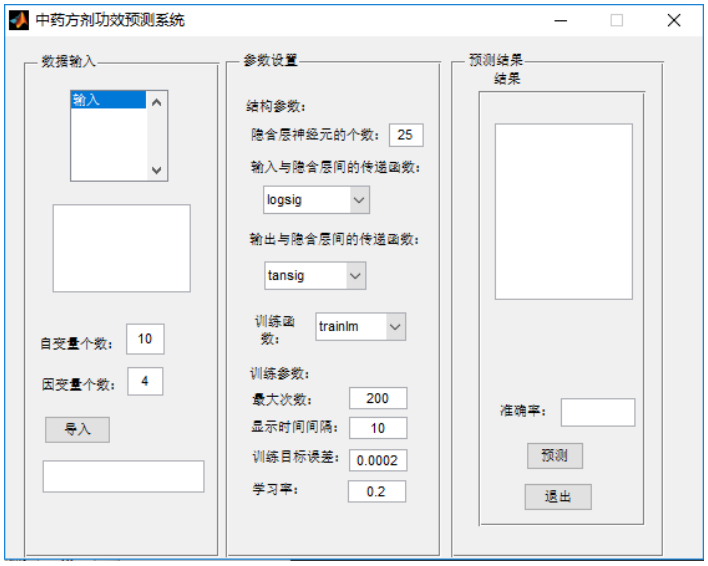


Figure 4-3. System function design.

(a) Data input module. For this module, we click the import button to import the required prescription quantization database, data format requirements.xlsx. The data will automatically display the data and its path after being imported.

(b) Parameter setting module. It includes two parts of network structure parameters and training parameters. The structure parameters include the number of hidden layer neurons, the transfer function between the input layer and the hidden layer, and the transfer function and training function between the output layer and the hidden layer. The training parameters include the maximum number of training, display time interval, training target error and learning rate.

(c) Prediction result module. This module includes training, testing and prediction functions. After importing the data, we click the prediction button, the system will automatically display the prediction result, and output the prediction accuracy rate. Then, we click the exit button to shut down the system.

(3) System development environment

Based on MATLAB's powerful function and the simplicity of language, the system development environment is Matlab2014a, and the operating system is Windows10. For good interactive design, the system integrates MATLAB object-oriented programming method and GUI design.

4.1.4 Results and analysis of the efficacy prediction system of TCM prescriptions

After the parameters are set, we click the "Import button" to select the training samples of the model. The system automatically trains the samples. The results are shown in Figure 4-4. In the prediction result module, we click the prediction button, select the sample to predict the efficacy of the prescription. The system automatically generates the prediction result, and outputs the corresponding accuracy rate. The system prediction interface process is shown in Figure 4-5.

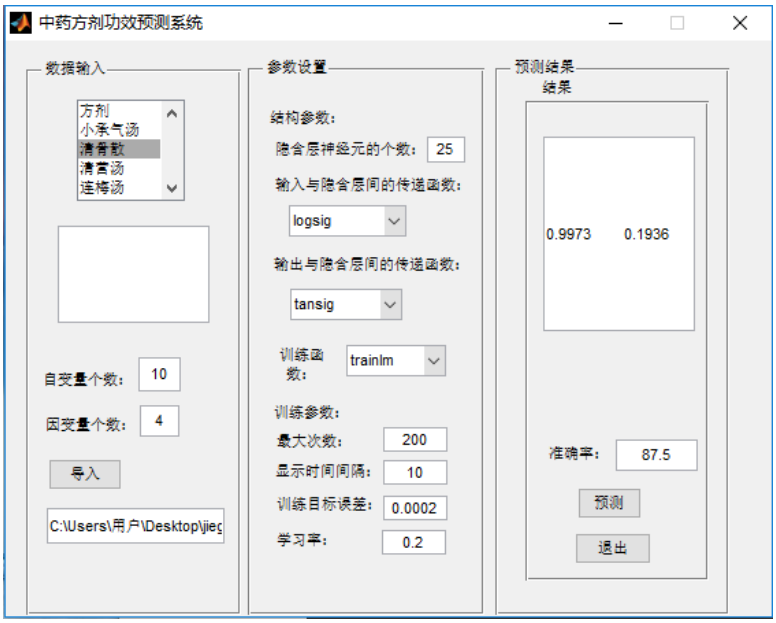


Figure 4-4. System data import interface.

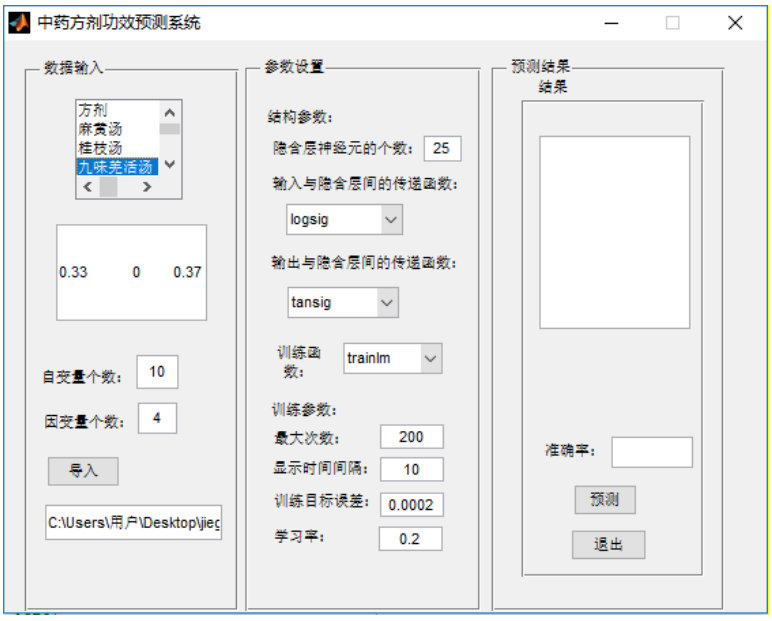


Figure 4-5. System prediction result interface.

Studies have shown that the of prediction system of TCM prescription efficacy based on the neural network is effective. The system utilizes the ambiguity and fitting of the neural

network, mines the relationship between the drug properties and efficacy of TCM prescriptions. This deals with the nonlinear problems between the basic attributes of TCMs and the efficacy of TCM prescriptions, and realizes the rapid prediction of the efficacy of TCM prescriptions. With the development of science and technology, the neural network method is increasingly used in science. It is still in the development stage, and there are few studies in the research of TCM. I believe that in the near future, the methods, parameters, training samples, learning process and prediction results of neural networks will be more developed, and its application in the field of TCM research will be more extensive. The system only tests and predicts the efficacy of a single class of prescriptions. The prediction of the efficacy of multiple combinations of prescriptions will be the next research direction. The design and implementation of the system provide a reference for the research of TCM compound.

4.2 Study on the drug property, flavor, channel tropism of TCM and efficacy of deficiency-nourishing drug based on BP neural network

The theory of TCM is summarized by ancestors refined in the long-term practice. The drug efficacy is a high-level summary of the role of TCM in the treatment process. It is an important part of the theory of TCM. All the doctors of the past dynasties emphasized the holistic view of TCM of "the combination of natural property and efficacy". The drug property-efficacy of TCM together constitute an organic whole. With the development of computer technology, we use modern data mining methods to study the natural property-efficacy of TCM. This helps to explain the mechanism of TCM treatment of diseases. It reveals the essential connotation of modern scientific significance of TCM, and provides an effective theory for the internationalization of TCM compound.

The artificial neural network has the characteristics of self-learning and adaptive function. It can process nonlinear data well, which corresponds to the nonlinear and high-dimensional characteristics of TCM data. This has been widely used in TCM in recent years. Xiang Cuiyu *et al.* combined the near-infrared diffuse reflectance spectroscopy technique with the artificial neural network method to identify rhubarb from different habitats, and the recognition accuracy rate can reach 95%. Liu Xing *et al.* used the near-infrared spectroscopy data of Coix seed from different habitats and varieties as the research object. The learning vectorization neural network was used to establish the classification identification model of Coix seed, and the prediction accuracy rate is greater than 90%. Wang Liqiong *et al.* used the information provided by the HPLC fingerprint of Ephedra sinica to establish a BP-ANN model. This can help to classify ephedra from different habitats. The prediction accuracy rate of the model for different types and different origins of ephedra is 94.4%. Tang Yanfeng *et al.* used Fourier infrared spectroscopy to scan 42 samples of purple radix, and used radial basis neural network to establish a classification model to identify wild purple diced and cultivated purple diced. The recognition rate reaches 95.24%. The research of artificial neural network in the field of TCM has achieved good results. Therefore, we use BP neural network to construct the prediction model of drug efficacy classification of deficiency-nourishing drugs, and take the commonly

used property, flavor and channel tropism of 76-deficiency-nourishing drugs as the research object. Then, we use the Python language to realize the intelligent prediction of the drug efficacy of TCM by computer.

4.2.1 Overview of the four qi and five flavors of TCM

The preface of the "Shen Nong's Herbal Classic" said "drug has five flavors of acid, salty, sweet, spicy and bitter, and four qi of cold, hot, warm and cool". Each TCM has different four -qi and five- flavors. Thus, there are different therapeutic effects. When discussing the drug efficacy, pervious doctors first marked their "qi" and "flavor". The "qi" and "flavor" are one of the important signs to understand the performance of various drugs.

Four qi, also known as four properties, refer to four different medicinal properties of cold, warm, hot and cool. It reflects the tendency of drugs to the yin and yang of the human body and the change of cold and hot. It is an important part of the theory of medicinal properties. Among the four qi, cold and warm are two opposite properties. Among them, the differences between cold and cool, or temperature and heat are not significant. Some books even use the "very cold", "very hot", "slight acid" and "slight bitterness" to describe the drug properties. Besides to the four properties, there is a class of flat drugs, such as licorice, yam, codonopsis, *etc*. The cold and heat limits of these drugs are not obvious, the drug is calm, and the drug efficacy is moderate. Therefore, the doctors have the different view on whether flat should be included in drug properties.

Five flavors, namely, five different kinds of medicinal tastes of sour, bitter, sweet, spicy and salty, some of which have a light or astringent taste. Thus, there are more than five kinds of flavors. However, these five flavors are the most basic medicinal tastes, and we still call five flavors. The first production of the five flavors was tasted by people. It is not only a reflection of the taste of the drug, but also a high-level summary of the role of the drug in the long-term. In addition, the five flavors have the attribute of yin yang wu xing, that is, "spicy and sweet are scattered as yang, sour and bitter are yin, salty is yin, and light taste is vented to yang". Sour taste is wood, bitter taste is fire, sweet taste is earthy, spicy taste is gold and salty taste is water.

The channel tropism indicates the site of drug efficacy, that is, a drug has a special affinity for certain meridians. Therefore, the drug plays a major or special therapeutic effect on the lesions in these sites. Mastering the theory of channel tropism of TCM can effectively improve the accuracy of medication. The theory of channel tropism of TCM is based on the theory of internal organs and meridians, and is based on the specific symptoms and signs of drug treatment. Since the meridians can communicate with the human body inside and outside. Thus, the surface lesions can affect the internal organs through the meridians, and the internal organs can also be reflected on the body surface.

4.2.2 BP neural network principle and training process

(1) The basic principle of BP neural network

The BP (back propagation) neural network is a concept proposed by scientists led by Rumelhart and McClelland in 1986. It is a multi-layer feedforward neural network trained according to the error back propagation algorithm. It is the most widely used neural network.

BP neural network is mainly divided into input layer, hidden layer and output layer, and its structure is shown in Figure 4-6. The input layer is the set of vectors formed by the input data set. The output is the dot product of the input layer and the weight, and then the value is calculated with the activation function. The hidden layer can be one layer or multiple layers, generally no more than two layers. Its input is the dot product scalar of output product of the previous layer and the weight. Its output is the calculation result of the scalar and activation function. The output layer has only one layer, its input is the dot product of the output layer of the previous layer and the weight. The calculated method of output layer is the same as that of the hidden layer. The result of the calculation is the final expected classification weight.

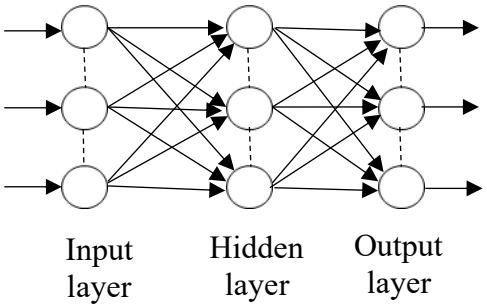


Figure 4-6. The structure of BP neural network.

The training of BP neural network is divided into two steps. One is to calculate the output value of each layer in the propagation process of the previous item according to the current parameter value. The other is to backpropagate the calculation according to the difference between the actual output and the expected output. The error propagation term on one layer combines the partial derivatives of the layer parameters with each layer to update the parameters of each layer. The two steps are repeated until the process converges.

(2) Training process of BP neural network

BP neural network training consists of three phases of the forward propagation process, the calculation of error between the expected and actual classification, and the calculation of the backpropagation process.

(a) Forward propagation process. The input information is passed through the input layer, the hidden layer to the output layer. The actual output value of each layer of neurons is calculated layer by layer.

$$net = w^T o + b \tag{4-1}$$

In Equation (4-1), the weight vector of each layer w is the output vector of the previous layer. For the input layer, it is the training sample set x. b is the threshold value.

Activation function:

$$f(\text{net}) = \frac{1}{1 + e^{-net}} \tag{4-2}$$

(b) The calculation of error between the expected and actual classification. We calculate the error between the actual output and the expected output, and determine if the error is below the tolerance. If it is above the tolerance, backpropagation is performed.

Error vector:

$$\text{error} = d_o - y_o \tag{4-3}$$

Global error function:

$$f_{error} = \frac{1}{2}\sum (d_o - y_o)^2 \tag{4-4}$$

(c) The calculation of backpropagation process. If the forward propagation process fails to obtain the expected output value, the error value between the actual output and the expected output is calculated layer by layer, and the weight is adjusted according to the error value. Dlogit function:

$$f'(\text{net}) = \frac{1}{1+e^{-net}} - \frac{1}{(1+e^{-net})^2} = y(1-y) \tag{4-5}$$

Output Layer Error: we calculate the gradient and differentiation of the output layer of the error backpropagation for updating the output layer weights. The error of the output layer can be written as follows:

$$\frac{\partial e}{\partial w_{ho}} = \frac{\partial e}{\partial y_i}\frac{\partial y_i}{\partial w_{ho}} \tag{4-6}$$

where w_{ho} is the connection weight of the hidden layer and the output layer, and ∂y_i is the actual output of the hidden layer.

Hidden layer error: we calculate the gradient and differentiation of the hidden layer of the error back propagation for updating the implicit layer weights. The error of the hidden layer can be written as follows:

$$\frac{\partial e}{\partial w_{ih}} = \frac{\partial e}{\partial h_i}\frac{\partial h_i}{\partial w_{ih}} \tag{4-7}$$

where w_{ih} is the connection weight of the input layer and the hidden layer, and h_i is the input vector of the hidden layer.

(4) The correction of the weight of each layer

The connection weights are corrected using the gradients and differentials of the neurons in each layer.

Update the hidden layer:

$$w_{ho}^{N+1} = w_{ho}^{N} + \eta\delta_o h_o \tag{4-8}$$

Update the output layer:

$$w_{ih}^{N+1} = w_{ih}^{N} + \eta\delta_h x \tag{4-9}$$

In Equations (4-8) and (4-9), N represents the current number of iterations, η is the learning factor, δ_o is the output layer gradient, and δ_h is the hidden layer gradient. The training process of the entire BP neural network model is shown in Figure 4-7.

4.2.3 Construction of BP neural network model and its application

(1) Quantization of drug properties indicators of TCM

This chapter is based on the sample data of the 76-deficiency-nourshing drugs of the National TCM College planning textbook “Chinese Pharmacy” (Fourth Edition). The model in this chapter mainly describes the relationship between 22 attributes of the four qi, five flavors, channel tropism and drug efficacy by mathematical method. It is necessary to quantify the drug properties of TCM. In this chapter, we use the binary quantification method to characterize the

drug. For example, about the four qi properties of TCM are “cold, hot, warm, cool, and flat”, if there is any property of each TCM, it will be recorded as 1, otherwise it will be 0. In the same way, the four types of efficacy of the five flavors, the channel tropism and the tonics of TCM, namely nourishing qi, yang, blood and yin, are also quantified by binary method. The partial quantified results are shown in Table 4-5.

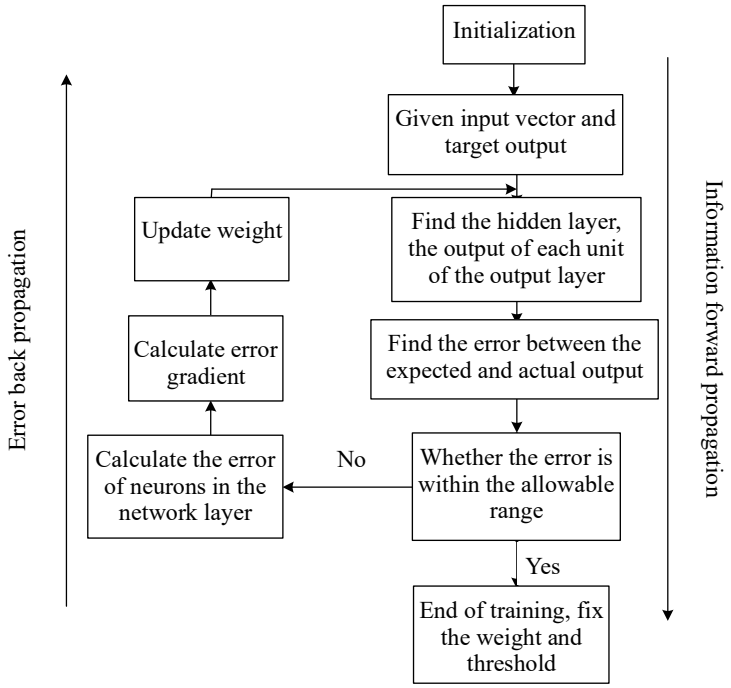


Figure 4-7. The training process of BP neural network model.

(2) Construction and application of BP neural network of deficiency-nourishing drugs

Python has been a popular programming software recently. Because of its rich extended library, simple and easy to learn, and high speed, Python is more and more widely used in data mining. This chapter uses Python’s numpy library to construct a BP neural network model. First, we determine the number of neurons in the input, hidden, and output layers. According to the four qi, five flavors and channel tropism of TCM, 22 parameters are selected as the input unit of neural network in this chapter. The four functions of nourishing TCM, namely nourishing qi, nourishing yang, nourishing blood and nourishing yin, are taken as output unit with output range of 0-1. After reviewing a large number of documents, the number of hidden layers of the model is considered as one. The number of neurons in the hidden layer is determined according to the empirical formula. The model finally determines the number of neurons in the hidden layer is 44. Secondly, the mathematical model of BP neural network for the efficacy of deficiency-nourishing drugs is constructed, trained and tested. According to the number of neurons in the input layer, hidden layer and output layer, and the number of hidden layers, this chapter constructs a three-layer BP neural network with the input node of 22, hidden node of 44, and output node of 4. The activation function uses the Sigmoid function. Because it can well reflect the input and output characteristics of the artificial neural network. We randomly select 64 from 76 types of TCMs in this model as training data, and the remaining 12 types are used as test data.

Table 4-5. Quantitative results of binary values of TCM.

TCM name	Cold	Hot	Warm	Cool	Flat	Sour	Bitter	Sweet	…	Heart	Nourishing qi	Nourishing yang	Nourishing blood	Nourishing yin
1 Donkey hide gelatin	0	0	0	0	1	0	0	1	…	0	0	0	1	0
2 Morinda officinalis How	0	0	1	0	0	0	0	1	…	0	0	1	0	0
3 White hyacinth bean	0	0	1	0	0	0	0	1	…	0	1	0	0	0
4 White peony root	1	0	0	0	0	1	1	0	…	0	0	0	1	0
5 Atractylodes	0	0	0	0	0	0	1	1	…	0	1	0	0	0
6 Lily	1	0	1	0	0	0	0	1	…	0	0	0	0	1
7 Coasiai giehnia root	1	0	0	0	0	0	1	1	…	0	0	0	0	1
8 Flos lablab	0	0	0	0	1	0	0	1	…	0	1	0	0	0
...	...	...	...	...	...	...	...	...	…	...	...	...	...	...
75 Polygonatum	1	0	0	0	0	0	0	1	…	0	0	0	0	1
76 Placenta Hominis	1	0	0	0	0	0	0	1	…	0	0	1	0	0
77 Fluoritum	0	0	1	0	0	0	0	1	…	0	0	1	0	0

(3) Test results of BP neural network of deficiency-nourishing drugs

After 1000 iterations, the weight error is optimal, and the actual output is shown in Table 4-6. The comparison between the expected output and the actual output of the test sample is shown in Table 4-7. As shown Table 4-7, the accuracy rate of the model is as high as 83.33%, and the effect is good.

Table 4-6. The expected output and actual output of test sample.

Name	Expected output				Actual output			
	Nourishing qi	Nourishing yang	Nourishing blood	Nourishing yin	Nourishing qi	Nourishing yang	Nourishing blood	Nourishing yin
Curculigo	0	1	0	0	0.1017	0.9086	0.1683	0.0074
Dendrobium officinale	0	0	0	1	0.1955	0.0069	0.0068	0.9998
Gynostemma	1	0	0	0	0.7272	0.2725	0.1461	0.1741
Ginseng	1	0	0	0	0.9792	0.0261	0.0442	0.0001
Licorice	1	0	0	0	0.9802	0.0005	0.2088	0.3002
Cistanche	0	1	0	0	0.0194	0.9998	0.1911	0.081
Antler glue	0	1	0	0	0.0105	0.9843	0.0522	0.014
Actinolite	0	1	0	0	0.002	0.9999	272	0.0013
Propolis	1	0	0	0	0.7676	0.9889	0.1077	0.2772
Black sesame	0	0	0	1	0.0152	0.3822	0.3375	0.9875
Sea buckthorn	1	0	0	0	0.091	0.1152	0.1753	0.2343
Dendrobium	0	0	0	1	0.1955	0.0819	0.0826	0.9998

Table 4-7. The comparison between expected output and actual output of test.

Name	Nourishing qi	Nourishing yang	Nourishing blood	Nourishing yin
Curculigo	Accurate	Accurate	Accurate	Accurate
Dendrobium officinale	Accurate	Accurate	Accurate	Accurate
Gynostemma	Accurate	Accurate	Accurate	Accurate
Ginseng	Accurate	Accurate	Accurate	Accurate
Licorice	Accurate	Accurate	Accurate	Accurate
Cistanche	Accurate	Accurate	Accurate	Accurate
Antler glue	Accurate	Accurate	Accurate	Accurate
Actinolite	Accurate	Accurate	Accurate	Accurate
Propolis	Accurate	Wrong	Accurate	Accurate
Black sesame	Accurate	Accurate	Accurate	Accurate
Sea buckthorn	Wrong	Accurate	Accurate	Accurate
Dendrobium	Accurate	Accurate	Accurate	Accurate

Note: Accurate rate 11/12 = 83.33%.

4.2.4 Prospect of research on the property, flavor, channel tropism of TCM and efficacy of deficiency-nourishing TCM

The results of this study show that only two of the 48 data corresponding to the efficacy prediction of the 12 TCMs in the test set are wrong, and the prediction accuracy rate is as high as 83.33%. At the same time, the model and method have certain practical value in the classification and prediction of the efficacy of TCM. The model can be extended to the study of the efficacy of TCM. This model studies the relationship between the four qi as well as five flavors and the efficacy of TCM. The data set used is the deficiency-nourishing TCM. The properties of most deficiency-nourishing TCM are warm, sweet, and belong to the channel

tropism of liver, spleen, kidney, lung and heart. This results in a large similarity of the data set, which requires more precise model parameter settings. Moreover, the relationship between the efficacy of TCM and the components of TCM is non-linear, and the mechanism is complicated. It is difficult to achieve comprehensive analysis by traditional research methods. Artificial neural network has strong self-learning ability, self-adaptive ability and the characteristics of dealing with nonlinear relationship. It is effective to use BP neural network to systematically study the efficacy of TCM.

Because of the complexity of the composition and efficacy of TCM, we will consider more independent factors in the future to improve the accuracy rate of model classification. For example, in addition to the four qi and five flavors, TCM has the toxicity and other influencing factors. In the four qi and five flavors, the "cold" can be subdivided into "very cold" and "slight cold". This model divides it into cold. In addition, the data quantization used in this model should be automatically quantized. This will be more conducive to the processing of large data sets. Therefore, a faster quantization method can be adopted in the future research. The types of training data sets can be expanded to improve the robustness of the model and promote the wider application of this model in TCM research.

4.3 Prediction of the efficacy of TCM compound based on BP neural network

TCM compound is a bridge connecting basic with clinical of TCM. It is the main method and characteristic of TCM for treating diseases. There are many methods for studying TCM compound preparations, and most of them use decomposed analysis methods, such as medicine versus research method, single-flavor research method, *etc*. However, because of the variety of compound preparations, the differences in drug properties and compatibility are significant, and they belong to complex nonlinear problems. This kind of method cannot be used to study the efficacy of TCM compound preparations from a holistic perspective. In addition, there is no clear mathematical formula to express the relationship between the compound efficacy and the drug property. Artificial neural networks have been widely used in the field of medicine in recent years. Among them, BP neural network can solve the nonlinear problem of the characteristics and efficacy of compound preparations with its good nonlinear fitting function. To seek a better research method, this chapter takes the medicinal characteristics data of the nourishing compound prescription as the research object, and aims at the complex and uncertain relationship between the drug properties and efficacy of the TCM compound, and uses the theoretical viewpoint of drug properties to conform the traditional thinking of TCM. We use BP neural network to explore the intrinsic relationship between the four properties, five flavors, channel tropism and their efficacy in TCM. The GUI simulation is designed to study the efficacy of TCM compound and the modernization of TCM.

4.3.1 Overview of neural network and TCM theory

Artificial Neural Network (ANN) is an information processing system that simulates the function and thinking of the human brain. The Back-Propagation Network is the most common

multi-layer parallel information feedforward network used in artificial neural networks. It consists of three parts of input layer, hidden layer and output layer. There is a teacher's learning style, and its main running process has the forward transmission of signals and the back propagation of errors. The running process is in the forward transmission of information. The signal is processed and transmitted layer by layer. Then, the error between the predicted output and the target output is checked to meet the preset accuracy requirement. Otherwise, it returns along the original path and transfers to the back propagation by adjusting the weights and thresholds between the layers. Therefore, the actual output of the network is constantly approaching the target output. It has a high degree of adaptability and self-learning, as well as nonlinearity. This will enable it to better simulate human intelligence and behavior, and its ability to deal with nonlinear problems. This is better than the traditional pharmacological theory and mathematical statistics analysis methods. Therefore, the BP neural network method is used to explore the potential relationship between the drug properties and efficacy of TCM compound from the overall perspective. This provides technical support for deepening the relationship between the medicinal properties and efficacy of TCM compound.

The drug properties of TCM s are closely related to their efficacy. The efficacy is an important source of drug formation. For example, the ancient “Shen Nong’s Herbal Classic” records “treating cold with hot medicine, treating hot with cold medicine”, and “Yellow Emperor’s Internal Classic” records “spicy scans, sour converges, sweet slows, bitter strengthens, salty softens.” In the same way, the complex drug properties and efficacy of TCM compound are also closely related. It is an organic whole formed by the TCM to determine the pathogenesis by dialectics, and select drugs according to the formula structure and drug compatibility principles. As the main content of TCM, the four qi, five flavors and channel tropism can explain the efficacy of TCM from different aspects. Therefore, to comprehensively study the efficacy of TCM compound, it is necessary to consider all aspects. Therefore, in this chapter, we analyze the data of properties, flavor and channel tropism of nourishing TCM compound, and use BP neural network method to predict the efficacy of it. The three-layer topology of the neural network is shown in Figure 4-8.

4.3.2 Construction of BP neural network model for the efficacy of nourishing prescriptions

(1) Numerical quantification of the drug properties and efficacy indicators of TCM prescriptions

In this chapter, we take the nourishing TCM prescriptions library as the sample data, including the composition table of the nourishing TCM prescription, the compatibility table of jun-chen-zuo-shi of prescriptions, and the compound information quantification table. The drug properties of TCM are described by text in the theory of TCM. Thus, in this chapter, mathematical methods should be used to study the drug properties and efficacy of TCM. The drug properties must be processed into numerical types. Therefore, we use the method of binary quantification to treat the drug properties. For example, the five drug properties of TCM are cold, hot, warm, cool, and flat. If the medicine has this property, it will be recorded as 1, otherwise it will be recorded 0. In the same way, the five flavors and 12 items of TCMs are

classified as follows. When they have a certain medicinal flavor or channel tropism, they are recorded as 1, otherwise they are 0. The nourishing compound mainly has four functions of nourishing qi, nourishing blood, nourishing yin and nourishing yang. If one kind of TCM has a certain function, it is recorded as 1, otherwise it is recorded as 0. Then the quantification of the compound is based on the binary quantification of the single drug, according to the compatibility characteristics of the TCM. Each corresponding attribute value of the different nourishing prescription are calculated by taken weighting calculation methods using the jun, chen, zuo of 0.5, 0.3, and 0.2, respectively. The role of inter-drug interactions is not considered in this chapter. Table 4-8 shows the partial results of the quantified process of the 126 nourishing TCM compounds included by the above method.

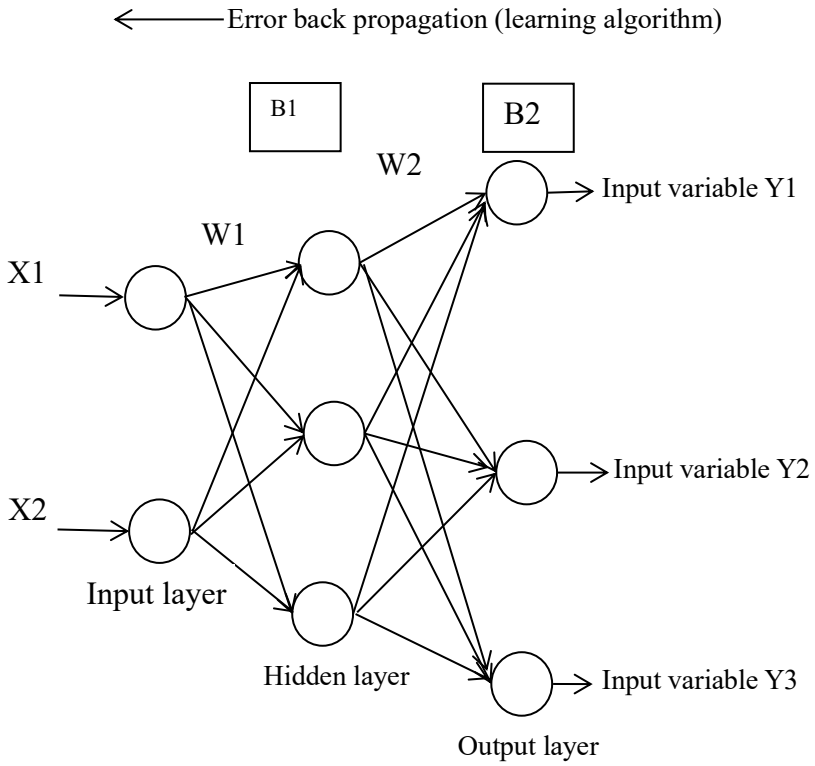


Figure 4-8. Three-layer topology of BP neural network.

(2) BP neural network construction

(a) Determination of the number of neurons in the input, output and hidden layers. The input layer of this chapter is the basic attribute characteristic value of TCM, including the 22 parameters of four properties, five flavors, and channel tropism of TCM (22 neurons). The output layer is the four functional neurons that represent the four functional indicators of the nourishing qi, nourishing blood, nourishing yin and nourishing yang of the nourishing compound TCM, and the output range is from 0 to 1. The number of layers of the hidden layer is one. Many studies have shown that the neural network with one hidden layer can represent any continuous function with arbitrary precision. The number of neurons in the hidden layer is selected according to the comparison of the empirical formula of Kolmogorov's law and the simulation results of the last hidden layer. After continuous testing, the number of neurons

finally determined by the hidden layer of the model is 45.

Table 4-8. Compound quantification results.

No.	Name	Warm	Hot	Cold	Cool	Flat	Stomach	Nourishing qi	Nourishing Blood	Nourishing Yin	Nourishing yang
001	Pure ginseng decoction	1	0	0	0	0	0	1	0	0	0
002	Qiongyu cream	0.5	0	0.3	0	0.2	0.15	1	1	0	0
003	Ginseng decoction	0.5	0.3	0.1	0	0.1	0.4	1	0	0	0
004	Zanyu bolus	0.33	0.38	0	0	0.3	0.08	0	0	0	1
005	You gui wan	0.3	0.33	0	0	0.37	0	0	0	0	1
006	You gui-yin	0.1	0.5	0	0	0.4	0.1	0	0	0	1
007	Zhengyang lilao decoction	0.92	0.04	0	0	0.04	0.34	1	0	0	0
	……	.....									
123	Shengxian docoction	0.5	0	0.43	0	0.07	0.37	1	0	0	0
124	Angelica Decoction for Replenishing Blood	1	0	0	0	0	0	0	1	0	0
125	Deficiency-nourishing astragalus saponin decoction	0.85	0.05	0.05	0	0.05	0.2	1	0	0	0
126	Da ying jian	0.8	0.04	0.04	0	0.12	0.04	0	1	0	0

(b) Construction, testing and training of network models. The constructed BP neural network is a three-layer structure with 22 input neurons, 45 hidden layer neurons, and 4 output neuron nodes. The transfer function between the layers is a sigmoid function with non-linear features and differentiable characteristics, because it can well reflect the characteristics of the input and output of artificial neurons. In this chapter, 116 sets of data from 126 compound data samples will be used as training samples, and 10 sets of data will be used as test samples.

(c) Prediction results. In this chapter, 126 kinds of nourishing compound data are taken as samples, among which 1-116 groups are training samples, and 117-126 groups are test samples. The property, flavor, channel tropism of the compound are taken as the input layer data, and the four functions of the nourishing compound of nourishing qi, nourishing blood, nourishing yin and nourishing yang are taken as the output layer data. Then, the constructed neurological model of the nourishing compound is performed network training. After 11 iterations, the expected error value is 0.001, and the network training is completed. The training error curve is shown in Figure 4-9. The output result of final 10 sets predicted data is bpoutput. The prediction effect of the model is satisfactory, and the accuracy rate is as high as 92.5%. The expected output and actual output of the nourishing compound test sample are shown in Table 4-9. The comparison between the expected output and the actual output of the nourishing compound test sample is shown in Table 4-10.

Table 4-9. Expected output and actual output of the nourishing compound test samples.

Name	Expected output				Actual output			
	Nourishing qi	Nourishing blood	Nourishing yin	Nourishing yang	Nourishing qi	Nourishing blood	Nourishing yin	Nourishing yang
Yangjing zhongyu decoction	0	1	0	0	0.000	0.6497	0.0000	0.0124
Qifu decoction	1	0	0	0	0.0382	0.0002	0.0000	0.0000
Modified wuyao decoction	0	0	0	0	0.0012	0.0000	0.0000	0.0003
Guilu erxian glue	0	0	1	0	0.0000	0.0001	0.9934	0.2179
Changchun dan	1	1	0	0	0.0000	0.9999	0.0002	0.0002
Lingshu siwu decoction	0	1	0	0	0.0123	0.0005	0.0000	0.0000
Shengxian decoction	1	0	0	0	1.0000	0.0000	0.0000	0.0000
Angelica decoction for replenishing blood	0	1	0	0	0.0215	0.8322	0.0000	0.0000

Table 4-10. The comparison between expected output and actual output of nourishing compound test samples.

Name	Nourishing qi	Nourishing blood	Nourishing yin	Nourishing yang
Yangjing zhongyu decoction	accurate	accurate	accurate	accurate
Qifu decoction	wrong	accurate	accurate	accurate
Modified wuyao decoction	accurate	accurate	accurate	accurate
Guilu erxian glue	accurate	accurate	accurate	accurate
Changchun dan	wrong	accurate	accurate	accurate
Lingshu siwu decoction	accurate	wrong	accurate	accurate
Shengxian decoction	accurate	accurate	accurate	accurate
Angelica decoction for replenishing blood	accurate	accurate	accurate	accurate
Buxu huangqi decoction	accurate	accurate	accurate	accurate
Da ying jian	accurate	accurate	accurate	accurate

4.3.3 GUI simulation experiment results of efficacy prediction of TCM compound

There is relatively few analysis software for TCM data in the software market in China. People commonly use common statistical analysis. The main feature is that the interface is relatively complicated and needs to be more professional. The statistical background has made it difficult for many TCM scholars to master. In addition, the use of SAS and MATLAB software requires scholars to master certain computer programming techniques. Therefore, the research on the drug property and efficacy of TCM by using neural network method needs to design a tool that is easy to learn, and operate. This can quickly construct a network model, and save the network model. The model output is more intuitive. In addition, because the above network prediction model is implemented in MATLAB2013a through network programming.

It is not convenient whether it is applied to the actual compound power prediction. Therefore, this chapter is based on the graphical user interface GUIDE toolbox in MATLAB. A graphical user interface (GUI) based on BP neural network for predictive analysis of compound TCM efficacy is designed. Thus, users do not need to understand how to write and run the whole program. They only need to interact with the interface correctly, such as simple input data and set various parameters of the network. Then, we can click each indicator button to obtain the prediction analysis results of TCM compound neural network.

(1) Detailed module function of efficacy prediction of TCM compound

The block diagram of the efficacy prediction module of TCM compound is shown in Figure 4-10. The detailed description of each module in the interface is as follows.

(a) Data input module. In this module, the user can manually input the sample data that needs to be trained and tested, or import the corresponding sample data from the outside through the import button. The percentage of the training set and test set data in the sample should be set by the user.

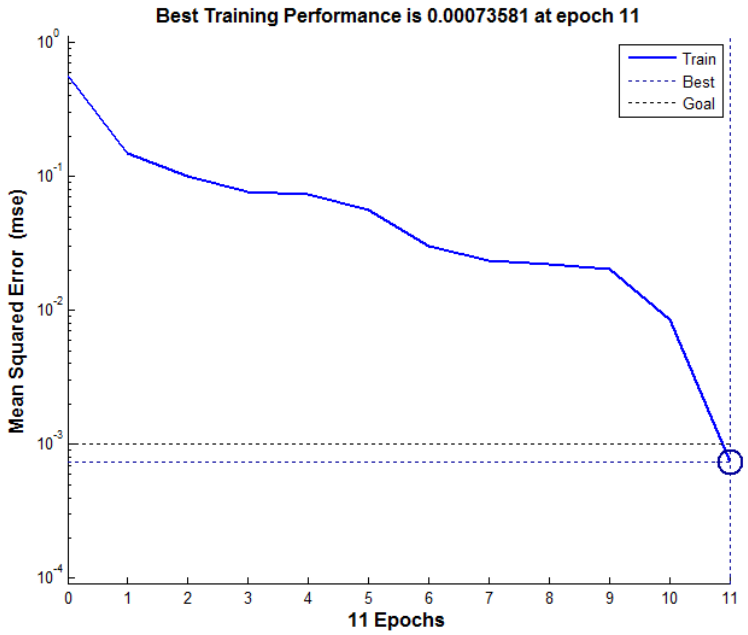


Figure 4-9. Training error curve.

(b) Network parameter setting module. This module is used to set the network parameters for network construction, including network structure parameters and network training parameters. The structural parameters of the network include the number of neurons in the hidden layer, the transfer function between the input layer and the hidden layer, the transfer function between the output layer and the hidden layer, and the training function. The training parameters of the network include the maximum number of trainings, the display time interval, the training target error, and the learning rate. Users can set parameters by themselves or use default parameters. The default network structure is a three-layer neural network, in which the number of hidden layer neurons in the structural parameters is 45. The transfer functions between the input layer and the hidden layer, between the output layer and the hidden layer are Logsig and tansig, respectively. The training function is trainlm. The maximum number of

training sessions in the training parameters is 100 with the time interval of 10. The target error is 0.001, and the learning rate is 0.01. The block diagram of prediction module of the TCM compound efficacy is shown in Figure 4-10.

(c) Implementation of the network. The implementation of the network trains the network through the training button. The predict button is used for performing the network simulation test and simultaneously outputs the simulation result data. The function of saving button of the network is to save the trained network model for the future use. Finally, we launch the interface through the launch button.

(d) Result output module. Users can see the prediction data of the network and the accuracy of its prediction under this module.

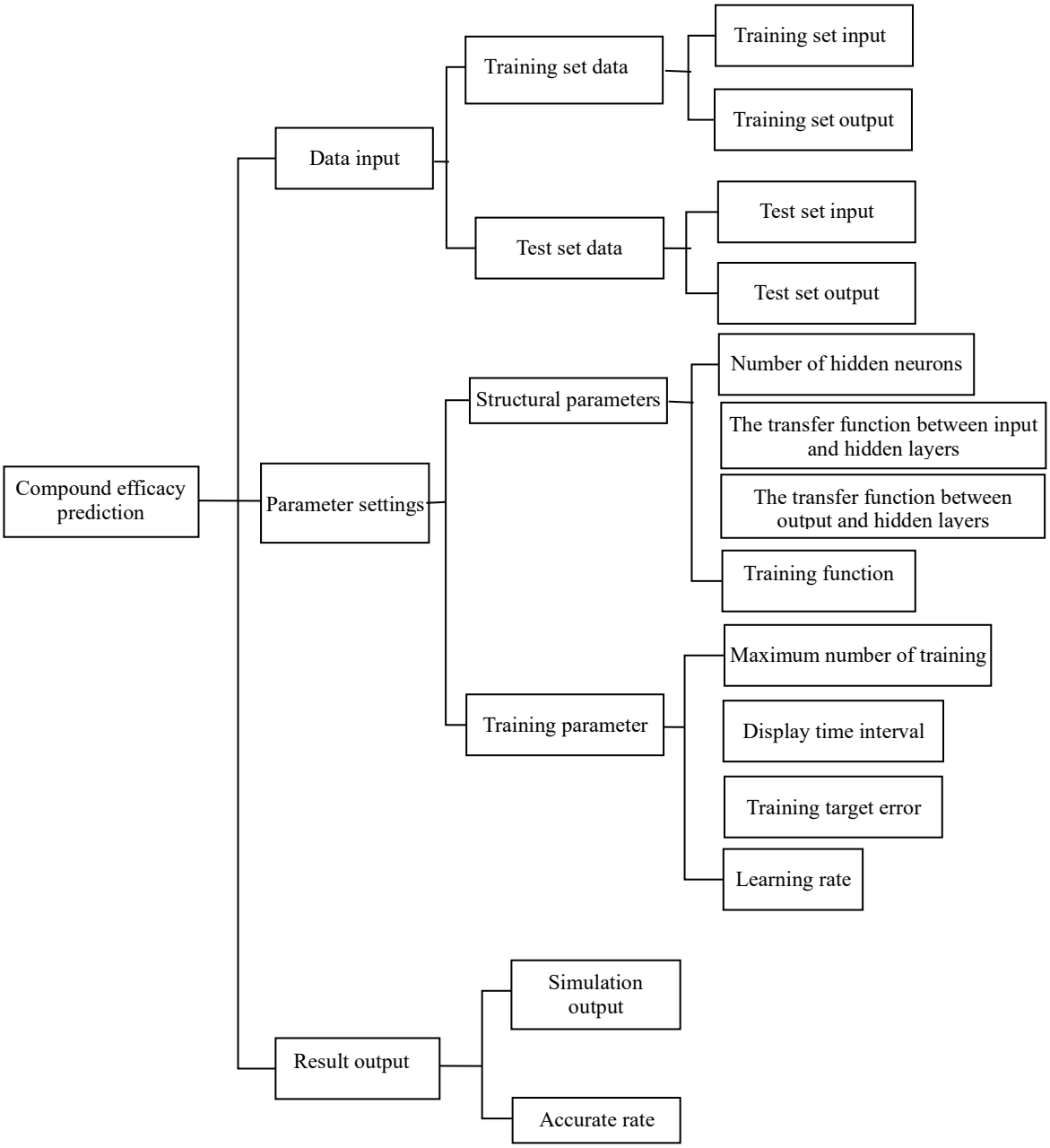


Figure 4-10. Block diagram of prediction module of TCM compound efficacy.

(2) Realization of GUI simulation interface for efficacy prediction of TCM compound

In this chapter, data are imported from an instance sample with 126 sets of data, and each contains 22 input eigenvalues. The first 116 sets of data are used as the training set, the next 10 sets are used as the test set data. The network's structural parameters and training parameters use the default values. Then, we click the training button to train the network. After the network training is completed, we click the predictive button to perform the network simulation test. After the network simulation is completed, the predicted output value of the network is shown in the "Result Output". Because this chapter does not constrain the format of the data to display in the network, the display of the data in the network is the default output of Matlab. By comparing the actual value with the predicted value, the percentage of the accurate rate of the network prediction is shown in the "predictive accurate rate". The interface diagram of the simulation results of efficacy prediction of TCM compound is shown in Figure 4-11.

Figure 4-11. Interface diagram of simulation results of TCM compound efficacy prediction.

4.3.4 Discussion about the efficacy prediction of TCM compound based on BP neural network

The results of this study show that there are 40 data in the prediction of the efficacy of the 10 compound preparations in the test set, of which 37 are accurate. The prediction accuracy rate is as high as 92.5%. The prediction results of this model are high in accuracy and good in simulation performance. The model and method have certain practical value. This is precisely because the TCM compound is used in combination with a variety of TCMs. The relationships between each component and efficacy as well as between the interaction among components are nonlinear. Its mechanism is complicated. Thus, it cannot be done with traditional research methods. Using the powerful self-learning, adaptive and nonlinear mapping capabilities of artificial neural networks, based on the compatibility principle of Jun-chen-zuo-shi and the weighted operation method of the proportion, the prediction model of drug property and efficacy of TCM compound based on BP neural network is constructed. This model is used to analyze the complex nonlinear relationship between the property, flavor, channel tropism and the efficacy of the TCM compound. It is effective to systematically study the whole compound.

However, because of the complexity of the TCM compound, we do not include the relationship among compound drugs in this chapter. Although the collected drug sample data has certain representativeness, it is not enough in quantity. Therefore, it may have some influence on the prediction of drug performance. Therefore, in the future research, we will further consider the relationship among drugs to find a better weighting method to process the compound data. In addition, we will include more representative sample data for training. This can promote the refinement and accuracy rate of the model. Different types of neural network methods will be further incorporated to realize the property-efficacy analysis of TCM compound to enrich the function of the system.

References

1. L. Zhao, J. Zhang, S.L. Nan, etc. Discussion on the Teaching Design of the Course "Prescription of Chinese Materia Medica" under Chinese Pharmacy Major. *Journal of Shaanxi University of Chinese Medicine* **2017**, *40*, 134.
2. J.Y. Yu, S.Y. Gu, Y.Q. Wen, *et al*. Discussion on the four inverse dispersive sources based on the framework theory of prescriptions. *Lishizhen Medicine and Materia Medica Research* **2017**, *28*, 2493.
3. Y.Q. Wen, B. Jia, T. Shen, *et al*. Analysis on the concept of prescription efficacy. *Medical Innovation of China* **2015**, *12*, 96.
4. X. Zhang. Application of artificial neural network in the study of traditional Chinese medicine. *Journal of Guangdong Pharmaceutical University* **2011**, *27*, 653.
5. Y. Luo, B. Lin, C.X. Wen. Research on algorithm of correlation model of chinese physique and physical examination index based on neural network. *Lishizhen Medicine and Materia Medica Research* **2018**, *29*, 763.
6. M.C. Yao, Y.L. Zhang, Y.M. Yuan, *et al*. Study on the Prediction of the Effect Attribution of the Deficiency-Nourishing Drugs Based on the Quantification of TCM Drug Properties. *Journal of Beijing University of Traditional Chinese Medicine* **2004**, *27*, 7.
7. H.B. Zhou, Y.H. Zhou, L.J. Zhu. Implementation and comparison of improving BP neural network based on MATLAB. *Computing Technology and Automation* **2008**, *27*, 28.
8. W.W. Li, X.Y. Zhang, Y. Yan, *et al*. Effects of Traditional Chinese medicine compound based on bp neural network. *Guiding Journal of Traditional Chinese Medicine and Pharmacy* **2016**, *22*, 38.
9. Y.Y. Wu, X.Y. Zhang. Research on Quantization of Traditional Chinese Medicine Information. *Journal of Jiangxi University of Traditional Chinese Medicine* **2008**, *20*, 56.
10. X.D. Wu. Numerical model of efficacy of traditional Chinese medicine based on BP neural network and subjective evaluation. *Gansu Science and Technology* **2011**, *27*, 19.
11. W. Ye, X. Li, Q.G. Liao, *et al*. BP for character, taste, channel of Chinese medicine and antihyperlipidemia. *Computer Engineering and Applications* **2008**, *44*, 222.
12. J. Peng, C.J. Tang, T. Zeng, *et al*. A Chinese traditional medicine prescription effect reduction algorithm based on artificial neural network and property distance matrix. *Journal of Sichuan University (Engineering Science Edition)* **2006**, *38*, 92.

13. F. Yuan, L.L. Ren, H.M. Jiang, *et al*. Application of MATLAB neural network toolbox in runoff simulation. *Yangtze River* **2003**, *34*, 38.
14. P. Li, L.K. Zeng, A.Z. Tax, *et al*. Design of forecast system of back propagation neural network based on MATLAB. *Computer Applications and Software* **2008**, *25*, 149.
15. Y. Li, W. Li, F.Z. Xue, *et al*. Discrimination of properties of Chinese traditional medicines based on an artificial neural network. *Journal of Shandong University (Health Sciences)* **2011**, *49*, 57.
16. Q.P. Mai, X. Li, Y.H. Wu, *et al*. BP model for relation between anti-aging and four natures, five flavors and meridian tropism of Chinese prescription medicine. *China Journal of Chinese Materia Medica* **2010**, *35*, 337.
17. Y. Li, X. Feng, W. Song, *et al*. Progress in research methods of medicinal properties of traditional Chinese medicine. *Medical Information* **2018**, *31*, 38-40.
18. D.X. Mao. Three Views on the Holistic View of Traditional Chinese Medicine. *China News of Traditional Chinese Medicine* **2017**, *4*.
19. R.R. Zhou, Yan Runhong, Wang Liping, *et al*. Application of data mining technology in scientific problems of TCM prescriptions research. *China Journal of Traditional Chinese Medicine and Pharmacy* **2018**, *33*, 4016-4020.
20. C.Y. Xiang, X.Y. Song. Near-in Identification of rhubarb by NIR and ANN. *Lishizhen Medicine and Materia Medica Research* **2012**, *12*, 3168-3169.
21. X. Liu, D.Z. Mao, Z.W. Wang, *et al*. Rapid identification of Coix seed varieties by near infrared spectroscopy. *Spectroscopy and Spectral Analysis* **2014**, *34*, 1259-1263.
22. L.Q. Wang, Q. Fan, Z.K. Yi, *et al*. Identification of different ephedra plants using HPLC fingerprints in combination with back-propagation artificial neural network and discriminant analysis. *Journal of Southwest China Normal University (Natural Science Edition)* **2012**, *37*, 73-77.
23. Y.F. Tang, Z.Z. Hou, Z.B. Wang, et al. Identification of Viola philippica Car. based on infrared spectrum and radial basis function neural network. *Hubei Agricultural Sciences* **2014**, *53*, 132-134.
24. W.W. Li, X.Y. Zhang, Y. Yan, *et al*. Effects of traditional Chinese medicine compound based on BP neural network. *Guiding Journal of Traditional Chinese Medicine and Pharmacy* **2016**, *22*, 38-41.
25. J. Wang, N. Zeng, H.L. Xia, *et al*. New thoughts on the theoretical research model of traditional Chinese medicine. *Journal of Traditional Chinese Medicine* **2013**, *54*, 99-102.
26. R. Jin, B. Zhang. Analysis on complex characteristics of traditional Chinese medicine property theory. *China Journal of Chinese Materia Medica* **2012**, *37*, 3340-3343.
27. S.H. Chen, G.Y. Lu. Study on the properties of traditional Chinese medicine at the level of "combination of nature and taste". *Pharmacology and Clinics of Chinese Materia Medica* **2008**, *4*, 58-62.
28. X.H. Xiao. Speculating about scientific investigation on nature of Chinese Materia medica. *Chinese Traditional and Herbal Drugs* **2008**, *4*, 481-484.
29. Z.G. Wang. Modern research on the theory of traditional Chinese medicine: problems, ideas and methods. *Journal of Shandong University of Traditional Chinese Medicine* **2011**, *35*, 195-198.

30. Y.S. Zhong. The Chinese Materia medica. China press of traditional Chinese medicine: Beijing, China, 2016.
31. X. Shen, L. Wang, D. Han. Application of BP neural network optimized by artificial bee colony in intrusion detection. *Computer Engineering* **2016**, *2*, 190-194.
32. H. Das, A.K. Jena, J. Nayak, *et al*. A Novel PSO Based Back Propagation Learning-MLP (PSO-BP-MLP) for Classification. Springer: India, 2015; pp. 333-344.
33. P.L. Nie, D. Fei, H.J. Wang, *et al*. Short-term power load forecasting based on EMD-BP neural network. *Control & Instruments in Chemical Industry* **2016**.
34. Z, Zhao, Q. Xu, M. Jia. Improved shuffled frog leaping algorithm-based BP neural network and its application in bearing early fault diagnosis. *Neural Computing & Applications* **2016**, *27*, 375-385.
35. X.J. Liu, S.B. Su. Study of combination methods for formula composition of Chinese herbal medicines and their components. *Journal of Chinese Integrative Medicine* **2009**, *7*, 6011.
36. X. Lu. The Establishment of the Chinese Herb and the Research in the Properties of Anns. Ph.D. Thesis, East China Jiaotong University, Jiangxi, China.
37. L.M. Wang, Q.X. Fu, B.H. Kong, *et al*. Evaluation of artificial neural networks models for predicting ovarian malignancy in patients with adnexal masses. *Shandong Medical Journal* **2007**, *47*, 19-21.
38. Z.W. Shang. Shennong Materia Medica Note. Academy press: Beijing, China, 2008.
39. D.H. Tian. Huang Di Nei Jing Su Wen. People's medical publishing house: Beijing, China, 2005.
40. L.B. Qu, B.R. Xiang, D.K. An. Artificial neural network technology and its application in the analysis of pharmaceutical compound preparations and traditional Chinese medicine. *Chinese Journal of Pharmaceutical Analysis* **1996**, *16*, 201-203.
41. B. Tang, L. Yang, J. Du. Study on the identification of Chinese herbal medicine rhubarb by using improved artificial neural network. *Lishizhen Medicine and Materia Medica Research* **2009**, *20*, 1019-1021.
42. Z.H. Xu, W.F. Yang, W.J. Jiang, *et al*. Improvement and application of the learning algorithm of BP neural networks. *Computer Engineering & Science* **2004**, *26*, 621.
43. M.C. Yao, Y.J. Qiao, Y.M. Yuan, *et al*. Study on the pharmacodynamic classification of traditional Chinese medicine based on artificial neural network. *China Journal of Chinese Materia Medica* **2003**, *28*, 689-691.
44. Y.H. Yin, Y.Q. Jin, Z.J. Yi. Application of Artificial Neural Network in Modernization Research of TCM. *Guiding Journal of TCM* **2006**, *12*, 83-85.

Chapter 5 Research and System Design of Traditional Chinese Medicine Data Mining Based on Strategy Pattern

The information resources of TCM have become an important strategic resource for competing. Strengthening the acquisition of information resources of TCM enables us not only to obtain valuable information, but also to find our precise positioning in the fierce market competition. This provides a basis for making decisions quickly and accurately, accelerates the global dissemination, share and utilization of TCM information, speed up the conversion of effective information into knowledge, and promote the rapid update of TCM information knowledge. The use of information resources of TCM not only promotes the exchange of information among TCMs, but also broadens the field of information resources services of TCM. This enhances the level of information resources services of TCM and the utilization rate of information of TCM. In addition, this increases both social and economic benefits. However, the statistics indicate that the overall utilization rate of information resources of TCM is not high. In today's information society, the main focus is on an information war. The security of information is particularly important, as is the Traditional Chinese Medical Information Security (TCMIS). As "three points depended on technology and seven points depends on management", it is conceivable that how to manage information security is very important. There are many factors affecting the TCMIS, such as software vulnerabilities, misuse of TCM information resources, environmental impacts, human causes, viruses and hacker attacks, all of which are serious threats to the TCMIS. Gradually, people have realized the important role of TCMIS in the management of medical and pharmaceutical industries. The data mining of TCM provides knowledge and decision-making for TCM database or collection of TCM data. It is an important part of decision support system of TCM. TCM is important to the Chinese medical and health industries. Many clinical and scientific practices, as well as ancient literatures, have accumulated rich data on TCM, such as science, law, prescription, medicine, *etc.* How to effectively use these TCM data, how to find the hidden relationships among these massive data, and how to mine the potential knowledge in the data are important topics for improving the use value of TCM. To solve this problem, we adopt the method of data mining of TCM based on the strategic pattern idea. This mining method is different from pure data mining. It can solve problems that cannot be solved by using pure data mining methods, which improves the effect of data mining in TCM. The strategy pattern idea has established the one-to-many, many-to-one, many-to-many solutions among different TCM problems and different data mining methods, and broadened the way of TCM data mining. In this chapter, we will provide a new idea for the development and utilization of TCM based on the above research.

5.1 Development and utilization of TCM information resource

The term "Information Resources Management" (IRM) was formed in the late 1970s, and

was first used in government departments of the United States. Since then, it has rapidly been applied to businesses, research institutes, and institutions of higher learning. After more than 30 years of development, its theoretical and practical research has expanded rapidly around the world. As a part of social information resources, TCM information resource inherits the commonality of information resources and have its characteristics. The management of information resources of TCM is a product of the combination and mutual intersection of TCM and information technology. It mainly obtains real and reliable data through the acquisition, analysis and screening of information resources of TCM, which provides the basis for the development of TCM enterprises, and the support for the requirement of TCM information of users. Nowadays, the value of TCM information resources is increasingly reflected. This is because the public now pays more attention to their physical and mental health, disease prevention and lifestyle optimization than ever before. Thus, the demand for health information is becoming more and more urgent.

5.1.1 Acquisition of TCM information resource

(1) The practical significance of the acquisition of information resources of TCM

The current society is already an information society. With the development of information-related technologies, the information resources of TCM are growing rapidly. The way of obtaining them is constantly changing. The integration of information resources of TCM is beneficial to doctors searching for relevant literatures of TCM in their offices and wards covered by network. However, without the acquisition of TCM information resources, there will be no integrated organization of TCM information resources, and no sharing of TCM information resources. Therefore, the acquisition of TCM information resources is the focus, as well as the starting point and destination of TCM informatization. In addition, the acquisition of TCM information resources is a prerequisite for the management and use of TCM information resources.

(2) Problems in the acquisition of information resources of TCM

We still have some problems in the process of obtaining information resources of TCM. These problems will affect the quality of TCM information resources. We have to find solutions to solve these problems.

(a) The recognition degree about the accuracy rate of obtained information resources of TCM is not enough. We must ensure the authenticity of the obtained TCM information. Otherwise, we will waste our energy. If we obtain wrong information about TCM, we will make a wrong decision and cause serious consequences. We must pay attention to the channels of TCM information sources and the way that TCM information is obtained. This is an important aspect for the acquisition of TCM information resources. To obtain accurate information, we must repeatedly verify to ensure the reliability of it.

(b) No clear purpose before obtaining TCM information. Sometimes, people's access to information resources of TCM is blind, and there is no clear purpose. This may result in that no information is obtained. This is because the society is an information explosion society, and various kinds of information are different. It is impossible to take all the information. Therefore, we must have a clear purpose for the acquisition of information resources of TCM according to

actual requirements. Different from the general information resource acquisition, the most important purpose of the acquisition of TCM information resources is to meet the needs of specific people. This means that the acquisition of TCM information resources must be targeted.

(c) Ignoring the update of TCM information. Nowadays, the society is changing every day. The information of TCM is even more so. The information resources of TCM have a certain time. One information is more valuable at this time, it may lose its value tomorrow. Many people are not aware of this change. We should understand that the value of TCM information resources is mainly to be able to be used in a timely manner. The information resources of TCM can only be given the greatest extent, if they are transmitted to the users in the first time. The acquisition of information resources of TCM requires us to have a keen insight, and timely discover hidden and valuable information on TCM.

(d) Lacking systemicity for the collection of information on TCM. The information dissemination has the discrete and intermittent characteristics. Therefore, in the process of collecting TCM information, it is necessary to broaden the source of information. We must pay attention to the accumulation of TCM information, and strengthen the systematic characteristic of TCM information to ensure the quality and adequacy of the collection. In addition to querying, retrieving, investigating and sampling various information sources of TCM, we need to systematic collect some special information to make up for the shortcomings in our conventional collection methods.

(e) The lack of comprehensive consideration for the acquired information resources of TCM. Before obtaining information resources of TCM, we must comprehensively consider and strive. Only a comprehensive access to TCM information can make the most realistic decisions. The comprehensiveness of TCM information acquisition requires us to collect relevant information around. In addition, we must pay attention to the comprehensiveness of information carriers and TCM information types to avoid omissions. Of course, the information on TCM that we have obtained cannot be absolutely comprehensive and complete under the circumstances. Therefore, making scientific decisions based on incomplete information of TCM is a problem that we should seriously consider.

5.1.2 Management of TCM information resources

(1) Strengthening the necessity of management of information resources of TCM

TCM information is the blood of the TCM market. The imperfect information of TCM will directly affect the orderly operation of the mechanism of TCM market. Strengthening the management of TCM information resources and eliminating the barriers to TCM information are important to improve the TCM market. By increasing the development and utilization of TCM information resources, we ensure the two-way flow of TCM information and make it more effective in the TCM market. The continuous deepening of the management of TCM information resources and the implementation of TCM information resource sharing can break the closed state of TCM information. The management of TCM information resources is a gradual deepening process. With the improvement of people's understanding of TCM information, the thinking of management of TCM information resource will become more and more abundant, and the management methods will become more and more diverse.

(2) The status in the management of TCM information resources

According to research on the management of TCM information resources, there are several key issues that need to be resolved. These issues are affecting the sustainable development of the TCM information industry.

(a) The low quality of talents in TCM information. The development of TCM information industry is inseparable from talents. The quality of talents is directly related to the sustainable development of TCM information industry, and directly affects the competitiveness of TCM enterprises. The training of talents in TCM information is not enough. The training methods are not appropriate. The awareness of self-cultivation is relatively weak, which has caused the phenomenon of low quality of talents in this field. In general, we must attach importance to the cultivation of talents in the medical information industry, establishing and improving the personnel training mechanism. Thus, the graduates of TCM majors have the ability to retrieve, manage, and utilize TCM information through the TCM database or other information resources. Thus, we can solve problems in medicine and related fields, make correct decisions, establish a nationwide library of information technology talents for Chinese medicine, continuously strengthen international exchanges, gradually integrate China's training model and examination system with international standards, and strive to cultivate internationally advanced technology talents of TCM information.

(b) Not enough researches on processing of TCM information database. With the rapid development of the information industry, the construction of the TCM information database has become an important part of the development of the information industry. China has invested a certain amount of energy in the development of the TCM information industry in recent years. There is no unified planning for the construction of the TCM database. Researchers are searching for information. It is impossible to obtain all the information in one database. In addition, the development and utilization of information products of TCM are lagging behind. The main reason is that the development of information resources database of TCM lacks the mechanism for cooperation between TCM information departments. The content of TCM information database is relatively simple. The repetition is quite serious, and the depth of processing is deep. It is not enough to form a comprehensive or professional TCM information resource database that can be directly utilized. In the future, research and development should focus on this type of database to meet the needs of the society for the information resources database of TCM. In addition, most people's application of the database only stays in the query and statistics, and does not fully reflect the true use value of the database, and we should develop a wider application of the database.

(c) The infrastructure is not well established. The development of the TCM information industry mainly serves the people's health and socio-economic development. Then, the establishment and improvement of the TCM information network platform have become a task we must complete to develop, acquire, and manage the information resources of TCM. The quality of the infrastructure is directly related to the development of the TCM information industry. We must increase investment in TCM research, TCM management and other fields, and create a practical and standardized TCM information support system and network environment.

(d) The awareness of innovation needs to be further improved. Competition in the

international pharmaceutical market is becoming more and more fierce, and depends on technological innovation. The success or failure of competition lies in the awareness of technological innovation of TCM enterprises. Many TCM companies follow the old-fashioned methods of operation and management, and gradually lose their competitiveness in the market. From the current competitive characteristics of TCM, innovation should become the reason for the development of TCM. Therefore, the TCM information industry should establish a sense of innovation in the process of continuous development. The technical advantages of the information industry should be used to actively participate in the competition of TCM research and market. In addition, the TCM information industry should make full use of its own advantages and strive to enhance the core competitiveness of it under the background of internationalization.

5.1.3 Utilization of TCM information resource

(1) The use of information resource of TCM

The utilization of information resources of TCM plays an important role in promoting the informatization of TCM. In recent years, China has made many achievements in the construction of information resources for TCM, and has made great progress in the utilization of information resources of TCM with the rapid development of information technology. It has initially satisfied the needs of the medical sector and society. For example, the Digital Museum of TCM opens a window for the world to understand the culture of TCM, allowing TCM to transcend time and space and spread to the world. The key problem is to reduce the loss caused by the waste of information resources of TCM, improve the level of utilization of information resources of TCM, give full play to the role of information resources of TCM, realize the effective circulation of information resources of TCM and the sharing of information of TCM.

(2) Analysis of factors affecting the retrieval of information resources of TCM

The most crucial thing to realize the effective circulation and sharing of TCM information resources is to involve the retrieval of TCM information resources. The process of TCM information retrieval is the process of identifying, searching and using TCM information. According to the demands for TCM information, we adopt scientific and effective methods, apply specialized retrieval tools, and then obtain required information from a huge amount of information ocean in a quick and accurate way. The result of each individual search may vary, and is affected by a number of human factors.

(a) The retrieval ability of the searcher. Whether the manual retrieval system or the computer retrieval system is used, the retrieval is performed by the searcher. Thus, the retrieval ability of the searcher has a great influence on the retrieval efficacy. The searcher must have the scientific and cultural knowledge of TCM information, can correctly conduct subject analysis, can skillfully use the retrieval tools of TCM literature, and apply flexible methods of retrieval. Because the entire process of retrieval is controlled by the searchers, we must pay attention to the cultivation of the quality of them.

(b) The performance of the retrieval language. When searching for TCM information resources, we must use the search language. The vocabulary and grammar of the search language have a direct impact on the search results. For example, the reference categories in

the classification language, and the semantic reference system in the language are beneficial to improve the recall rate of TCM information. If we can use a better retrieval language, we can obtain the better retrieval efficacy.

(c) The quality of the retrieval sigh in the literature. The retrieval sigh is the basis for our literature retrieval. The retrieval sigh has a great influence on the recall rate and precision. If the retrieval sigh is not stored in the retrieval system or is inaccurate, the related literature information is difficult to retrieve. Therefore, we must pay attention to the use of the retrieval sigh. If the retrieval sigh is used properly, we can obtain the valuable literature information.

5.2 The security and countermeasures of TCM information resources

5.2.1 Problems in information security of TCM

(1) The information security management of TCM is not enough, and there is a lack of an effective management plan.

Network information management must be designed based on information security and thus better serve information users. The same is true for the security management of TCM information. The rules and regulations governing the management of information security of TCM cannot meet the requirements of current demand. For example, regarding the issues of patient privacy, the illegal "unification" of pharmaceutical companies, the disclosure of information on TCM, there are no clear and targeted security management requirements and regulations. Some TCM units have formulated some rules. However, these rules are only used within the units, lack unified standards, and cannot be used as common guidelines. This seriously affects the effective role of TCM informatization in China.

(2) TCM has poor information security protection capability and does not have a complete security protection system.

After the development of TCM information system in recent years, it has gradually become the main reason for the informatization of the TCM industry, promoting the scientific management of TCM enterprises and improving the quality of TCM services. TCM information systems are facing a series of security issues, such as server security, operating system security, computer work environment security, storage device security, system management security, data exchange security and cash settlement security between medical institutions and banks or social security organizations, etc. These problems not only affect the social image of medical units, but also cause certain economic losses. The TCM information system is not guaranteed in terms of security. Thus, it has become a short-board in the process of informatization construction of TCM.

(3) The construction of TCM database is not perfect, and there is the lose of TCM information data.

With the continuous deepening of informatization of TCM, more and more database systems are established. To improve the level of service, data interaction platforms have gradually emerged. However, the information security issue of TCM has appeared at the same time. For example, the state of continuous construction is not good during the TCM data

exchange, many databases have not maintained, security protection measures are not good, or the database itself has security backdoors, *etc.* Some people even tamper with this information or spread it, causing great harm to individuals and medical units.

(4) The effectiveness of TCM information security supervision is not obvious, and the security supervision is not strong enough.

The entire process of circulation and use of TCM information lacks an effective security supervision system. The national electronic information supervision system for TCM and the related information system of TCM enterprises cannot be connected. Unreasonable and illegal use of information on TCM usually happens. The construction of software and hardware supporting environment for providing information security of TCM is insufficient. The understanding of the supervision of information security of TCM is even indifferent. The lack of integration of information resources of TCM and not unified information standards for TCM increase the difficulty of security supervision of TCM information, and affect the effective operation of the TCM industry.

5.2.2 Countermeasures to solve the problem of information security of TCM

(1) Continuously improving the information management system of TCM

The use of TCM information is inseparable from the management system. The security of TCM information system mainly means that the software, hardware and information data of the TCM information system can be protected. The information system can operate normally and resist external damage. In the process of using the information system of TCM, it is necessary to strengthen management to ensure that it can obtain the correct operation. We often check the equipment and surrounding environment involved in the information system and the virus, and strengthen the repair of operating system vulnerabilities. We have learned to use the application properly, and regularly test the security of the TCM information system to ensure that the information system can run in an orderly and safe manner.

(2) Increase the use of security protection technology

The use of security protection technology is important to ensure the security of TCM information. It runs through the beginning and end of security protection. Now the most used security technologies are access control technology, security audit and monitoring technology, computer virus prevention technology, firewall technology, intrusion monitoring and monitoring technology, system backup and fault recovery technologies, *etc*. They have the reliability, integrity and controllability characteristics, which can improve the intensity of information security. In addition, we must constantly improve the existing security technology, accelerate the research on new security technologies, and promote the upgrading of security technologies.

(3) Improve the information security awareness of medical staff

The people who contact most with TCM information are medical staff. They are directly related to the security of TCM information, vigorously carry out information security education and improve relevant laws and regulations. As human preventive measures, they cannot be ignored. Therefore, it is necessary to strengthen their information security training and prevention skills training for TCM. We improve their security awareness, and make them fully

aware of the importance of strengthening the information security protection of TCM. We need to accelerate the construction of information security defense system for TCM. In the training process, we strengthen their training in operation, the use of TCM information, computer technology update learning, and methods to prevent information leakage, *etc*. This gradually establishes a high-level and high-quality security management team in information security of TCM.

(4) Strengthening information security measures for TCM

The TCM Information Database System adopts a computer system that manages many existing TCM information in the form of the database. It can realize the sharing of data and the efficient use of data. The security of the database is to prevent the theft, modification or destruction of data caused by illegal users on the basis of ensuring the availability and integrity of the database information. To strengthen the security protection mechanism of the database, we can take security measures such as data encryption, identity authentication or password authentication at login, definition of user access rights, and audit trail of data to ensure the security of the TCM information database.

(5) Establishing a security management organization of TCM information

The establishment of a special security management organization can improve the efficacy of information security management of TCM. This organization must manage the information security industry of TCM in accordance with the national laws and policies. In this way, we can implement the national policies, formulate relevant development strategies, coordinate security work, analyze the status of TCM information security, and monitor the implementation of TCM information security measures. TCM information security is a part of national strategic security. It is related to the process of national informatization. It can promote communication between national government departments and help determine the direction suitable for the development of TCM information security.

(6) Accelerating the development of information security standards for TCM

The information security standard is a technical specification and basis for ensuring the information security of TCM, and is an important part of the information security system of TCM. The establishment of this standard can play an important technical support role in solving the information security problem of TCM. At present, the Chinese government is exerting great influence on the information security industry from various aspects such as technology, standards and management. In terms of the construction of information security standards, we must combine the current national conditions, through research and development, to develop a security standard suitable for the development of TCM information, and provide legal protection for information security of TCM.

5.3 TCM information quantization and regulation

5.3.1 Quantization processing of TCM information

In recent years, domestic scholars have focused on the research of TCM information database system. The TCM information is divided into two types. One type is increasing information, such as new natural product components, pharmacological activity data, *etc.* The

other type is information that is basically stable and needs to be standardized, such as the efficacy of TCM, processing methods, *etc*. Most of the research on TCM information database is aimed at short-formed data with exclusive and relatively independent connotation, such as efficacy, categorization, usage and dosage, toxicity, chemical composition and other information on TCM establishment. Most of this type of data can be directly quantified by computer. The other type is a complex large-text TCM data or picture composed of multiple Chinese characters, such as pharmacological effects, clinical diagnosis, various discourses, and pictures of TCM. This type of data has an important relationship with the deep information of TCM. However, most of the content is not suitable for computerized quantitative processing without special processing. The traditional method cannot effectively solve the extraction of TCM information that describes vagueness. Improving the accuracy of quantitative extraction of TCM data requires an accurate description of the fuzzy concept. This study introduces the concept of "quantization" of TCM information in the design of TCM database. This concept refers to the refinement of the original complex TCM data through reasonable analysis, which is composed of several Chinese characters or numbers. The smallest unit of information with relatively independent connotations cannot be subdivided and exclusive. This minimum unit of information is called the "quantum" of TCM information. The process of quantification of TCM information is called "Quantization". After the quantification of TCM information, a large number of ambiguous, noisy and redundant data in TCM database are cleaned and filtered to ensure the certainty, consistency and uniqueness. This is beneficial to meeting the demands of data mining. The separation and removal of non-medicinal parts of TCM are quantified, which are refined into the removal of root, stem, branch, stalk, peel, shell, hair, heart, core, reed, flesh, head, skin, tail, bone, feet, wing, meat, impurities, mildew. The information about the Chinese medicine picture is quantized by the first letter plus the number of the pinyin. For example, the picture of the yam can be coded as SY001. By "quantizing" the original data of TCM, the TCM data with complex original content can be subdivided into much "quantum" information with independent concepts. The quantum information of TCM can be established based on the "quantum" information of these TCM. The database lays the foundation for deep mining of Chinese medicine data. The process of TCM quantization is shown in Figure 5-1.

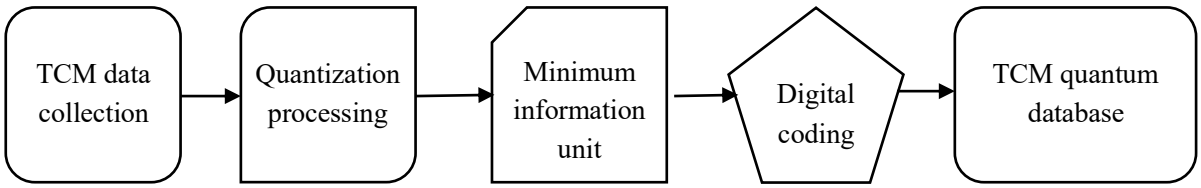


Figure 5-1. The process of TCM quantization.

5.3.2 Statistical analysis of quantum information data of TCM

This database system uses the 2015 edition of the "Chinese Pharmacopoeia" and the authoritative TCM as the data sources. For example, through the study of 617 kinds of TCM s contained in the "Chinese Pharmacopoeia", it was found that the quality standards of 24 processed products of TCM are listed separately. 23 TCMs are attached to the catalogue of original medicines. The processing methods in the 2015 edition of the "Chinese Pharmacopoeia"

are classified into purifying processing, cutting and preparation, quick frying and roasting, *etc*. The purifying processing means that the Chinese herbal medicines can be separated into medicinal parts by removing the non-medicinal parts, removing the impurities of the sediment, and performing the size division. The cutting and preparation refer to softening the purified selected medicinal materials, cutting into different specifications of silk, segments, blocks, tablets, *etc*. Thus, the effective ingredients of the medicinal materials after cutting are more convenient to fry, and the quality of the omelet is improved and easy to smash. The method of quick frying and roasting in the general rule includes frying, simmering, charcoaling, steaming, boiling, stewing, braising, and calcining. Other methods include sputum, frosting, powder-refining with water, germination, and fermentation. The database deals with the "quantization" of processing method (Figure 5-1). Some of these Chinese herbal medicines and decoction pieces contain one type of processing specifications, and some contain several processing specifications. More details are shown in Table 5-1.

Table 5-1. The specifications statistics of processing methods of TCM.

Project	Frequency	Percentage
0	180	29.2
1	227	44.9
2	140	22.7
3	17	2.8
4	2	0.3
5	1	0.2
total	617	100.0

Table 5-2. The quantization of drug properties of TCM.

Project	Quantization
drug properties	cold, hot, warm, cool, flat, very cold, very hot, slightly cold, slightly hot

As shown in Table 5-1, 180 Chinese herbal medicines have no decoction pieces in the database, and most Chinese herbal medicines contain two kinds of processing specifications, and a few of them contain more than two processing specifications.

5.3.3 Quantization of TCM information

Quantization of the raw data not only improves the independent and exclusive characteristics of TCM information, but also lays a good foundation for the mining of TCM data. According to the flow of Figure 5-1, the TCM is quantized. The results of the quantization process of this database are shown in Table 5-3.

Table 5-3. Quantization process of TCM flavor.

Project	Quantization processing
flavor	sour, bitter, sweet, spicy, salty, light, glutinous, slightly spicy, slightly bitter, slight glutinous, slight sweet, slight sour

Table 5-4. Quantization process of toxicity of TCM.

Project	Quantization processing
toxicity	extremely toxic, slightly toxic, non-toxic, toxic

Table 5-5. Quantization process of species of TCM.

Project	Quantization processing
TCM species	Mongolian medicinal drugs, Tibetan medicinal drugs, Uighur medicinal drugs, Yi medicinal drugs, Traditional Chinese medicinal materials, TCM decoction pieces, processed products, *etc*

Table 5-6. Quantization process of TCM dosage.

Project	Quantization processing
Dosage	9-15 g, 3-6 g, 1-3 g, 0.15-0.35 g, 3-9 g, 6-12 g, 15-30 g, 6-9 g, 5-10 g, 3-10 g, 2-5 g, 15-60 g, 1-1.5 g, 1.2-2.5 g, 6-15 g, 3-15 g, 9-12 g, 1-2 g, 1.5-3 g,1.5-2.5 g, 5-12 g, 9-30 g, 2-6 g, 0.3-0.6 g, 3-6 g, 0.06-0.6 g, 6-12 g, 10-15 g, 0.3-0.9 g, 9-27 g, 0.9-1.2 g, 0.15-0.35 g, 6-10 g, 0.1-0.3 g, 15-30 g, 2-10 g, 2-11 g, 0.5-1.5 g, 0.15-0.3 g, 6-12 g, 9-18 g, 5-10 g, 9-16 g, 6-10 g, 30-60 g, 12-30 g, 9-20 g, 1.5-3 g, 0.1-0.5 g, 0.1-0.2 g, 1-2 g, 3-12 g, 0.6-l.5 g, 0.6-0.9 g, 4.5-9 g, 0.3-1 g, 0.8-1.6 g, 9-24 g, l-5 g, 3-5 g, 3-4.5 g, 4.5-9 g, 15-45 g, 15-25 g, 0.6-1.5 g, 0.5-1 g, 2-3 g, 10-25 g, 1-1.2 g, 0.5-2 g, 5-15 g, 2-3 g, 1-6 g, 12-40 g, 20-30 g, 10-30 g, 120 g, 0.03-0.06 g, 0.05-0.1 g, 10-20 g, 15-21 g, 5-7 g, 0.2-0.6 g, 25-50 g, 0.9-2.4 g, 30-45 g, 0.015-0.03 g, 0.03-0.1 g

Table 5-7. Quantization process of TCM usage.

Project	Quantization processing
Usage	Another fried, grind powder and swallow, pill powder, external use, tea, first fried, long fried, wrap, ingredient added later, simmer, and dip with 40% ethanol

Table 5-8. Quantization process of TCM precautions.

Project	Quantization processing
Precautions	Pregnant women use with caution, pregnant women are prohibited, pregnant women and hemoptysis, vomiting patients should not take, pregnant women and those who are allergic to lacquer, edema should be used with caution, pregnant women and menstrual period, breastfeeding use with caution, raw food should be cautious, not suitable for long-term service and used by athletes; toxic ingredients can be absorbed through the skin, external application should not be applied in large areas, pregnant women, infants and kidneys are banned, children and the elderly should be used with caution, heart disease, tachycardia, glaucoma patients and pregnant women are prohibited, bleeding tendency and pregnant women use with caution, not suitable for more to avoid poisoning; glaucoma patients are prohibited; pregnant women and severe hypertrophy of the prostate use with caution and should not be taken excessively; pregnant women, patients with weak stomach, deficiency of cold, liver and kidney dysfunction use with caution, avoid drinking strong tea when taking medicine, patients with yin deficiency and blood dryness use with caution, dilated pupils use with caution, should not be taken excessively, pregnant women, patients with exogenous and hot cough for glaucoma, hypertension, and tachycardia, when smashing, pay attention to protection, so as to avoid allergic reactions, avoiding use of real evils, easy to addiction, not suitable for routine use; pregnant women and children are prohibited; athletes use with caution

Table 5-9. Quantization process of TCM medicinal parts.

Project	Quantization processing
Medicinal part	shell, carapace and plastron, nest, mature fruit, germinated and dried processed products, fermented processed products of mature seeds, mature seeds, germinated and dried processed products, dried insects of females, fallopian tubes of female frogs, dried products of bile, secreted wax, dried sterile fruits, dried grass stems, dried mature spores, dried mature fruit, dried ripe core, dried mature fruit shell, dried mature fruit peel, dried ripe fruit, dried ripe fruit, vascular bundle of dried mature fruit, dried mature fruit order, dried mature seed, dried mature seed, dried belt stems, dried flower heads, dried scaly fleshy stems, dried leafy stems, dried cane stems, dried gallstones, dried aerial parts, dried secretions, dried skin, dried root bark, dried skin or bark, dried roots, dried roots and rhizomes, dried roots and rhizomes or stems, dried roots and rhizomes, dried rhizomes, dried rhizomes and roots, dried rhizome and petiole residue, dried rhizome nodes, dried root bark, dried root bark or dried skin, dried root bark or near bark, dried peel, dried fruit, dried ears, dried flowers, dried pollen, dried corolla, dried flowers and flower buds, dried flower buds, dried flower buds and inflorescences, dried flower buds or flowers with initial opening, dried receptacles, dried inflorescences, dried inflorescences or flower buds, dried spines, dried pseudobulbs, dried near mature fruits, dried near mature or mature fruits, dried near mature or mature ears, dried near mature seeds, dried stems, dried stem pulp, dried stems and leaves, dried stems, dried stems and leaves, dried sclerotia, dried roots, dried tubers, dried bulbs, dried knobs section or branching branch, dried egg sheath, dried mother root, dried inner shell, dried shoots, dried whole grass, dried whole, dried whole plant, dried fleshy stem, dried fleshy scales, dried sandbag inner wall, dried bark, dried bark and root bark, dried resin, dried sputum, dried scorpion or fruit stalk, dried cane, dried body, dried flower head, dried outer fruit , dried immature fruit, dried fine leaves, dried leaflets, dried heartwood, dried stamens, dried leaves, dried petioles, dried leaves and leafy shoots, dried leaves or leafy shoots, dried fronds, dried young fruits, dried fruit or immature fruit peel, dried algae, dried branches and leaves, dried branches or dried skin, dried shoots and leaves, dried shoots, dried intermediate layer, dried seed kernels, dried stigma, dried fruiting bodies, a dried mass of the secretions in the stalk, a wood containing resin, a dried complex of parasites and larvae of parasitic larvae of the bat moth, aril, healthy placenta, horn, and dried roots and rhizomes, scales, bulbs, antlers to colloidal horns, stuffed honey, peeled branches, dried extracts, whole grasses, husks that fall off when nymphs emerge, mature males, dried secretions in the sachet, the dried heartwood of the trunk and roots, the fragrant resin exuded from the trunk, the dried resin after processing, the dried resin of the resin, the pearl formed by the bivalve, and the underarm, dried epidermis, immature or near-ripe dried outer skin, fresh rhizomes, fresh or dried aerial parts, fresh or dried roots, fresh or dried stems, fresh or dried roots, fresh whole grass or dried aerial parts, fresh leaves, fresh leaves Processed crystals, fresh branches, leaves are extracted and processed, the stag is not ossified, the horns of the stalks, or the horned horns, the horns that fall off in the spring after the saw, the dried snakes, and the dried ground. Part or root, dried roots and rhizomes, juice concentrates, lethal dried bodies, processed roots

Table 5-10. Quantization process of TCM storage method.

Project	Quantization processing
Storage method	set in a dry place; placed in a cool and dry place; avoid light, placed in a dry and ventilated place; dried product in a dry place, fresh buried in wet sand; dried product placed in a dry and ventilated place, fresh placed in a cool and humid place; buried in sand; sealed, kept below 30 °C; sealed, dry place; sealed, placed in a cool place; sealed, avoid light, placed in a cool dry place; sealed, moisture-proof, avoid light, placed in a cool place; sealed; placed in a dry place; cool place; sealed, placed in a cool place; ventilated and dry place; cool and dry place; sealed with wooden box, commonly used pepper mixed, placed in a cool dry place, anti-mite; shading, sealed, dry place; set in the dry place, moisture-proof; placed in the wooden box lining with oil paper, moisture-proof, anti-mite; set in the ventilation and dry place; set in the ventilation in a cool and dry place; set in a cool and humid place; set in a cool place; set in a cool dry place

Table 5-11. Quantization process of TCM genus.

Project	Quantization processing
Genus	Benzolidaceae, white mushroom fungus, phylum, genus Liliaceae, cypress, deciduous family, medusa, pearl family, genus Fernaceae, primulaceae, botanical, gecko, scorpion, scorpion Aphididae, silkworm moth, scorpionidae, scorpionidae, psyllidae, tamarisk, scutellaria, scorpionidae, Labiatae, Euphorbiaceae, rushaceae, hollyaceae, leguminous, azalea, Eucommia, Eucalyptus, Polyporaceae, Polygonaceae, Hydrangea, Duckweed, Olive, Queraceae, Turtle, Kelp, Halal, Haijinsha, Hailong, Aphididae, Gramineae, Black-triangular, Taxaceae, Vespa, Piper, Walnut, Caragana, Cucurbitaceae, Saxifragaceae, Grey-grayed fungi, Polygonaceae, Apocynaceae , Zingiberaceae, Amaranthus, Asteraceae, Amaranthaceae, Malvaceae, Asteraceae, Sedum, Campanulaceae, Compositae, Polygonaceae, Selaginaceae, Jujube, Chicory Moss, bitter wood, scorpion, orchid, scorpion, scorpion, scorpion, scorpion, genus, genus Aphididae, gentian, deer, sylvestris, radix, ephedra, verbena, dentate, aristolochia, equine, mazonic, sargasso, ergotaceae Fungi, geranium, ranunculaceae, bee, oyster, magnolia, kapok, phylum, osmanthus, bovidae, grape family, horse chestnut, lacquer tree, medlar, yarrow, rose Branch, Solanaceae, Caprifoliaceae, Myristica, Rutaceae, Sanbaaceae, Umbelliferae, Mulberry, Moraceae, Cyperaceae, Hawthorn, Shanglu, Provincial Oysteraceae, Cruciferae , pomegranate, stone pine, Amaryllidaceae, Dianthus, Gentianaceae, Persimmonaceae, Rhamnaceae, Dioscoreae, Water Dragon Orthopaedics, Minkidae, Nymphaeaceae, Pinaceae, Cynomoraceae, Sandalwood, Polygonaceae, Myrtaceae, Garcinia, Araceae, Frog, Squid, Sapindus, Sycamore, Polygonaceae, Araliaceae, Polygonaceae, Camphoraceae, Scrophulariaceae, Convolvulaceae, Eucalyptus Branch, Flax, Radix, Cobra, Ginkgo, Papaveraceae, Serpentidae, Iris, Polygalaceae, Rutaceae, Ze , Lauraceae, pearl Beca, mussels Branch, Division Zhi Ma, Phasianidae, porcine, comfrey, Zijin bovine, Osmundaceae, Bignoniaceae, palms, a total of 166 quantum. Benzolidaceae, white mushroom fungus, phylum, genus Liliaceae, cypress, deciduous family, medusa, pearl family, genus Fernaceae, primulaceae, botanical, gecko, scorpion, scorpion Aphididae, silkworm moth, scorpionidae, scorpionidae, psyllidae, tamarisk, scutellaria, scorpionidae, Labiatae, Euphorbiaceae, rushaceae, hollyaceae, leguminous, azalea, Eucommia, Eucalyptus, Polyporaceae, Polygonaceae, Hydrangea, Duckweed, Olive, Queraceae, Turtle, Kelp, Halal, Haijinsha, Hailong, Aphididae, Gramineae, Black-triangular, Taxaceae, Vespa, Piper, Walnut, Caragana, Cucurbitaceae, Saxifragaceae, Grey-grayed fungi, Polygonaceae, Apocynaceae , Zingiberaceae, Amaranthus, Asteraceae, Amaranthaceae, Malvaceae, Asteraceae, Sedum, Campanulaceae, Compositae, Polygonaceae, Selaginaceae, Jujube, Chicory Moss, bitter wood, scorpion, orchid, scorpion, scorpion, scorpion, scorpion, genus, genus Aphididae, gentian, deer, sylvestris, radix, ephedra, verbena, dentate, aristolochia, equine, mazonic, sargasso, ergotaceae Fungi, geranium, ranunculaceae, bee, oyster, magnolia, kapok, phylum, osmanthus, bovidae, grape family, horse chestnut, lacquer tree, medlar, yarrow, rose Branch, Solanaceae, Caprifoliaceae, Myristica, Rutaceae, Sanbaaceae, Umbelliferae, Mulberry, Moraceae, Cyperaceae, Hawthorn, Shanglu, Provincial Oysteraceae, Cruciferae , pomegranate, stone pine, Amaryllidaceae, Dianthus, Gentianaceae, Persimmonaceae, Rhamnaceae, Dioscoreae, Water Dragon Orthopaedics, Minkidae, Nymphaeaceae, Pinaceae, Cynomoraceae, Sandalwood, Polygonaceae, Myrtaceae, Garcinia, Araceae, Frog, Squid, Sapindus, Sycamore, Polygonaceae, Araliaceae, Polygonaceae, Camphoraceae, Scrophulariaceae, Convolvulaceae, Eucalyptus Branch, Flax, Radix, Cobra, Ginkgo, Papaveraceae, Serpentidae, Iris, Polygalaceae, Rutaceae, Ze , Lauraceae, pearl Beca, mussels Branch, Pedaliaceae, Phasianidae, porcine, comfrey, Zijin bovine, Osmundaceae, Bignoniaceae, palms

Table 5-12. Quantization process of the TCM harvest.

Project	Quantization processing
Harvest	harvested when the fruit is ripe in winter; harvested when the fruit matures and becomes red from October to November; captured from November to March; harvested from November to March; harvested before the flower is unearthed in December or before freezing Digging; harvested from the beginning of April to November; harvested from April to May; stripping from April to June; harvested when the fruit turns red from April to June; collecting fruits from May to June; collecting young fruit from June; collecting from May to July; harvested when flowering is open from June to August; harvested from June to September; mining from late June to early August; harvested from late June to the first ten days of September; harvested in July and August, the trunk is cut, the resin is discharged, and it is harvested from October to April; the skin is harvested when the skin is still green in July; when the fruit is not cracked from August to November, the fruit branches are cut; when the fruit matures in red-yellow in November, it is harvested; in September-November, the flowers are harvested in batches; the mature fruits are picked; the early spring flowers are picked when they are not open; the first seedlings are just germinated or the leaves are withered at the end of autumn; excavation in early summer when the plants are withered; stripped in spring and autumn; harvested in spring and autumn; roots in spring and autumn; excavated plants in spring and autumn; spring and autumn production; catching in spring and summer; germination in spring and summer, or excavation when stems and leaves are wilted in autumn, cultivated products are harvested in the third year of September after the planting or in the middle of April of the fourth year; in autumn, the stems and leaves will be excavated when they are withered; they will be harvested at the beginning of spring or when the stems and leaves are withered at the end of autumn; the roots will be excavated at the beginning of spring or autumn; the springs will be harvested; when the spring flowers are just opened, the flowers will be picked; harvested when spring flowers are in full bloom; harvested when spring flowers are not open; harvested after emergence in spring or early summer; harvested in spring when germinated or autumn plants withered; harvested after the leaves are withered, the excavation is carried out; when the spring seedlings are unearthed or before the autumn frozen soil, the excavation is carried out; when the spring is not pumped, the excavation is carried out; when the heights of spring seedlings are 6 to 10 cm, the harvested or the autumn flower buds are grown until the flower is opened; mining at the end of the flowering season; harvested at the end of spring and early summer; harvested in the late spring and early summer will be batched in batches; harvested in late spring or early summer or early winter; full bloom from early spring to early summer to early fruit; late spring to autumn Captured at the beginning; harvested from late spring to early autumn; harvested when the flower blooms from spring to autumn; when the flower bud turns from green to red Winter harvested; winter harvested; harvested after winter fruit ripening; harvested when winter fruit is ripe; harvested when winter fruit is ripe; mining when winter stems and leaves begin to wither; winter leaves and leaves wilting; winter stems and leaves withered Mining in the lower part; the lower part of the winter leaves are yellow and the upper part is brittle; the winter is to the next spring; the immature fruit is harvested from winter to spring; the winter to the next spring when the fruit is ripe; the winter to the next spring Excavation when the leaves are withered or not stalked; harvested in the winter to the next spring; harvested when the flowers are not open at the end of winter and spring; more than 7 to 9 months; more than spring picking; autumn stripping; autumn harvesting; digging; catching in summer and autumn; capturing in summer and autumn; collecting in summer; harvested in summer and autumn; capturing from spring to autumn; capturing in summer; harvested when ripe; harvested after ripening; collecting after ripening; harvested from green to black; harvested when flower is opened; excavated after winter to the next year; harvested in autumn and winter; harvested in autumn and winter; mature fruit harvested in autumn and winter; harvested mature seeds in the second season; harvested in autumn and winter; in autumn and winter, the excavation is carried out; in the autumn and winter, the ground part is withered; in the autumn and winter, the fruit is harvested when it is ripe or frozen; when the fruit is ripe in autumn and winter, the fruit is changed from green to red in autumn and winter, it will be harvested in yellow; picking at the beginning of flowering in autumn and winter; harvested when stems and leaves are wilting in autumn and winter; mining in autumn and winter when leaves are withered; harvested in autumn and winter, and harvested in early autumn; spores are not cut off when harvested; autumn harvested mature

fruit in autumn; autumn mining; autumn mining; autumn picking; harvested mature fruit in autumn; cutting stem in autumn; harvested when it is purple-black; harvested when the fruit turns red in autumn; harvested when the fruit is ripe in autumn, red or orange-red when the fruit is ripe; the whole plant is taken when the fruit is ripe and not cracked in autumn; the fruit is harvested after the fruit is ripe in autumn; the skin is collected after ripening; the fruit is harvested when the fruit is ripe; the fruit is harvested when the fruit is ripe in autumn; the fruit is harvested when the fruit is ripe in autumn; the fruit is harvested when the fruit is ripe in autumn; harvested when the first fruit is still green in autumn; harvested when the fruit has not turned yellow or yellow in autumn; harvested when the fruit turns from green to yellow in autumn; harvested when the autumn fruit turns from tender green to dark green; harvested when the autumn flower is first opened; harvested in autumn; excavation after flowering withered in autumn; excavation before flowering in autumn; harvested in autumn when flowering is in full bloom; harvested in autumn when flower is in full bloom; ripening fruit or cutting juice in autumn, harvested after the ripe fruit is removed; the autumn stems are withered or harvested in the second spring; the autumn leaves are withered when the leaves are withered; the autumn leaves are harvested when the leaves are flourishing; the autumn leaves are harvested when the leaves are green; harvested in the autumn to the next spring; autumn seeds are harvested at the end of autumn; harvested at the end of autumn; harvested at the end of autumn and winter; mature fruits harvested at the end of autumn and winter; fruits harvested when the skin becomes red at the end of autumn and winter; harvested after ripening at the end of autumn and winter; ripening at the end of autumn and winter; plants are harvested when they have not been cracked; when the fruit is ripe at the end of autumn, the plants are harvested when the shells are not cracked; the stems and leaves are withered at the end of autumn or excavated before the second spring germination, the fine roots are removed, the outer skin is scraped off, the flaps or segments are cut, and the strings are worn in a string; the roots are harvested from the fall of the autumn to the time of the second spring; the fruit is harvested from dark to green in the late autumn to the next spring; harvested in the late autumn to the next spring; harvested throughout the year; stripped throughout the year; harvested throughout the year; harvested throughout the year; harvested all year round; collected from late autumn to the next spring; the trunk is damaged by natural damage or the trunk is cut in summer and autumn, and the resin is collected; the flowers are harvested when the flowers are opened in April and May; harvested in summer and autumn; stripped in summer and autumn; harvested in summer and autumn; captured in summer and autumn; harvested in summer and autumn; harvested in summer or autumn; harvested in summer and autumn; harvested mature fruits in summer and autumn; harvested in summer and autumn; picked in summer and autumn, harvested when red and yellow in summer and autumn; harvested when the fruit is ripe, the peel becomes yellow, and the interior is picked during dryness; the plants are harvested when the fruit is ripe in summer and autumn; harvested when the fruit is ripe in summer and autumn; the fruit is collected when the fruit is ripe in summer and autumn; when the fruit is red in autumn, the fruit is harvested; in summer and autumn, the fruit is near mature and harvested; harvested in summer and autumn when the fruit is green and yellow; harvested in summer and autumn, when the fruit is green and yellow; in summer and autumn, harvested in the flowering and fruiting season; in summer and autumn, the flower is cut to the top and the ear is green; in summer and autumn, the flower is harvested; harvested at the opening; in the summer and autumn, the flowers can be harvested before and during the flowering period; in the summer and autumn, the flowers are harvested in the summer and autumn; in the summer and autumn, the flowers are harvested; in the autumn or after the snow melts, it will be excavated; in the summer and autumn, the stems and leaves will flourish or bloom until the three rounds; in the summer and autumn, the stems and leaves will be harvested when the leaves are flourishing; in the summer and autumn, harvested when the stems and leaves are flourishing; in the summer and autumn, leaves are excavated; in the summer and autumn, the velvet antlers are sawed; in the summer and autumn, the excavation is carried out before flowering; in the summer and autumn, the harvest is collected; in the summer and autumn, the leaves are harvested in the lush spring; the leaves are harvested in the autumn; in the summer and autumn, the leaves are harvested when the leaves are flourishing; the branches are harvested when the leaves are flourishing in the summer and autumn; the ears are harvested when the seeds are ripe in the summer and autumn; and the fruiting bodies are mature in the summer and autumn; in summer and autumn, fruits are harvested when they are ripe; in summer and autumn, leaves are flourishing; ripe fruits are

	harvested in summer and autumn; harvested in summer from green to red; harvested early in summer; excavation when stems and leaves are wilted in early summer; excavation after wilting in early summer; excavation in Xiachuzi, excavation when spores are not diverged; summer stripping; mature fruits are harvested in summer; picking mature fruits in summer; harvesting plants in summer when the fruit is ripe, when the peel is not cracked; harvested when the fruit is ripe in summer; harvested plants when the fruit is ripe in summer; harvested when the fruit is near mature in summer; harvested in summer when the ears are brown-red; harvested in summer flowering; harvested in summer flower opening or flower bud formation; harvested when summer flowers are not open; picking when summer flowers are not open; picking when summer flowers turn from yellow to red; picking when most of the stems and leaves are withered in summer; flowers are harvested when sunny, and summer flowers are selected; harvested in the flowering season; harvested before the flowering in the summer; mining in the summer when the leaves are dry; harvested in the summer when the leaves are not open; harvested in the summer when the leaves are flourishing; harvested in the summer when the leaves are flourishing; plants are harvested when the fruit matures in late summer and early autumn; cutting stems from late summer to autumn; harvested in early summer, stems and leaves are flourishing, flowers are not open; fresh products can be harvested all year round; dry goods in summer, the leaves are flourishing, and the ear is harvested in many times; harvested when the leaves are flourishing

Table 5-13. Quantization processing of TCM processing methods.

Project	Quantization processing
Processing methods	stir frying, bran frying, stir-frying with rice, stir-frying with earth, stir-frying with sand, stir-frying with clam powder, stir-frying with talcum powder, stir-frying with charcoal, wine-processing, vinegar-processing, salt-processing, ginger-processing, honey-processing, oil-processing, forging charcoal, forging, quenching, steaming, boiling, stewing, simmering

Table 5-14. Quantization processing of TCM preparing methods.

Project	Quantization processing
Preparing methods	After the excavation, the sediment and the waste rock are removed; after the excavation, the miscellaneous stones are removed; after the excavation, the miscellaneous stones and sediment are removed; after the excavation, the impurities are removed; after the excavation, the heat is melted to remove the impurities; after the excavation, wash, dry, and remove the stones; pick the fruit, expose under the sun, lay the seeds, sieve the peel, branches, and dry; cut fresh into thick slices or small pieces; dry; freshly sliced, dried or dried at low temperature; mature fresh fruit and skin nitrate are processed; except for the ground part and sediment, dry; remove white sapwood, dry; remove sapwood, dry; remove residual roots and impurities, wash, slightly steamed or dried after drying; remove residual stems, fibrous roots and sediment, dry; remove residual stems, fibrous roots and sediment, and dry; remove residual leaves, bundle into fresh or cut sections, dry; remove thick stems, cut into sections, drying, or steaming and drying; removing thick stems and sediment, cutting the section to dry; removing the rough skin, drying, or simmering freshly sliced, dried; removing the rough skin, washing, drying; removing the rough skin and fibrous roots, drying; or steamed (boiled) and dried; remove the aboveground part and sediment, separate into large size boiling water pot to cook until the heart, dry; remove the ground stem, wash, dry; remove roots, leaves, dried; remove roots and impurities, drying; removing roots, fibrous roots and sediment, drying; remove roots and roots, drying; remove roots and small roots, washing, drying, or slicing fresh slices, drying; removing roots and fibrous roots, washed, dried; remove the roots and fibrous roots, and bake with a slight fire semi-dry, stack “sweating” until the interior turns green, then dry; remove the stems, leaves, set the boiling water slightly hot or slightly steamed, remove, dry; remove the stems and impurities, steam to the upper gas or set the boiling water slightly hot, remove, dry; remove the peel, remove the seed, wash, remove the hard shell (exocarp), dry; remove the peel, dry or low temperature drying; remove the peel and fleshy aril, wash, dry; remove Pulp, washed, dried; remove the pulp and core shell,

remove the seed, dry; remove the pulp and core shell, remove the seed, dry; remove the pulp and core shell, collect the seeds, dry; remove the fruit, dry; remove pedicel and sediment, dry; remove inflorescence, cut into sections, dry; remove stems, leaves and fibrous roots, wash and dry; remove stems and fibrous roots, wash and dry; remove shoot tips; cut, dry; remove stems and leaves, dry; remove stems; remove stems and impurities, dry; remove old stems, dry; remove old branches, dry; remove reed, roots and sediment, bake or dry to semi-dry, pile back to run, then dry or dry; remove the internal organs, wipe off, use bamboo to open, make the whole flat and straight, dry at low temperature; remove the sediment, air dry; remove the sediment, dry; remove sand, dry, or remove the hair (scale); remove the sediment, dry; remove hard roots, petiole and golden yellow fluff, cut thick slices, dry, steam and then sun until six or seven percent dry; cut thick slices , drying; removing sediment, drying; removing sediment, drying or drying, then removing the roots; removing the sediment, drying; removing the sediment, drying, removing the roots; removing the sediment, drying, and then removing the roots And the outer skin; remove the sediment, dry, hit the fibrous roots; remove the sediment, dry; remove the sediment, dry or dry; remove the sediment, dry after drying, then remove the roots; remove the sediment, wash , dried; remove the sediment, place in boiling water or boiling brine, cook until the body is stiff, remove, set the ventilation Dry; remove the sediment; sunburn, stack "sweating" until the surface is reddish yellow or grayish yellow, spread out to dry, or directly dry without sweating; remove the black skin and dry when fresh; remove the sediment and coarse skin, peel off the root bark, dry; remove the sediment and fine roots, cut into longitudinal or diagonal pieces, and dry; remove the sediment and fine roots, steam or boil until thoroughly, dry; remove the sediment and the roots, cut sections, large longitudinal sectioning into petals, dry and then hit the rough skin; remove the sediment and fibrous roots, dried or sliced to dry; remove sediment and impurities, and dry; remove silt, dry; remove fleshy peel, dry, remove core shell and wooden diaphragm; remove fleshy aril, wash and dry; remove fleshy Skin, washed, slightly steamed or slightly boiled, dried; remove the outer skin, cut into large pieces, add water to cook, concentrate, dry; remove the outer skin, slightly dry, cut or cut longitudinally or cut obliquely into thick Tablets, dry; remove the outer skin, fibrous roots and sediment, dry or low temperature drying; remove the skin and pulp, wash, dry, remove the seeds; remove the skin and impurities, dry; remove fine roots, wash, freshly slicing, dry; remove fine roots and sediment, peeling roots, drying or scraping; remove the rough skin, remove the wooden heart, dry; remove the fibrous brown hair, dry; remove the roots, dry; remove the roots, sun to six or seven percent dry, gently flatten, dry; remove the roots, wash and fresh scrap off the rough skin, wash and dry; remove the roots, wash, simmer fresh cut short or thick slices, dry; remove the roots, wash, dry; remove the roots, wash, dry, or freshly sliced, dry; remove the roots, wash and dry; simmer freshly into thin slices, dry; remove the roots, wash, slice, and dry; remove the roots, wash, slice, and dry; remove the roots, wash, dry; remove the roots wash, dry or simmer freshly sliced, dry; remove the roots, wash, dry or dry at low temperature; remove the roots, wash, set the boiling water slightly hot, dry; remove the roots, wash, set the boiling water slightly hot or steam to the heart, dry; remove the roots, wash, set a little hot in boiling water steam to no white heart, remove, dry; remove the roots, wash, boil in boiling water or steam until no white heart, sun to half, remove the skin, dry; remove the roots, wash, boil in boiling water when there is no white heart, take it out and dry it; remove the roots, wash it, put it in boiling water until it is white, remove it, scrape off the skin, rinse, dry; remove the roots, rough skin and sand, dry or dry at low temperature; remove the roots, sediment and roots of the gel, dry; remove the roots and residual scales, wash, cut into sections, and dry; remove the roots and roots, and dry them; remove the roots and sediments, and after the water has evaporated slightly, bundle them into small pieces, shed them, slowly dry them with pyrotechnics; remove the roots and sediment, dry; remove the roots and sediment, dry , smash the residual roots; remove the roots and sediment, dry; remove the roots and sediment, bake until semi-dry, stack for 2 to 3 days, soften and then dry until completely dry; remove the roots and sediment, bundle into small After drying to dry wrinkles, cut the top and dry it; remove the roots and sediment, cut into pieces or slices, dry or dry; remove the roots and sediment, dry; remove the roots and sediment, and dry or remove the skin, dry; remove the roots and sediment, dry or low-temperature drying; remove the roots and sediment, dry or low-temperature drying; dry slices or dry at low temperature; remove the roots and sediment,

	sunbathing remove rough skin, dry; remove the roots and sediment, ventilate and dry until the outer skin is dry; remove the roots and sediment; remove the roots and skin, dry; remove the roots and sediment, dry; remove the roots and sediment, dry; remove roots and skin, dry; remove buds, fibrous roots and membranous leaves, use or dry; remove the leaves, dry, or slice to dry; remove the knotted roots and sediment, braided into a dried shape, or directly dried; remove impurities, dry; remove impurities, dry; remove impurities, drying or steaming and drying; removing impurities, drying in time; remove impurities, cutting off some fibrous roots, twisting into a spiral or spring shape while heating, drying; or cutting into sections, drying or low-temperature drying; remove impurities, drying; remove impurities, dry, or freshly sliced, sun-dried; remove impurities, dry; remove impurities, dry or low-temperature drying; remove impurities, dry; remove impurities, dry; remove impurities, when the leaves are soft, stack them until the leaves turn purple brown, dry; remove impurities, wash, dry; remove impurities, dry; remove the roots; remove the fruit, remove the peel, dry; remove the seeds, dry; soak, wash, steam; remove the skin in time, cut into pieces or cut into thick slices; take the ginger, use the sand to heat it to the bulge, and the surface is brown

After the quantization processing, we can summarize the drying method of TCM according to the above information. The drying of TCM can preserve specific effects of TCM and facilitate storage. In the database of TCM quantum information, the statistics of the drying methods used in the processing of 532 TCMs are shown in Table 5-15.

There are some TCMs that do not have a clear processing and drying method. As shown in Table 5-15, most of the TCMs are botanicals. The medicinal parts are mostly the roots and rhizomes of plants, and the drying methods of them are drying in the sun or drying. For TCMs that their medicinal parts are flowers or flower buds, drying methods of them are drying in the shade or drying at low temperature.

5.3.4 Quantization process of medical information

Medical information “quantization” processing refers to the decomposition of the original complex medical information resources into the smallest information units that can be subdivided without reasonable subdivision through reasonable analysis. This can help to realize the digitization of medical information. For example, the hospital grade can be divided into ten quanta of “three grades and ten grades”, which are 1-Level A-Class, 1-Level B-Class, 1-Level C-Class, 2-Level A-Class, 2-Level B-Class, 2-Level C-Class, 3-Level A-Class, 3-Level B-Class, 3-Level C-Class, 3-Level Special-Class. Through the “quantization” subdivision, the construction of the database can be facilitated, the data in the database can be optimized, users can obtain more detailed medical information and improve the effect of medical information query.

(1) Design of medical quantum information table

The design of the medical quantum information table has a series of processes. First, we must understand the purpose before designing the table. We determine the required table in the database system according to this purpose. Then, we consider the field in the table that is used as the primary key. Then, we establish the relationship between the tables and implement referential integrity. Then, we check the table to find deficiencies, optimize the table, and finally input data.

Table 5-15. drying methods used in the processing of 532 TCMs.

Drying method	Amount	TCM
Sun drying	286	no record
drying	142	no record
drying in the shade	31	Jiulixiang, Xiheliu, Zhuru, Benzoin, Red Cardamom, Agarwood, Awei, Artemisia, Maple Balm, Arborvitae, Melon, Niuhuang, Mutong, Poria, Melon, Pine, Cinnamon, Tianshan Snow Lotus, Manshan Red, Asarum, Honglian, Xu Changqing, Senecio, Mallow Fruit, Citron, Wing Grass, Xinyi, Quanju, Coltsfoot, Fragrant, Valerian
Sun drying or drying beside the fire	5	Achyranthes, Aphis, Atractylodes, Scorpion, Sakamoto
Sun drying or drying beside the fire at low temperature	1	Eucommia leaves
Sun drying or drying at low temperature	25	Dried ginger, Fritillaria, leeches, Fritillaria, white peony, Cordyceps sinensis, Dilong, Dihuang, American ginseng, malt, Wusong, bergamot, grain bud, tangerine peel, grass fruit, clam shell, medlar, magnolia, Amomum villosum, citron, anterior humilis, golden flower, scorpion, puzzle, scorpion
Sun drying or drying in the shade	15	Wood thief, epimedium, orange, lotus, mint, safflower, ramie, Wushan Epimedium, thief, invertebrate, cloth slag, Shiwei, Gansong, celandine, Shanglu
drying in the shade or drying beside the fire	2	Velvet
Sun drying or drying with a desiccator	1	Musk
drying in the shade or drying at low temperature	1	Rose flowers
Sun drying or direct Sun drying	2	Radix Pseudostellariae
drying or direct drying	3	Star anise, rhubarb, northern sand ginseng
Sun drying or drying beside the fire for half a day, retaining 3-6 days	1	Scrophularia
repeated drying	1	Ganoderma
drying in the shade or drying beside the fire at 40-50 ℃	1	Hypericum
drying in the shade or drying beside the fire at low temperature	1	Dendrobium
drying at low temperature	5	Plum, dragonfly, rose, gastrodia, mangosteen
drying beside the fire at low temperature	1	Dark plum
air dry	1	Antler
drying beside the fire	3	Chuanxiong, ginkgo, continuous
air dry	1	Hawthorn
air dry or drying beside the fire	1	Ginseng leaf
drying in the shade or drying beside the fire	1	Chrysanthemum
drying in the shade or drying	1	Aspongopus

(2) The relationship between medical quantum information tables.

Based on the design process and practical principles of medical quantum information table,

and the characteristics of medical quantum information, we design three tables of hospital quantum information table, doctor quantum information table, disease quantum information table. This is shown in Figure 5-2.

(3) Content design of medical quantum information table

The content design of the medical quantum information table is based on the classification of medical quantum information, and the effective storage of information is realized through the reasonable classification of the abstract data stored in the information table. The system processes and classifies the collected quantum information to realize efficient storage of quantum information, which is convenient for users to query. The collected data is verified to ensure the accuracy of the medical quantum information. For example, by categorizing, the hospital quantum information table includes the hospital name, hospital level, hospital address, hospital introduction, and bus route. The details of this table are shown in Figure 5-3.

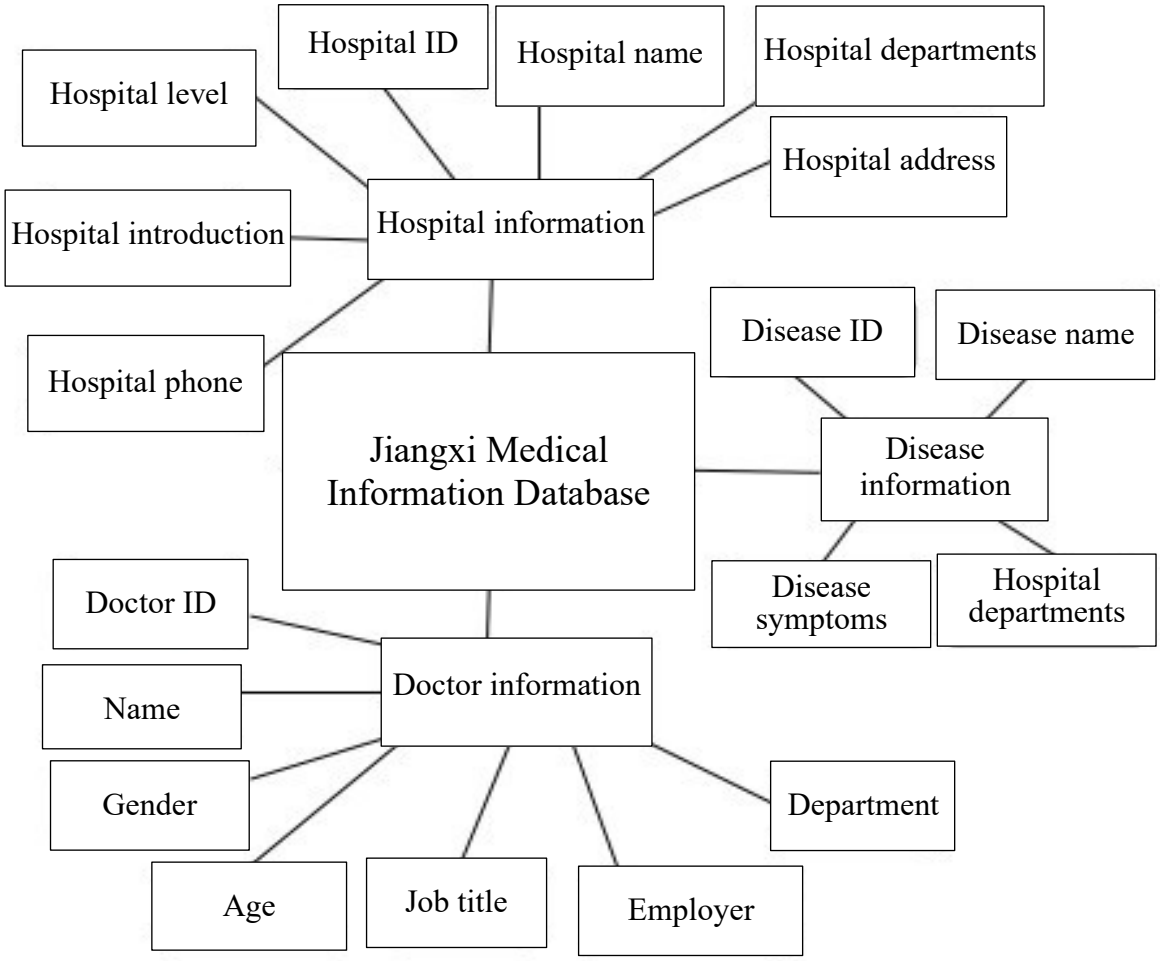


Figure 5-2. Three tables of medical quantum information.

(4) Query design of medical quantum information table

Database is the basic project to support knowledge discovery. Query is an important application of knowledge discovery in the database. The query can query a table according to the conditions set by the user, or can associate multiple tables to perform an overall query. Its data viewing and data statistics functions are powerful. Thus, users do not need to view the data

of all the tables. We enter the conditions of the information that users want to view. Users can filter out the information that meets the conditions, and help users to make the best decision.

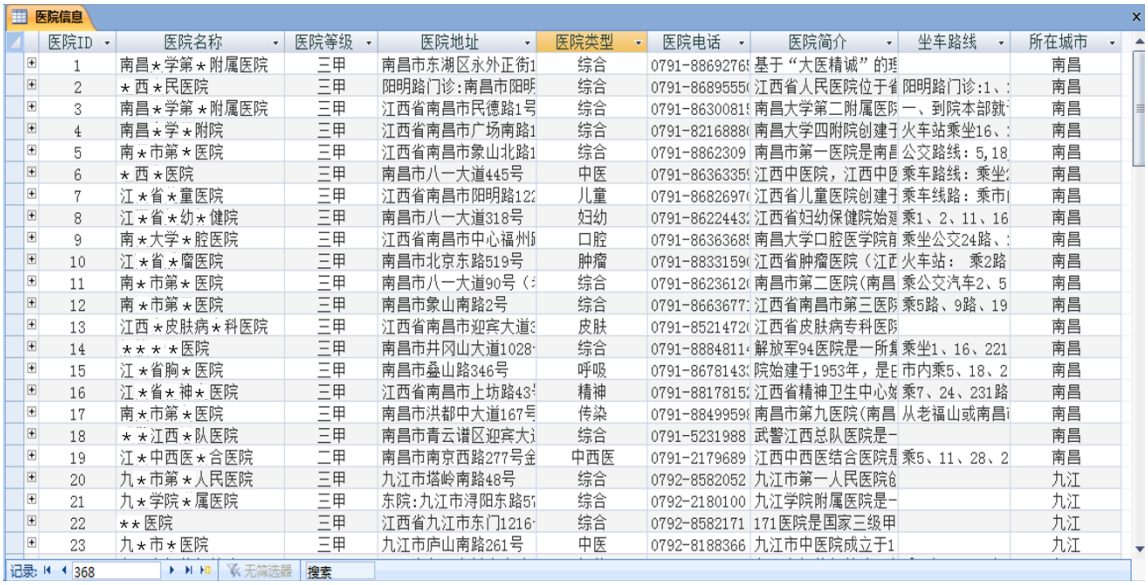


Figure 5-3. Hospital quantum information table.

(a) Query design of quantum information single-table. The query of quantum information single-table is mainly for the operation of single information table. In the single-table query process, the direct query can be performed without establishing the relationship between the tables, and the required quantum information is filled into the condition box. After running, the system will automatically display the results of the query, and the operation of this query is quite simple. For example, to find a well-known expert in orthopedics in Jiangxi Province, you can enter the word “orthopedics” in the condition box of the doctor information query in this system. After running, the experts are filtered out, and the results of the single-table query are shown in Figure 5-4.

Figure 5-4. Quantum information single-table query.

(b) Query design of quantum information multi-table. Quantum information multi-table

query is performed in the presence of multiple quantum information tables. Before the query, the administrator needs to establish the relationship between the tables through the primary key and implement referential integrity. On this basis, the user can perform multi-table query. Through this query, the required information in multiple tables can be extracted and displayed. The most important feature of this query is that it can obtain more detailed medical quantum information. This provides a more reliable basis for the user to make the final decision. For example, to find out the name of the well-known expert in pediatrics, the name of the hospital, the hospital level and the city where the hospital is located, we enter the relevant conditions in the condition box to query the required medical information. The results of the quantum information multi-table query are shown in Figure 5-5.

Figure 5-5. Quantum information multi-table query.

5.4 Data mining of TCM information and its application

5.4.1 Overview of data mining technology of TCM

Data mining is a process of discovering novel, effective, potentially useful, and ultimately understandable knowledge from large amounts of data. It is designed to help decision makers find potential relationships between data and identify factors that are difficult to discover manually. It is different from traditional information processing methods. Data mining methods are mostly based on knowledge of machine learning, pattern recognition, statistics, *etc*. Data mining can realize data analysis in a linear or non-linear way, and then discover potential usefulness knowledge. The application scope of data mining technology covers almost all the fields of TCM research, especially in the fields of prescription compatibility and quality identification. This has become a hot spot in current research. There are many methods of data mining. Commonly used data mining methods include association rules, regression analysis, decision tree method, cluster analysis, frequency analysis, artificial neural network, *etc.* Data mining technology deals with fuzzy and nonlinear characteristics of TCM data. At the time, it has a distinct advantage over traditional processing methods. Data mining technology can effectively realize the in-depth study of TCM by analyzing a large amount of data. For example,

the use of data mining technology can deeply explore the theories of many medical scientists, which can further promote the unified understanding of TCM theory and practice. The clustering analysis method of data mining technology is used to cluster four qi and five flavors. Some TCMs were classified and predicted. The association rules were used to identify the medicinal characteristics of TCM. Some of them were not classified and predicted. In this chapter, we explore the application of data mining methods in the field of TCM. The "China Knowledge Network" is used as a search database, and the "data mining method" and "traditional Chinese medicine" are used as search terms to search for relevant literatures published in the past 10 years. These literatures are processed to eliminate the irrelevant one. 103 literatures are selected as the research objects. Throughout the process, the randomness of the literature selection is guaranteed as much as possible to improve the credibility of the data. Finally, the statistical analysis is performed on these literatures. For example, Qianfeng He *et al.* used the method of cluster analysis to explore the compatibility of TCM prescriptions from simple information to complex information. Hongyan Yu *et al.* used association rules to study the medicinal properties and classics of TCM, and reported that the close frequent itemsets are warm-spicy, cold-bitter, flat-sweet, cold-sweet, *etc*. Erxin Shang *et al.* studied the incompatibility of TCM based on association rules, and reported that the frequency of occurrence of some characteristic combinations of hot-lung, hot-bitter and hot-stomach in the drug pair is low, and that in the incompatibility is high. The statistical results of data mining methods and TCM related properties are shown in Table 5-16. The numbers represent the number of literatures.

According to the Table 5-16, the relationship between the data mining method and its main aimed TCM properties is summarized, as shown in Table 5-17.

The application advantages of data mining technology are introduced above. However, there are deficiencies. Each data mining method may aim only one or several problems, and it is not applicable to all problems. It has certain limitations in the use process. Therefore, it is necessary to solve the same problem with different data mining methods. The results obtained by a data mining method are not completely sure of the correctness. The simple data mining method has its fixed result presentation form, and some forms are not easy to understand. Through the joint application of different data mining methods, we can discover a suitable and easy-to-understand form. Commonly, researchers use a data mining method to solve a problem. There is a lack of comparison among different data mining methods, and a lack of in-depth analysis of problems. In addition, it is not convenient to use a data mining method to deal with different problems to maximize the role of a certain method. This data mining mode is relatively simple, and mining efficacy is relatively low. In response to these problems, data mining method based on the strategy pattern is proposed.

Table 5-16. Statistics of literatures of data mining methods and TCM related properties.

	TCM	Channel tropism	efficacy	Quality evaluation	TCM toxicity	Compatibility	Pharmacological effect	Prescription
Association rule	26	14	15	1	3	12	3	2
Frequency analysis	6	9	12	3	2	13	4	13
Cluster analysis	4	7	8	0	3	4	11	9
Decision tree	7	7	8	0	1	3	4	3
Principal component analysis	3	2	7	9	1	0	6	2
Artificial neural networks	9	7	3	11	2	1	12	3
Rough set method	2	6	7	0	5	5	1	3

Table 5-17. The relationship between the data mining method and its main aimed TCM properties.

Data mining method	Association rule	Frequency analysis	Cluster analysis	Decision tree	Principal component analysis	Artificial neural networks	Rough set method
Mainly aimed field	drug property of TCM, channel tropism, efficacy, compatibility	prescription, compatibility, efficacy, channel tropism	pharmacological effect, channel tropism, efficacy, prescription	efficacy, drug property of TCM, channel tropism	efficacy, quality evaluation, pharmacological effects	pharmacological effect, quality evaluation, channel tropism, drug property of TCM	efficacy, channel tropism, TCM toxicity, compatibility

5.4.2 Data mining method of TCM

Association rule refers to the existence of some correlation and regularity between two or more variables in the same event. It was first proposed by Agrwaal *et al.* in 1993. It is mainly used to discover hidden relationships among TCM data. In general, only the ARM that satisfy the set confidence and support are meaningful. The confidence and support here refer to the lowest reliability and the lowest degree of association, respectively. The association rule can be used to discover rules that statistical methods and traditional artificial intelligence cannot discover. The association rule plays an important role in the research process of TCM. For example, the research model based on association rule has become the primary choice for analyzing the implicit relationships among TCM prescriptions. In recent years, with the deepening of the research on association rule, some efficient algorithms for association rule data mining are proposed. The data mining process of association rule is shown in Figure 5-6.

Artificial Neural Network (ANN) is a simplification, abstraction and simulation of biological neural networks, reflecting the basic characteristics of biological neural networks. It is a nonlinear analysis model that can process TCM data by means of computer simulation of the human brain. ANN has the advantages of fault tolerance, robustness and self-organization. Difficult problems and difficult decision-making issues can be better solved by artificial neural network. The artificial neural network is divided into three layers of input layer, hidden layer and output layer. The layers are connected with each other through nodes. The artificial neural network has a wide range of applications in the research process of TCM data mining. For example, using artificial neural network-ultraviolet spectrophotometry, a variety of chemical components in TCM can be determined without separation. With the deepening of artificial neural network research and the continuous expansion of applications, the artificial neural network system can replace many human intelligent works. The structure of the three-layer artificial neural network is shown in Figure 5-7.

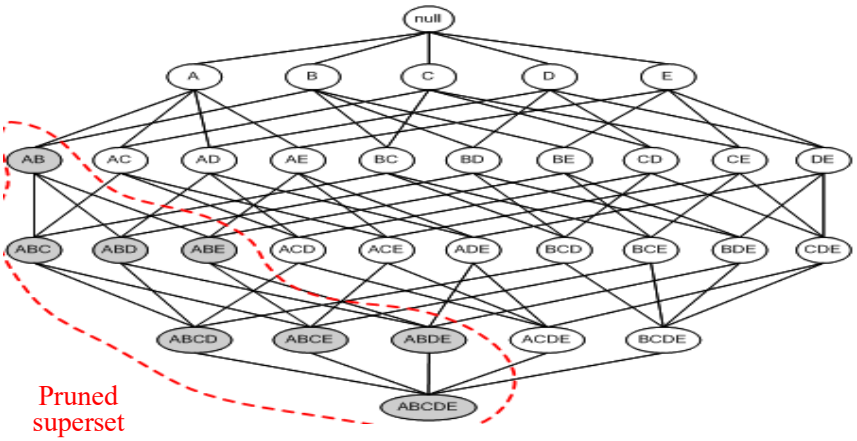


Figure 5-6. Data mining process of association rules.

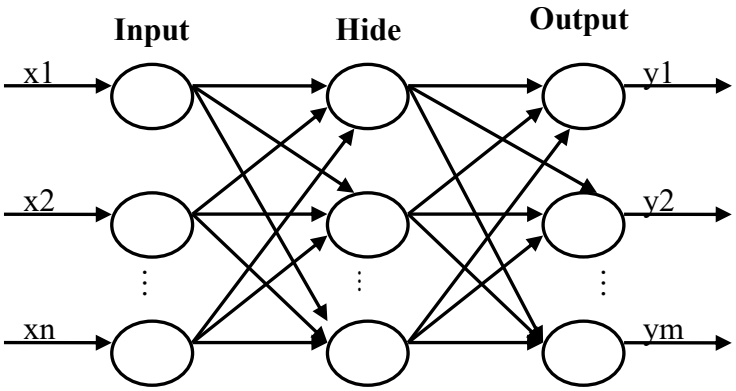


Figure 5-7. Three-layer artificial neural network.

The Decision Tree (DT) is a tree-like structure. Each internal node represents a test on a property. The branch represents the result of the test. Each node of the tree represents a class or class distribution. The decision tree is generated in two stages of tree construction and tree pruning. In the process of the decision tree, training the training set, taking tree pruning on the initial decision tree, and extracting the classification rules from the resulting decision tree are the key steps. In the classification analysis, the decision tree model is the most used one. Because it can display the results of the mining in a graphical way. The smaller the size of the tree is, the simpler the data is, the easier it is to be analyzed. However, not the smaller the scale of the tree, the higher the accuracy of prediction. Some researchers use the decision tree method to establish a classification tree for genes with known functional classification, and realize the classification and prediction of genes with unknown function classification. Decision tree data mining method is shown in Figure 5-8.

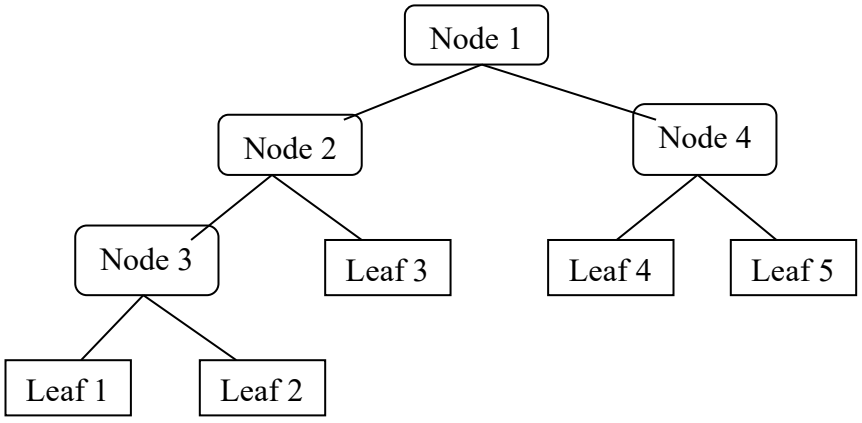


Figure 5-8. Decision tree data mining method.

Rough Set (RS) theory is a tool proposed by Professor Pawlak in 1982 to study inaccurate, incomplete information and knowledge. In term of artificial intelligence, rough set is a decision table. In term of the rough set, some sets are studied. From the aspect of programming, the rough set studies some special matrices. Rough set theory has strong complementarity with theories dealing with uncertainty problems such as probability theory, fuzzy mathematics and evidence theory. Rough set theory has strong advantages in solving large amount of data and

eliminating redundant information. It is widely used in data preprocessing and attribute reduction. The main idea of rough set theory is to characterize uncertain knowledge using the knowledge base. Rough set theory is a novel and effective soft science method. It has unique advantages in analyzing and dealing with incomplete information, and is a further development direction in the field of artificial intelligence. The rough set data mining theory is shown in Figure 5-9.

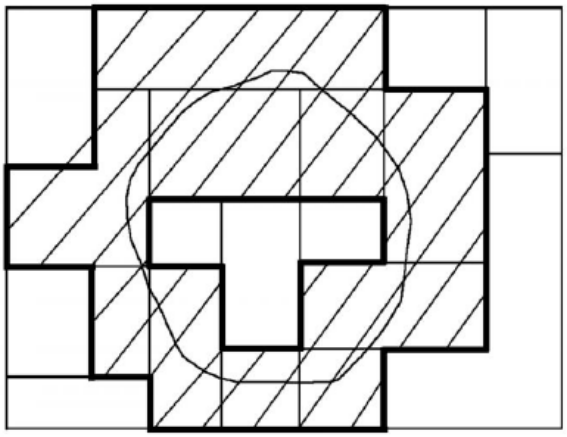


Figure 5-9. Rough set data mining theory.

Principal Component Analysis (PCA) is a method of statistical analysis, in which relatively few variables are selected from multiple variables by linearly transformation. It was first proposed by K. Pearson when studying non-random variables. Subsequently, K. Pearson extended it to the study of random variables. Researchers analyze all the variables, delete the closely related and repetitive variables by principal component analysis method, and establish a new variable. The new variables are independent of each other, and the information of the original variables is retained as much as possible. The steps of principal component analysis can be summarized as: (I) the standardization of the original data; (II) the determination of the number of principal components; (III) the interpretation of the meaning of the principal components; (IV) constructing a comprehensive evaluation function Y comprehensive with principal components and their variance contribution rates; (V) calculating the sample value of Y comprehensive and providing the order.

5.4.3 Application of data mining technology in TCM

(1) Application of data mining technology in new drug research and development

In the development of new drugs, the discovery of lead compounds is a key step of the research. The original mining methods are mainly random screening and accidental discovery. The efficiency is low. Data mining technology and knowledge discovery are gradually applied to the exploration of lead compounds with the development of computer technology and the deepening of new drug research. The drug development system created by data mining technology is used to discover an effective chemical substance foundation and establish the pharmacophore group. This can guide the research and development of new drugs and effectively shorten the development period of new drugs.

(2) Application of data mining technology in TCM literature

In the literature of TCM, there are a large number of qualitative descriptions and fuzzy concepts. Data mining technology can be used to correlate these descriptions to reveal their inherent laws. In addition, the application of data mining technology in the literature of TCM strengthens its links to literature processing, literature retrieval, and databases. For example, a lot of knowledge about the medicinal properties of TCM is described in the Chinese medicine literature. The correlation rule algorithm can be used to realize the correlation analysis of the medicinal properties, efficacy, and channel tropism of TCM. Some researchers have used the information extraction technique of Hidden Markov Model and the sample training method to extract potential knowledge from TCM cases.

(3) Application of data mining technology in TCM compound

TCM compound plays an important role in the treatment of diseases by TCM. During thousands of years of development, TCM has accumulated hundreds of thousands of TCM compounds and established several databases of TCM compounds. However, there are significant differences in the drug flavor and drug dose among these compounds, which makes it impossible to achieve a unified understanding of the disease law by TCM theory. However, through the data mining method to mine and sort out the valuable experience of many TCM experts, it will be able to fully realize the theory of TCM and a unified understanding of clinical practice. For example, using association rules and attribute-oriented induction methods can analyze the rules of drug compatibility in different compatibility relations, rough set theory can effectively deal with the simplification of compound and the extraction of compound features.

(4) Application of data mining technology in informatization of TCM

In the long-term development process, TCM has formed a unique professional expression terminology, which has maintained the characteristics of TCM. However, it has also become an obstacle to the informatization of TCM. Nowadays, it is urgent to apply modern science and technology to scientifically explain the knowledge of TCM, especially to realize the digitization and information processing of TCM. Text data mining of TCM knowledge is an important way to promote the digitization and informatization of TCM. In this process, the structural analysis of TCM knowledge based on plain text and ancient language is an important part of TCM information research. Some content can be implemented by data mining techniques for text. For example, the standardization of the TCM efficacy is a typical application of text data mining.

5.4.4 Design of data mining process of TCM

TCM data mining is a relatively complete process. This process is to mine unknown and valuable information from the TCM quantum information database. Before mining, users firstly should sort and classify TCM data mining methods or algorithms. Then, users design a pattern to implement the addition, deletion, and modification of data mining methods or algorithms to achieve the solution of “many-to-many” TCM problems or the mining of TCM data. Data mining generally goes through the following major steps. (1) Data collection. Users determine the purpose of data mining, recognize the problems to be solved by data mining, collect relevant internal and external TCM information according to the purpose of mining and problem solving; (2) Data sorting. The collected TCM data is classified and made into a form suitable for data

mining. The organized TCM data is transformed into an analysis model, which is based on the mining algorithm; (3) data mining. Users mine the collated TCM data, select appropriate mining algorithms before mining, and further develop and utilize TCM data based on mining algorithms; (4) mining results evaluation. The mining results are verified to ensure the accuracy of the mining results and displayed in a visual form to ensure the utilization of the mining results; (5) mining results application. The knowledge obtained through data mining technology is integrated into the user information system for user reference. The whole process of TCM data mining model is shown in Figure 5-10.

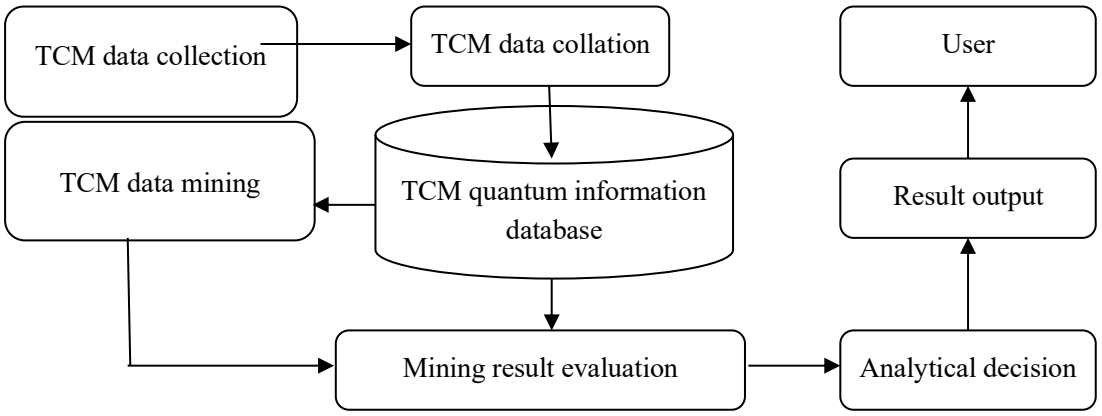


Figure 5-10. TCM data mining model.

5.5 TCM data mining based on the strategy pattern

5.5.1 Introduction of TCM data mining based on strategy pattern

At present, many data mining methods have been applied to the research of TCM. However, they are limited to specific problems. The specific method of data mining is one-to-one corresponding to a specific problem in the research of TCM. The relationship between the traditional data mining method and problem is shown in Figure 5-11. Although this data mining method has achieved some results, there is still a limitation of data mining for more valuable TCM.

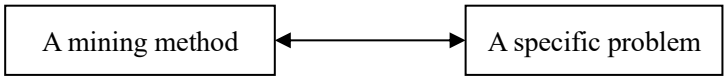


Figure 5-11. The relationship between the traditional data mining method and problem.

Thus, the application of the strategy pattern in TCM data mining methods is studied, the strategy pattern and method of TCM data mining are explored. Based on this new method, we design a TCM data mining technology that can flexibly realize the "many-to-many relationship" among different problems and different methods (Figure 5-12). For this processing method, both the TCM problem (data or information) and the data mining method (algorithm) are encapsulated, which can achieve following two results. First, different methods are used to solve a specific problem, which can avoid that a problem is being tied to a method, and can

compare the effects of different methods. Second, a specific method can be applied to different problems, which maximums the effect of a specific method for as many problems as possible.

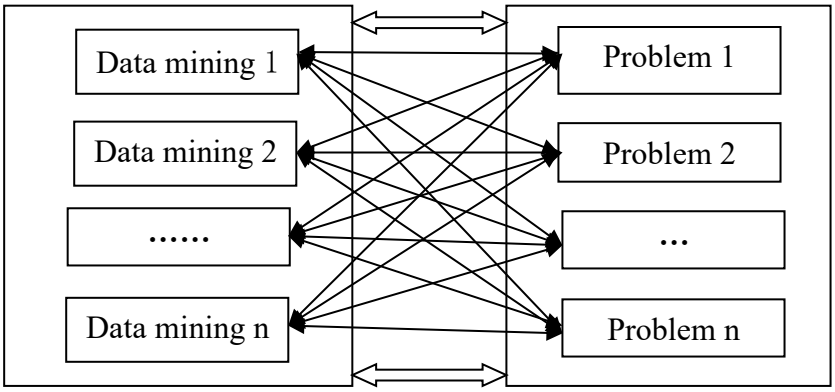


Figure 5-12. Many-to-many relationship of the data mining.

Strategy Pattern (SP) defines a series of algorithms, encapsulates them and makes them replaceable. Strategy pattern is composed of the abstract strategy role, the specific strategy role and the scene role. The abstract strategy role is responsible to the abstract class or interface, and provides the interface that the specific role requires to implement. The specific strategy role is to package the specific algorithm or behavior. The scene role refers to holding a reference to a strategy class, and finally implements a call for the client. The strategy pattern is one of the object behaviors patterns. The core of the strategy is the packaging of the algorithm. The responsibility of using the algorithm is separated from the algorithm itself in the packaging process, and they are delegated to different objects for management. A strategy pattern typically encapsulates a series of algorithms into a series of strategy classes as a subclass of an abstract strategy class. In other words, we define a set of algorithms and encapsulate each algorithm. Thus, they can replace each other. The flowchart of the strategy pattern is shown in Figure 5-13.

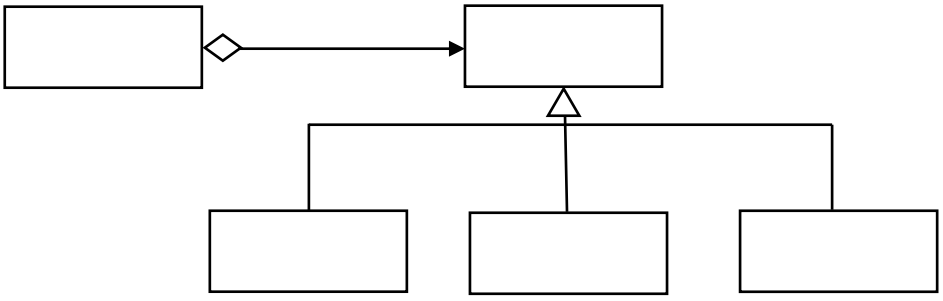


Figure 5-13. The flow chart of strategy pattern.

5.5.2 Application and advantages of TCM data mining based on the strategy pattern

(1) Discussion on the application of TCM data mining based on strategy pattern

The strategy pattern has unique advantages in the following cases. (I) An object has a large number of behaviors. If you do not choose the appropriate mode, these behaviors can only be

implemented with multiple addition selection statements. (II) Do not want the client to know the complex data structure related to the algorithm, and encapsulate the algorithm and data structure in the specific strategy class to improve the security and confidentiality of the algorithm. 3 When selecting an algorithm among several algorithms in a system. (IV) There are multiple classes in a system, and they only have behavioral differences. In this case, using a policy pattern allows an object to choose one of these behaviors.

(2) Advantages of TCM data mining based on strategy pattern

The strategy pattern can implement the update of the new algorithm or the replacement between algorithms without modifying the original system, and realize the multiple use of the code. This is a replacement inheritance in a new mode, thus effectively avoiding the shortcoming of difficult maintenance using multiple conditional transfers. In general, use behaviors or environment classes of algorithms may have subclasses, each of them provides different behaviors or algorithms. This leads to the confusion of behavior with the user of the algorithm. The logic of decision of behavior or algorithm is mixed with its own logic. The result is that it is impossible to develop independently. However, the introduction of the strategy pattern can solve such problems well. The way to dynamically change behaviors or algorithms can be well implemented. The hierarchical structure of the strategy class defines the algorithm or family of behaviors. The strategy pattern provides a way to manage related algorithms or behavior families. This can transfer public code to the parent class. Among them, it is possible to effectively avoid duplication of code. The strategy pattern embodies the principle of opening and closing, which means to close the development of the extension. In the case of the strategy pattern, when the new strategy pattern is added, it does not affect the modification of other strategy classes. This increases the scalability, and is developed for the extension. It only relies on abstract and does not depend on the specific implementation. Thus, the modification is closed. On the basis of the strategy pattern, the joint application of multiple data mining methods is discussed. Researchers use the strategy pattern to explore the data mining of TCM compound, and realize the goal of combining various data mining methods to explore the potential knowledge of TCM compound. The Bayesian network and support vector machine methods are used to understand the common law of TCM compound. The rough set theory is used to extract the characteristics of TCM compound. The association rule and decision tree methods are used to analyze the compatibility mode of TCM compound. Cluster analysis and association rules can achieve automatic classification of drug efficacy. It can study the association mode between drug efficacy and drug property. Rough set theory can simplify the medicinal characteristics based on drug classification. Association rules can be used to discover the rules that cannot be discovered by statistical methods and traditional artificial intelligence. Association rules play an important role in the research process of TCM. For example, the research model based on association rules has become a hidden method in the analysis of TCM prescription. On the basis of the primary choice of the relationship, the decision tree method is applied to the processing of this problem. The results of the mining can be displayed in a graphical manner quite conveniently. These show the advantage that the simple data mining does not have.

5.6 Design of TCM data mining system based on strategy pattern

5.6.1 Demand analysis of development of TCM data mining system based on strategy pattern

There are many data mining systems for TCM now. However, the TCM data mining system based on the strategy pattern is few. At present, with more attention from researchers, the development of TCM has reached an unprecedented speed. More and more people are choosing TCM to treat diseases and protect their health. Most hospitals have added TCM houses and increased the adjustment of TCM. Pharmaceutical companies have also expanded their research and development of TCM products and continuously developed new products. TCM plays an important role in safeguarding people's health. People urgently want to obtain TCM information conveniently and quickly through some ways in order to adopt corresponding strategies. To meet this demand from users, we have developed a TCM data mining system based on the strategy pattern. In this system, users can easily query TCM. It also provides a publicity function for TCM information, increasing the user's knowledge of TCM. The system implements the consultation function. It can communicate with users and answer questions encountered by users in a timely manner. In addition, the development of this system is of great significance for the discovery and deep mining of TCM information.

5.6.2 System development tool selection and network architecture design

(1) Tool selection of system development

We choose B/S (Browser/Server) form for the software structure of the system operation, which is a network structure model after the rise of WEB. The best advantage is that it does not require the user to install special software, only one browser is required. The browser automatically connects to the database through the web server. Considering the compatibility issue, the Oracle database is selected. It is one of the most popular databases in the B/S structure. It has complete database management functions, and it is reliable, shared, and massive in data. The persistence aspect of preservation has unique advantages, and also has certain data recovery functions, which can easily realize the operation of the data warehouse. Active Server Page Network (ASP.NET) is used in the development of server-side applications. It is a general language-based compile and run program. Thus, the functionality and applicability of it are powerful, and it can be run on WEB applications on all platforms of software development. The processing of the data interface, the basic library of the common language, and the message mechanism can be perfectly integrated into the ASP.NET WEB application. The ASP.NET development platform chose Microsoft Visual Studio 2010.NET because Microsoft Visual Studio 2010.NET integrates code writing, program debugging, compilation, operation, and other related operations on a platform, which greatly improves the efficiency of software design developers. ActiveX Data Object. Network, ADO.NET is used to access data in Microsoft technology. It is the interface for data access that is prioritized in the NET programming environment. Through the use of ADO.NET technology, you can effectively interact with the database. In the design of the front-end webpage, Dreamweaver (DW) is used. The great

advantage of Dreamweaver is that it combines webpage creation and website management. This software can easily create page that shows dynamic and static combination and cross-platform restriction.

(2) System network architecture design

The system is designed in a three-layer structure, which is the client layer, application management layer and data layer. The client layer is the page that the website displays to the user. The user can directly access the information displayed on the webpage through the browser. For the application management layer, the client cannot see it. Only the system administrator can access it. It can manage the client layer and control the content displayed by the client layer. The application management layer has its own management and setting page, which usually requires a password to log in. In the database layer, we can access the database. A large amount of data is stored in the database, and the information seen at the client layer is from the database. Through the database layer, the three-layer structure of the data in the database can be realized (Figure 5-14).

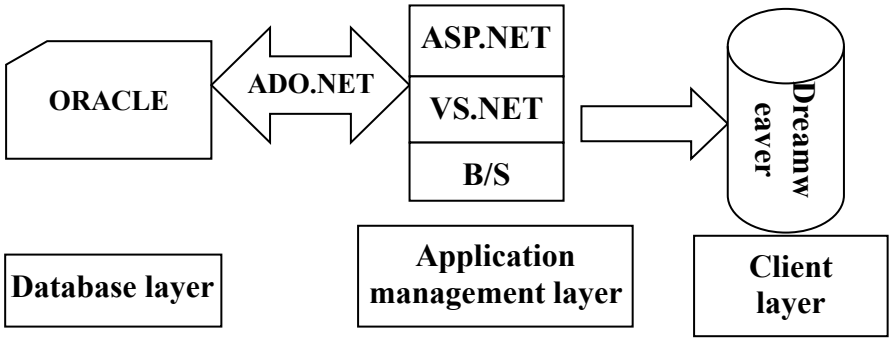


Figure 5-14. A schematic of the system of three-layer structure.

5.6.3 Functional model design of TCM data mining system based on strategy pattern

The strategy pattern-based data system will use knowledge of strategy pattern knowledge, statistical knowledge, database knowledge, artificial intelligence, computer technology, *etc.* It is not a simple combination of these knowledge and technology, and is a complete process of operation. It may need other technical support during the operation process. In this effective cooperation, we can complete our data collection, data processing, data analysis and output of results for the series of processes, and then present the results of the analysis to the user. Therefore, the TCM data mining system based on the strategy pattern is an application system that integrates data management, data analysis, data retrieval, expert evaluation, *etc.* Using the idea of the strategy pattern to study the TCM data mining model and method, the key is to realize a variety of TCM problems and multiple data mining methods to establish a "many-to-many" relationship. To achieve the maximum effect of a mining method or a problem, different solutions can be adopted. Finally, the in-depth extraction of TCM data information is achieved. The realization of these functions is based on the strategy pattern, encapsulating data mining methods and TCM problems, and implementing different methods or problems through the

same interface. The entire TCM data mining process based on the strategy pattern is shown in Figure 5-15.

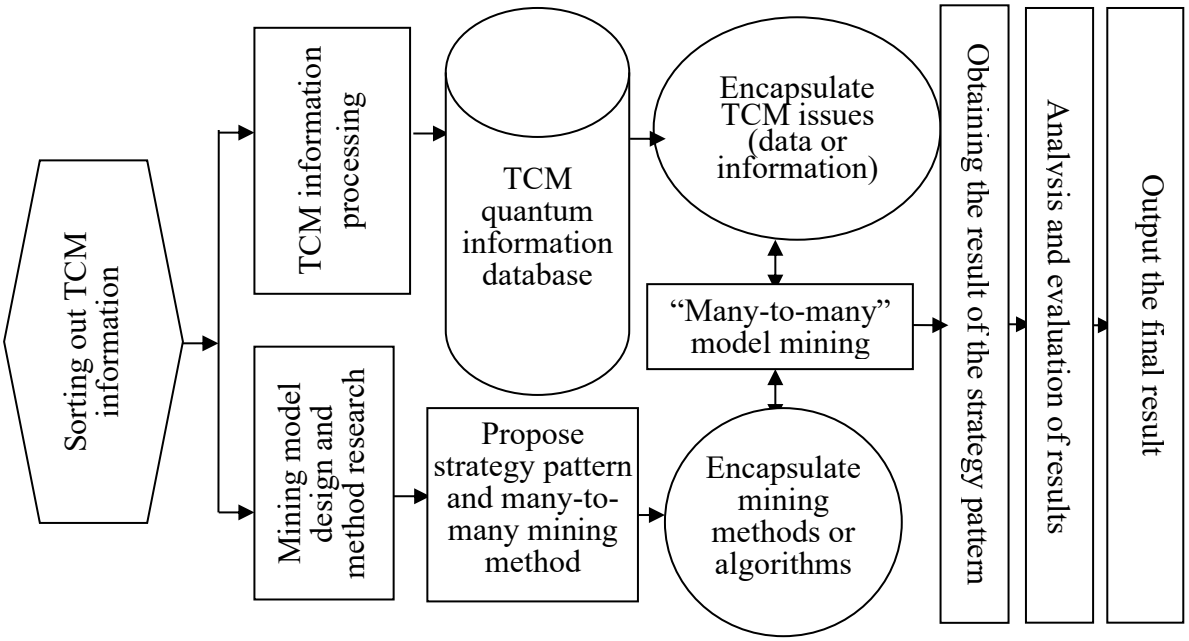


Figure 5-15. TCM data mining process based on the strategy pattern.

(1) The management module of TCM database

The TCM database is the basis for data mining of TCM data. Before the excavation, the TCM data was quantified and the TCM quantum information database was established. The TCM database management module is to maintain and manage the TCM quantum information database to ensure the smooth progress of data mining. In this module, there are two data processing libraries of TCM data warehouse and TCM mining knowledge base. They are responsible for cleaning and purifying external databases, guiding data mining process and evaluating data mining mode.

(2) The pre-processing module of TCM data mining

The collected TCM data is quite complicated, which does not have the characteristics of data mining. The module is responsible for sorting the data. The reorganization mentioned here includes cleaning up the noise, combining different data sources with each other, selecting the problem-related TCM data, and transforming selected data into a form suitable for data mining. If some data affects data mining, the evaluation mode discover the data, return it to the pre-processing of TCM data, and reprocess it.

(3) TCM data mining module

In this module, based on the strategy pattern, using various data mining algorithms, the algorithm and problems are encapsulated. Then, methods and rules in the data mining knowledge base are used TO achieve the mining and discovery of knowledge of TCM data. This module is the core of the TCM data mining system based on the strategy pattern, involving many data mining methods, techniques and the successful use of strategy pattern ideas.

(4) Evaluation module for TCM data mining

This module is mainly to analyze and evaluate the results of TCM data mining, because

there are many modes of data excavated by the TCM data mining module. The mode of interest to users may be only one or several of them, which requires that the user's concerns should be compared with these models. The value of the model is evaluated, and the shortcomings are analyzed. If the mined pattern is different from the user's concern, then the corresponding process needs to be returned to re-excavate the data. Finally, the mode that meets the user's concerns is transmitted to the TCM knowledge output module.

(5) TCM knowledge output module

This module realizes the interpretation of the patterns of TCM data mining. It provides the results of mining to decision makers in a way that is easy for people to understand. This module is a hub for users to communicate with data mining systems. The user can interact with the TCM data mining system through this interface. Entering the data mining keywords, the system will provide corresponding information. In this process, users can explore according to the results of each step in the TCM data mining process.

In summary, the TCM data mining platform based on the strategy pattern can flexibly combine data mining methods with TCM problems. This can compare the effects of different methods to deal with the same problem and the same method to process different problems. This displays the results to users. It can integrate existing data mining methods to make them more efficient. The TCM data mining method based on the strategy pattern is attracting more and more attention with its unique advantages.

5.6.4 Analysis of TCM data mining results based on strategy pattern

Throughout the study, we first analyze TCM for the treatment of "hyperlipidemia" and "obesity", and explore the core drugs for treating these diseases. Based on the strategy pattern, we use three data mining methods to solve this problem. The results are shown in Table 5-18. The TCM for treating the two diseases is basically the same. However, the core drugs are different. It is difficult to obtain such results only by using data mining technology, which indicates the superiority of TCM data mining based on the strategy pattern.

Table 5-18. Results of different data mining methods based on strategy pattern.

	TMC for the treatment of hyperlipidemia	**TMC for the treatment of obesity**
Association rule	Polygonum, Salvia, Alisma, Hawthorn	Poria cocos, Alisma, Atractylodes, Hawthorn
Text mining	Rhubarb, Hawthorn, Scutellaria	Alisma, Astragalus, Rhubarb, Atractylodes
Frequency analysis	Salvia, Astragalus, Hawthorn, Alisma	Hawthorn, Rhubarb, Alisma
Core drug	Hawthorn	Atractylodes

In addition, TCM data mining based on strategy pattern plays an important role in TCM prescriptions. Through the application of this model, one data mining method can be used to solve the problem of different prescriptions, and multiple data mining methods can be used to deal with the same problem. It can be targeted and comprehensive to investigate and discover the hidden rules in prescriptions. We reveal the compatibility relationship between the TCM in the prescription, the relationship between the drug flavor and the drug property, the relationship

between the prescriptions, and further deepen the understanding of the disease and the scientific laws of treating the disease. For example, 476 prescriptions related to Minor Radix Buplenri Decoction are collected from the ancient TCM books. Based on the strategy pattern, the association rules, cluster analysis and principal component analysis are used for data mining, and results show that the prescriptions are mainly related to more than ten drug pairs. The drug pair is guided by the basic theory of TCM. Finally, the core drugs found are Bupleurum, Pinellia, and Ginseng, and they are often used in combination with qi and heat-clearing drugs. In addition, about 60% of the prescription containing Bupleurum contains jaundice. In this chapter, we only list a few examples of the research process to illustrate the unique application of this research idea. In addition, the TCM data mining based on the strategy pattern has unique advantages in the research and development of new drugs, fingerprints, clinical application, pharmacological research, and pharmacological theory of TCM. It provides a new method for the development and utilization of TCM.

5.7 Implementation of TCM data mining system based on strategy pattern

5.7.1 System module composition and its functions

(1) System function module

This system is mainly to provide users with information based on the strategy model of TCM. In this system, the audit of the system administrator can ensure the reliability of the information, ensure that the information displayed on the website is true, reliable, and provide users with the powerful consulting service. The background function logic block diagram is shown in Figure 5-16.

(2) Functions of system function modules

(a) System login module. When the user wants to enter the system, they first login to the system administrator account. Then, users add the department, and the administrator is assigned to the department. The department administrators are divided into two types. One is for the hospital information entry clerk and the drug company information entry clerk. The other one is the auditor of system administrator, and is mainly responsible for reviewing the doctor information and drug information entered by the hospital information entry clerk to ensure the authenticity of the information and the credibility of the platform information. The system administrator can manage general administrators.

(b) Administrator management module. After the administrator enters the information about the hospital and the pharmaceutical company into the system, the administrator can assign one or more information input personnel to the system. Thus, it can manage and release the enterprise information, including the user name, password, contact information, address, profile, department, and administrator category of the information entry personal.

(c) News management module. Administrators at all levels can enter information here to release relevant information. The system administrator uses the added information of the article to determine which information entry the news belongs to. After the information is submitted by information entry personal, only the system administrator has the authority to modify, delete

and release it.

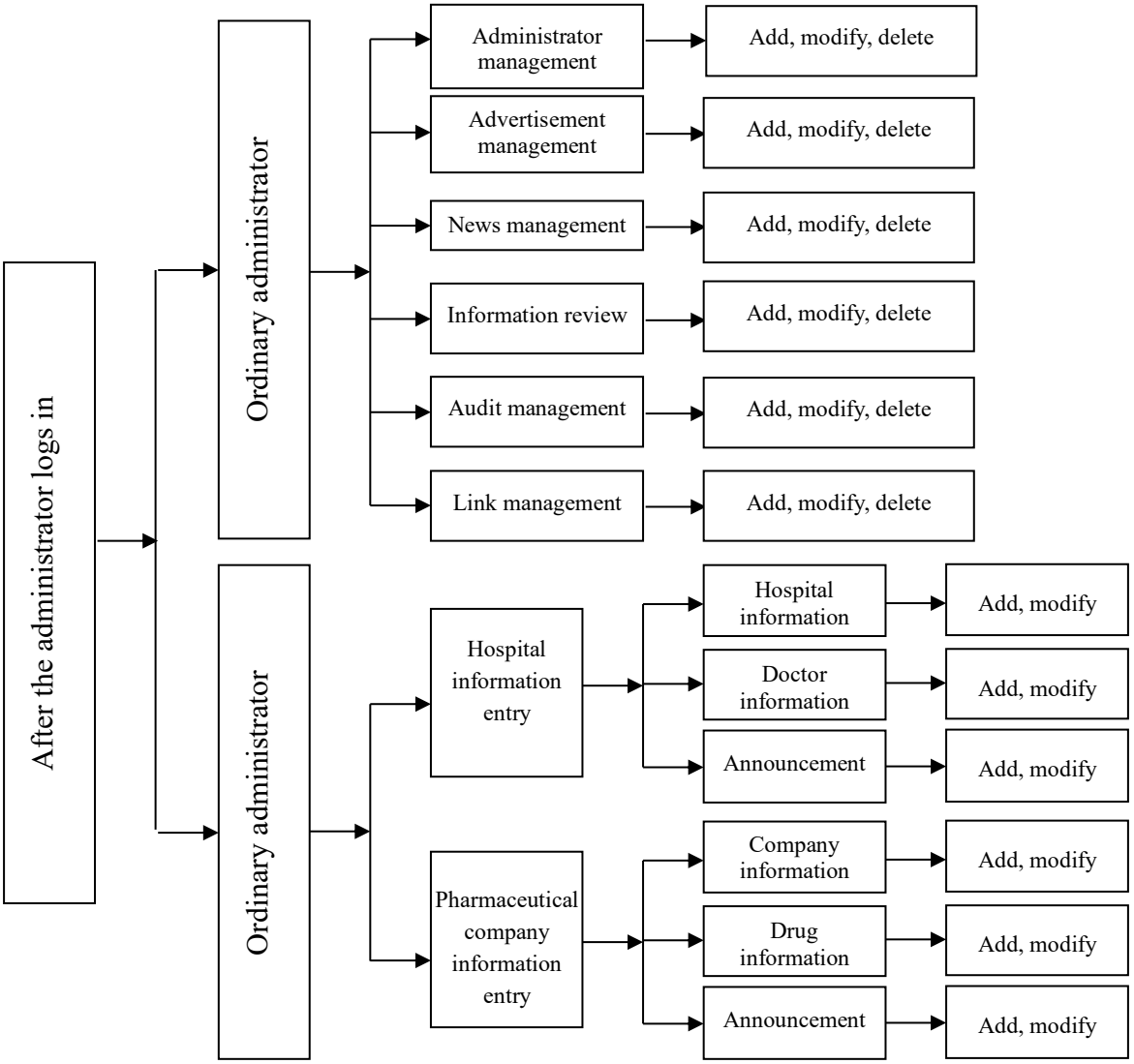


Figure 5-16. The background function logic block diagram.

(d) Doctor information management module. The hospital information staff enters the hospital's physician information into the system for users to query and access. The information entry personal can modify and delete the physician information, avoiding some erroneous operations and having unnecessary records in the database. After the information entry personal enter the doctor information into the system, the information passes the review and approval of the information review personnel directly under the Health Bureau. Then, the information can be displayed and inquired at the front desk of the system. If the information is not approved or the audit is not passed, the information cannot be displayed at the front desk. This ensures the credibility and authenticity of the information displayed in the system.

(e) Pharmaceutical information management module. The pharmaceutical company information entry personnel input the company's drug information into the system for users to search. The information entry personnel will undergo the review after entering the drug information into the system. Only after the approval is passed, the information can be queried at the front desk. Otherwise, it cannot be displayed at the front desk to ensure that users obtain

accurate drug information.

(f) Home module. The homepage of the website is the first page pop up when we open this website. Any user can directly enter the first page. The block diagram of the homepage design of this website is shown in shown in Figure 5-17.

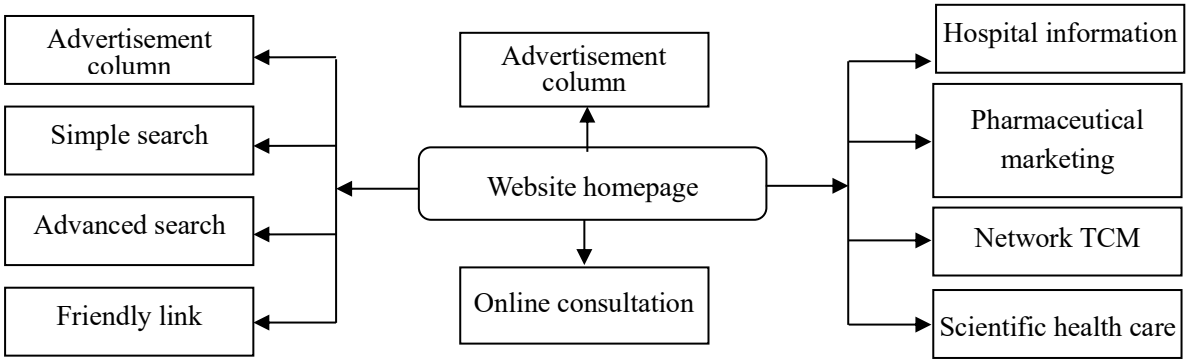


Figure 5-17. The structure diagram of homepage of website.

5.7.2 Page design and implementation

(1) Homepage

The homepage of this website is the first-level interface of system. Any user can view this page. It mainly includes simple search column, news list column, advanced search column, friendly link column and advertisement column. User can enter hospital information, pharmaceutical marketing, network TCM and scientific health care via the navigation at the top of the homepage. The homepage of the website is shown in Figure 5-18.

Figure 5-18. The homepage of website.

(2) Simple search

Simple search is a function bar that allows users to quickly query the required information

with keywords. The user only needs to input the name of any hospital, the name of a doctor, the attending disease, the Pharmaceutical company, drug name and other keywords in the text box, then the corresponding information can be queried. A simple search box page is shown in Figure 5-19.

Figure 5-19. The simple search box.

(3) Advanced search

The user selects "type", enters the keywords such as name, address, pharmaceutical company, medicine, *etc.* Then, the user can select the search type as "and" or "or" to find the corresponding result. The advanced search box page is shown in Figure 5-20.

Figure 5-20. The advanced search box.

(4) Announcement

The announcement column can be added in the background, and can be displayed on the homepage after passing the review. The announcement page is shown in Figure 5-21.

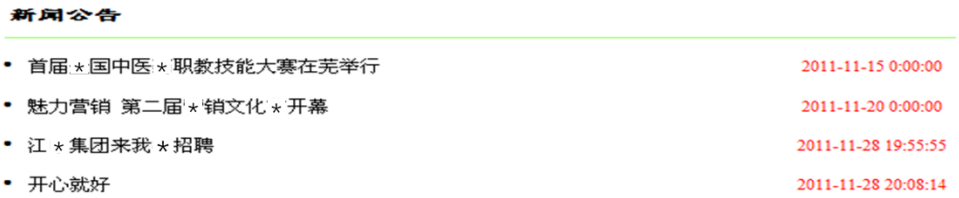


Figure 5-21. The announcement page.

(5) Hospital information page

The hospital information page mainly includes several modules, such as hospital list, hospital hotspot and famous doctor recommendation. It is added by the hospital representative in the background. After being audited by the administrator, it can be displayed in the page. A screenshot of the hospital information page is shown in Figure 5-22. Users click on the hospital name to enter the second-level page, as shown in Figure 5-23. Users click on the doctor's name to enter the doctor's detail page, which is the third-level page, as shown in Figure 5-24.

(6) Pharmaceutical marketing page

The pharmaceutical marketing page contains three modules of drug sales ranking, new drug recommendation, and drug list. Medical representatives should enter the background to add information. After being reviewed by the webmaster, it can be displayed in the page. The pharmaceutical marketing page is shown in Figure 5-25. Users click on the drug name to enter

the second-level page, as shown in Figure 5-26.

Figure 5-22. Hospital information page.

Figure 5-23. The second-level of hospital information.

Figure 5-24. The third-level of hospital information.

Figure 5-25. The pharmaceutical marketing page.

Figure 5-26. The second-level page of pharmaceutical marketing.

(7) Network TCM page

This page is for transmission of TCM theory, information and culture. The TCM page is shown in Figure 5-27.

Figure 5-27. TCM page.

(8) Scientific health care page

This page is for the transmission of health care methods, the recommendations of health experts and the experience of the elderly based on the guidance of the TCM theory, which can provide users with a healthier life. The scientific health care page is shown in Figure 5-28.

Figure 5-28. The Scientific health care page.

(9) Advertisement column page

The advertisement column is mainly for advertisement page of hospitals, drugs, famous doctors and pharmaceutical companies. In addition, there is a module that follows the page movement. The advertisement column page is shown in Figure 5-29.

Figure 5-29. The advertisement column page.

The current level of development and utilization of TCM information resources is still not high. There are some problems in the management of TCM information resources. At present, we are in the critical period of informatization of TCM resources. Can we seize and make good use of this opportunity? It is directly related to the development of TCM and its competitiveness

in the international community. In addition, the construction of TCM informatization requires further exploration and innovation, and strong support from the government. We must gradually promote the formulation of standards for TCM informatization, establish and improve the training model for TCM information resources, and improve the literacy and innovative spirit of the practitioners of TCM information industry. Then, the level of TCM informatization is improved, thereby promoting the development of TCM science and the modernization of human health. The information security issue of TCM has become an important factor affecting the development of health care. It has become an urgent problem to be solved in the process of TCM informatization. The regulations on information security of TCM, information security technology of TCM, information security standards of TCM and information security management system of TCM are the key research subjects in the field of information security of TCM. These has been discussed in this chapter. However, it needs to be further supplemented. In addition, we must strengthen the institutional reform of information security in TCM, so that it can constantly adapt to the rapid development of TCM informatization. The information security of TCM will play an increasingly important role in promoting the reform of the health system, safeguarding the interests of the public and protecting the security of the country. With the further study of TCM information, traditional research methods have been unable to meet the demands of the development of modern TCM. Many pharmaceutical workers are working hard to discover new methods to achieve breakthroughs in the information research of TCM. Based on the research of the strategy pattern and data mining method, the idea of combining modern technology with the research of TCM is put forward. The TCM data mining based on the strategy pattern is an effective way to realize the scientific management and effective utilization of TCM, especially to improve the depth and width of TCM data mining. In addition, it provides more information resources for the industrialization and modernization of TCM.

References

1. L. Sun, Y. Li, H.Y. Wang. Discussion on the common strategies and development of information security. *Science & Technology Information* **2013**, *19*, 11-12.
2. J. Yu, D.Y. Li, H.Y. Wang. Studies explore the data mining technology in pharmacy informatics application. *Chinese Medicine Modern Distance Education of China* **2010**, *15*, 59-60.
3. P.F. He, Y.H. Lu, X.F. He. The proposition of medical information resources distribution efficiency evaluation. *Journal of Medical Informatics* **2009**, *30*, 46-49.
4. Z.W. Cheng, Y.W. Chen. Suggestions for the construction of public health information platform based on medical information. *China Pharmacy* **2013**, *24*, 392-394.
5. L. Tang. Digitalization construction and integration of medical information resources in hospitals. *Journal of Medical Informatics* **2011**, *32*, 73-75.
6. J.R. Hu, Y.L. Zhang. Medical Information Literacy. Military Science Publishing House: Beijing, China, 2007.
7. Y.Y. Wu, X.Y. Zhang. Problems and solutions of the information systems for traditional Chinese medicine. *Lishizhen Medicine and Materia Medica Research* **2009**, *20*, 2583-2585.
8. H.Y. Zhou, Q.R. Zhang. The application of information technology in traditional Chinese

Medicine. *Chinese Journal of Information on Traditional Chinese Medicine* **2011**, *18*, 110-112.
9. F.H. Meng, F. Wan. Construction and development of traditional Chinese medicine in China. *Chinese Journal of Information on Traditional Chinese Medicine* **2010**, *17*, 3-5.
10. B. Sun. Analysis of information management based on modern network information Security. *Information & Communications* **2013**, *2*, 134-135.
11. L.L. Huang. Problems in construction of TCM databases in China and their solution measures. *Chinese Journal of Medical Library and Information Science* **2012**, *21*, 18-20.
12. B. Liu. Research on information security management system. *Computer CD Software and Applications* **2013**, 169-170.
13. B. Yang. On information security. *Digital Technology & Application* **2013**, 198-199.
14. W. Song, X.L. He, G.Y. Zhang. Application and simulation of association rules in medicine cloud data orientation. *Computer Simulation* **2013**, *2*, 239-242.
15. Y.Y. Wu, X.Y. Zhang. Research on quantization of traditional Chinese medicine information. *Journal of Jiangxi University of Traditional Chinese Medicine* **2008**, *20*, 56-57.
16. J.W. Han, M. Kanber. Data Mining: Concepts and Techniques. USA Morgan Kaufmann Publishers Inc.: San Mateo, CA, USA, 2001.
17. D. Cheung, H.Y. Hwwang. Efficient rule-based attribute-oriented induction for data mining. *Journal of Intelligent information System* **2000**, *15*, 175-200.
18. X. Wang. Research on the integration process of medical information based on data mining technology. *Journal of Qiqihar Medical College* **2010**, *6*, 911-912.
19. Q.F. He, X.Z. Zhou, Z.M. Zhou, *et al*. Cluster analysis based on the efficacy of traditional Chinese Medicine. *Chinese Journal of Information on TCM* **2004**, *11*, 561-562.
20. H.Y. Yu, C.G. Xu. Relationship between nature and other properties of traditional Chinese medicine based on association rule. *Chinese Journal of Experimental Traditional Medical Formulae* **2013**, *19*, 343-346.
21. E.X. Shang, X.S. Fan, J.Y. Duan, *et al*. Data mining study on incompatibility characters of Chinese herbal medicine in accordance with association rules. *Journal of Nanjing University of Traditional Chinese Medicine* **2010**, *26*, 421-424.
22. B. Zhang. Research on data-mining technology applied traditional Chinese prescription compatibility based on association rules. *Journal of Gansu Lianhe University (Natural Science Edition)* **2011**, *1*, 82-86.
23. C.B. Xiu. Artificial Intelligence. Mechanical Industry Press: Beijing, China, 2011.
24. M.P.S. Brown, W.N. Grundy, D. Lin, *et al*. Knowledge-based analysis of microarray gene expression data by using support vector machines. *PNAS* **2000**, *97*, 262-267.
25. Z. Pawlak. Rough set. *International Journal of Computer and Information Sciences* **1982**, *11*, 341-356.
26. W.X. Zhang, W.Z. Wu. An Introduction and a survey for the studies of rough set theory. *Fuzzy Systems and Mathematics* **2000**, *4*, 1-12.
27. C.D. Wu, Z. Li, Z.H. Han, T. Yan. Approach to data mining based on rough sets and decision tree. *Journal of Northeastern University* **2006**, *5*, 481-484.
28. Z.C. Zhang, X. Guan, Y. He, W.F. Guo. New research development on rough sets and its

applications. *Computer and Modernization* **2009**, *11*, 16-21.
29. A. Kusiak, J.A. Kern, K.H. Kernstine, *et al*. Autonomous decision-making: a data mining approach. *IEEE Trans Inf Technol Biomed* **2000**, *4*, 274.
30. E.M. Pierce. Developing and delivering a data are housing and mining course. *Communications of the Association for Information Systems* **1999**, *2*.
31. E.X. Shang, L. Ye, X.S. Fan, *et al*. Discovery of association rules between TCM properties in crude drug pairs by mining between datasets and probability test. *World Science and Technology (Modernization of Traditional Chinese Medicine and Materia Medica)* **2010**, *12*, 377-382.
32. X.Y. Zhang, Q. Yang, M. Zhou, *et al*. Development of the network multimedia courseware based on ADO.NET. *Journal of Jiangxi University of Traditional Chinese Medicine* **2009**, *21*, 88-91.
33. F. Usama, P. Gregory, S. Padhraie. Knowledge discovery and data mining: Towards a uniting frame work. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* **1996**, 82-88.
34. R. Jin, B. Lin, B. Zhang, *et al*. A study of association rules in three-dimensional property-taste-effect data of Chinese herbal medicines based on Apriori algorithm. *Journal of Chinese Integrative Medicine* **2011**, *9*, 794-803.
35. J. Han, M. Kamber. Data Mining: Concepts and Techniques. High Education Press: Beijing, China, 2001.
36. R. Hgmwal, T. Imielinski, A. Swami. Ming association rules between sets of items in large databases. *Proceeding of the ACM SIGMOD haft Conference Management of Data* **1993**, *2*, 207-216.
37. R.S. Agawal. Fast algorithms for mining association rules. *Proceeding of the 20th International Conference on Very Large Data Bases* **1994**, 487-499.
38. M. Yang, Z.H. Sun. An incremental urging algorithm based on pericentral list for association rules. *Chinese Journal of Computers* **2003**, *26*, 1318-1325.
39. J. Han, J. Pei, Y. Yin. Mining frequent patterns without candidate generation. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data* **2000**, 1-12.
40. S.F. Hl, Z.D. Hl. Fast mining maximum frequent itemsets. *Journal of Software* **2001**, *12*, 293-297.
41. Y.Q. Song, Y.Q. Zlau, Z.H. Sum, *et al*. An algorithm and its updating algorithm based on FP-Tree for mining maximal itemsets. *Journal of Software* **2003**, *14*, 1586-1592.
42. X.Y. Zhang, S.S. Luo, F.F. Xiao, W.W. Li. Design and implementation of traditional Chinese medicine sought consulting system. *Guiding Journal of Traditional Chinese Medicine and Pharmacy* **2014**, *13*. 25-29
43. S.S. Luo, X.Y. Zhang, F.F. Xiao, W.W. Li. Design and implementation of information system for adverse reactions of traditional Chinese Medicine. *Lishizhen Medicine and Materia Medica Research* **2015**, *1*, 240-242.
44. S.S. Luo, X.Y. Zhang, C.Q. Zhang, W.W. Li, C.C. Qi. Research and system design of traditional chinese medicine data mining based on strategy model. *Modernization of Traditional Chinese Medicine and Materia Medica-World Science and Technology* **2015**, *5*, 929-933.

Chapter 6 The Latest Research Achievements of Data Mining in Traditional Chinese Medicine

Mining and Correlation Analysis of Association Rules between Properties and Therapeutic Efficacy of Chinese Materia Medica Based on Strategy Pattern

Di-Yao Wu[1], Xin-You Zhang[1*], Xiao-Ling Zhou[2]

[1]College of Pharmacy, Jiangxi University of Traditional Chinese Medicine, Nanchang, Jiangxi 330004, China

[2]College of Computer Science, Jiangxi University of Traditional Chinese Medicine, Nanchang, Jiangxi 330004, China

To the Editor Chinese physicians often address the combination of the properties and therapeutic efficacy of Chinese materia medica (CMM). They believe that the properties and therapeutic efficacy of traditional Chinese medicines (TCMs) should be considered as an "organic whole." "Use based on therapeutic efficacy" can lead to omission of the properties of CMM. It also fails to meet the requirements of "differential diagnosis and treatment," thus restricting the flexibility in using CMM. The four natures, five flavors, channel tropism, and therapeutic efficacy of CMM are corresponded to visceral syndromes, which indicate the concept of "wholeness" in TCM.[1] Currently, many studies have been conducted to mine association rules between the properties and therapeutic efficacy of CMM. Some researchers mined text data for correlation between the properties and therapeutic efficacy of CMM. Although some results are obtained, limitations exist in the results, and it remains undetermined whether the obtained association rules are significant. In this paper, based on strategy pattern, we designed and developed a system for mining association rules between the properties and therapeutic efficacy of CMM. Strategy pattern defines a family of algorithms, encapsulates each one and makes them interchangeable. This pattern enables the implementations or interchange of different algorithms according to the option of the user.[2] Thus, we can choose the appropriate algorithm on the basis of the characteristics of the data set. Strategy pattern was fully utilized for the selection of proper mining algorithm, which could make the association mining process user-friendly, intelligent, and fast.

First of all, we need to carry out the steps of quantization of Chinese medicine data. Quantization refers that by reasonable analysis the complex CMM data are divided into independent and exclusive minimum information units that consist of several Chinese characters and numbers and cannot be subdivided.[3] Regarding research subject selection, the authoritative work Pharmacopoeia of the People's Republic of China 2015 edition (hereinafter referred to as Pharmacopoeia) was taken as reference, and the information of 619 CMM recorded in the Pharmacopoeia was extracted. The extracted information included the natures, flavors, channel tropisms, toxicities, therapeutic efficacies, and indications of CMM. Since the characteristics of ascending, descending, floating, and sinking of CMM are not included in the Pharmacopoeia, we also did not consider these characteristics in this paper. During information extraction, 13 CMM with incomplete information was excluded, and thus, a total of 606 CMM

were studied in this paper. In the initial extraction of CMM information, Chinese punctuation symbols such as comma (，), enumerated comma (、), full stop (。) and so on were used as split points for information extraction. If the toxicity of medicine is not given, it can be considered as nontoxic. In the descriptions concerning the therapeutic efficacies and indications of CMM, different words/phrases are often used to mean the same thing. Therefore, one word/phrase among these different words/phrases was selected for use. The information about the therapeutic efficacy and indications of CMM when used externally was kept.

According to the definition of quantization, vague descriptions should be processed to ensure the exclusivity and uniqueness of CMM data. On this basis, data mining can be performed. In the Pharmacopoeia, the descriptions about the natures, flavors, channel tropism and indications of each CMM have already met the requirements of quantization, thus quantization of these data is unnecessary. However, the descriptions about the toxicity and therapeutic efficacy of each CMM are relatively vague and not very exclusive. Therefore, the toxicity and therapeutic efficacy information was quantized. In the information table, the toxicity was classified into no toxicity, little toxicity, having toxicity, and high toxicity. "Having toxicity" means its toxicity is between "little toxicity" and "high toxicity". In order to compare the toxicity of each CMM directly, we used "medium toxicity" to replace "having toxicity" in this quantization research. In the Pharmacopoeia, the therapeutic efficacy of CMM is described using natural language, thus there is a certain degree of vagueness. For example, the therapeutic efficacy of Folium Callicarpae Formosanae is described as "cool blood/astringe/stop bleeding, remove blood stasis/eliminate toxicity/reduce swelling." However, "cool blood/astringe/stop bleeding" includes seven meanings: cool blood; astringe; stop bleeding; astringe by cooling blood; stop bleeding by cooling blood; astringe and stop bleeding; astringe by cooling blood and finally stop bleeding. "Remove blood stasis/eliminate toxicity/reduce swelling" also includes seven meanings: remove blood stasis; eliminate toxicity; reduce swelling; remove blood stasis and eliminate toxicity; remove blood stasis and reduce swelling; eliminate toxicity and reduce swelling; remove blood stasis, eliminate toxicity and reduce swelling. In order to guarantee the exclusivity of therapeutic efficacy, a total of 692 therapeutic efficacies involved in this study were quantized. For example, "remove heat/eliminate toxicity" was quantized into "remove heat," "eliminate toxicity," and "eliminate toxicity by removing heat." "Dispel wind-dampness" was quantized into "dispel wind," "dispel dampness," and "dispel wind and dampness."

Then, we set about designing the association mining system for properties and therapeutic efficacy of CMM. The association mining system was developed on Matlab2014a platform. The operating system adopted was Windows 10. In order to achieve the interactive design, the system integrated Matlab object-oriented programming and graphical user interface design.

Four algorithms, including Apriori, Eclat, DF-FIMBII, and CBM-Eclat, were used for frequent pattern mining of association rules between the properties and therapeutic efficacy of TCMs in the database of quantized TCM information. After frequent patterns were obtained, strong association rules needed to be found. For the above four algorithms, frequent pattern sets with the same form were generated: $L = L_1 \cup L_2 \cup \cdots \cup L_k$, where L_k denoted

the frequent k-itemset. Each frequent itemset was ordered. The simulation was performed in Matlab. Different minimum support thresholds (0–1) were set for the quantized TCM dataset.

With each minimum support threshold, each of the four algorithms was run for three times. The average of three running times was reported as the running time of algorithm (round to three significant figures). After comparing the running time of each algorithm, Apriori and DF-FIMBII turned out to be the two most efficient algorithms. When the minimum support threshold $\geqslant 0.07$, the efficiency of Apriori was higher than that of DF-FIMBII. When the minimum support threshold <0.07, however, the efficiency of DF-FIMBII was obviously higher than that of Apriori. At large values of minimum support threshold, the number of frequent patterns that met requirements was very small, and DF-FIMBII spent much time transforming data into a vertical data format. At small values of minimum support threshold, there were many frequent patterns that met the requirements. In this circumstance, Apriori would scan the database for several times and a large amount of candidate itemsets would be generated, thus leading to a long running time. Due to their high efficiencies, Apriori and DF-FIMBII were integrated into the system for CMM dataset mining. The user can choose the appropriate algorithm according to the characteristics of data set based on strategy pattern.

After algorithm selection, we began to design the system module. The system consists of three modules: data import module, parameter setting module, and result display module. In data import module, the database that needs to be mined can be imported by clicking the data import button. The data should be in XLS or XLSX format, and there should be no TID markers or column markers. After the data are imported, their properties can be automatically displayed: the number of affairs, the number of items, and the average length of affairs (the total number of items in all affairs divided by the total number of affairs). In parameter setting module, users can input the minimum support threshold (0–1) and the minimum confidence threshold (0–1). Then, additional correlation measures can be selected. If Chi-square test is chosen, level of significance (a) should also be selected. For Chi-square test, value 1 was defined as being relevant, and value 0 was defined as being irrelevant. Subsequently, mining method should be chosen. If users themselves do not select a mining method, the system can automatically choose a proper mining method according to the minimum support threshold. For TCM text data, Apriori will be selected if the minimum support threshold >0.07; otherwise, DF-FIMBII will be selected. After parameters are set, we can click the mining button to mine association rules. Moreover, the system will automatically mine association rules from dataset. The progress of the mining process can be seen, which is conducive to interaction between users and the system. When the association mining task is finished, results will be automatically displayed. Nonempty results will be saved as XLSX file. The name of the saved file is the name of the original XLSX file + minimum support threshold + minimum confidence threshold, which makes it convenient for users to further check the results.

After the system was designed, we could run the mining of properties and therapeutic efficacy of CMM and analyze the results. First, the quantized TCM information database was imported into the system of association rule mining. According to the results of the running time of algorithms, only a small amount of association rules can be obtained when the minimum support threshold $\geqslant 0.2$. Therefore, the minimum support threshold was set at 0.08, and the minimum confidence threshold was set at 0.6. All additional correlation measures were selected. For Chi-square test, level of significance a = 0.025. Ultimately, a total of 133 association rules were obtained the results of association rule mining are listed in Supplementary Table 1. As for

rule correlation analysis, different results will be produced due to the difference of evaluation methods adopted. The mining system designed in this paper contains six additional evaluation methods, and their meanings of results are as follows:

All confidence (A, B) refers to the minimum confidence of association rules "A=>B" and "B=>A" related to the two itemsets of A and B. If the value of all confidence (A, B) is <0.5, A and B are negatively correlated. If the value of all confidence (A, B) is equal to0.5, A and B are neutral with no obvious positive or negative correlation. If the value of all confidence (A, B) is >0.5, A and B are positively correlated. The final results showed that 14 of the rules were positively correlated and 119 were negatively correlated using all confidence to evaluate the association rules.

Max confidence (A, B) refers to the maximum confidence of association rules "A=>B" and "B=>A" related to the two itemsets of A and B. If the value of max confidence (A, B) is <0.5, A and B are negatively correlated. If the value is equal to 0.5, A and B are neutral with no obvious positive or negative correlation. If the value is >0.5, A and B are positively correlated. The final results showed that all the rules were positively correlated using the maximum confidence evaluation.

Lift (A, B) denotes the ratio of the probability of containing B under the condition of containing A to the probability of containing B without A. Lift (A=>B) >1 indicates that the rule context is positively correlated and Lift (A=>B) <1 indicates that the rule context is negatively correlated, while Lift (A=>B) = 1 indicates that the context is not correlated. The final results showed that 104 rules were positive correlation and 29 rules were a negative correlation. Chi-square test is the deviation degree between the actual observed value and the theoretical inferred value of the statistical sample. The Chi-square value is compared with the P value. If it is greater than or equal to the P value, the rule is of significance. If the value of Chi-square is less than the P value,the rule is not significant. The final results showed that all the Chi-square values were 0 and less than P values, so there was no significant difference in all rules.

Kulc (A, B) refers to the mean value of two conditional probabilities (the probability of containing B under the condition of containing A; the probability of containing A under the condition of containing B). If the value of Kulc (A, B) is <0.5, A and B are negatively correlated. If the value is equal to 0.5, and A and B are neutral. If the value is >0.5, and A and B are positively correlated. The final results showed that 82 rules were positively correlated and 51 rules were negatively correlated.

Cosines (A, B) can be viewed as the harmonic lift measure. It is similar to Lift except that the cosine takes the square root of the product of the probabilities of A and B. If the value of Cosine (A, B) is <0.5, A and B are negatively correlated. If the value is equal to 0.5, A and B are neutral. If the value is >0.5, A and B are positively correlated. The final results showed that 38 rules are positive and 95 rules are negative.

According to the experimental result, the system designed with proper mining algorithm selection enables to obtain good association rule mining results, and the operation is intelligent and fast. In addition, correlation measures are integrated into the system. During operation, users are free to choose among these measures for the evaluation of correlation between the antecedents and consequents of rules. For example, regarding the rule "cool blood => no

toxicity" listed in Supplementary Table 1, 8.5809% of 606 CMM are nontoxic and at the same time are able to cool blood. The 100% confidence level demonstrates all CMM that can cool blood are nontoxic. All confidence and cosine show that there is a negative correlation between the antecedent and consequent of this rule. However, max confidence, lift, and Kulc reveals positive correlation between the antecedent and consequent of this rule. The rule analysis results of the system are intuitive, diverse, and user-friendly. Correlation measures can be selected in accordance with mining targets. Alternatively, "voting principles" can be used. For example, if four out of the six correlation measures reveal that there is a positive correlation between the antecedent and consequent of a certain rule, a positive correlation is recorded as the evaluation result. In other cases, such as if three correlation measures reveal positive correlation and the other three reveal a negative correlation, "support-confidence" is used as standard [4] to determine whether the certain rule is significant.

As we know, many results have been obtained from CMM data mining, especially association mining of the properties and therapeutic efficacy of TCMs. Before association mining, researchers often need to process the data. However, during preprocessing, they may simply split the phrases into several shorter ones. For example, "remove heat and eliminate toxicity" is splitted d into "remove heat" and "eliminate toxicity." They may also combine phrases with similar meaning into one phase/word. For instance, "extremely cold" and "slightly cold" are combined as "cold." Such pretreatments may ignore the fact that some CMM have similar properties or therapeutic efficacy, but their strengths are different. These pretreatments may also fail to include other possible meanings of phrases describing the therapeutic efficacy of CMM. These shortcomings suggest that current CMM data preprocessing methods often lead to loss of information. Therefore, the concept of quantization was introduced here. Some vague descriptions about the therapeutic efficacy CMM were quantized. On this basis, a database of quantized CMM information was constructed, and this was conducive for subsequent CMM data. Due to the database, more information can be retained, and data can be described clearly and in more detail. The mining results are thus more accurate.

Generally speaking, the researchers often use one method for data mining of properties and therapeutic efficacy of CMM ("one to one" pattern, i.e., one method solves one problem).[5] In fact, there are many mining algorithms available. These algorithms have different efficiencies under different conditions, and the selection of the best algorithm depends on circumstance. If only one algorithm is used, the mining efficiency can be affected. This is also not conducive to the improvement of mining method. Moreover, many studies neglect to evaluate the obtained association rules. Therefore, even if strong rules are obtained, valuable conclusions still cannot be drawn since the rules might be insignificant. In order to address these problems, we designed and developed a system based on strategy pattern for mining association rules between properties and therapeutic efficacy of CMM. Users are free to choose among several mining algorithms and proper mining algorithm can be chosen using the strategy pattern. These make the association mining process more intelligent, fast, and convenient. The results are also more intuitive. Therefore, correlation analysis can be easily performed to determine whether these association rules are significant.

Although the database can retain as much information as possible, the quantization process is not aimed for specific CMM, thus limitations may exist. Moreover, the minimum support

threshold was still set in the system, and it may be not suitable for other datasets. Therefore, our future research will focus on how to overcome these limitations.

Supplementary information is linked to the online version of the paper on the Chinese Medical Journal website.

Financial support and sponsorship

This study was supported by a grant from the National Natural Science Foundation of China (No. 81660727).

Conflicts of interest

There are no conflicts of interest.

References

1. L.M. Sun. Synchronous treatment of the heart and brain and the holistic view of traditional Chinese medicine (in Chinese). *Journal of Traditional Chinese Medicine* **2012**, *53*, 1705-1706.
2. S.S. Luo, X.Y. Zhang, C.Q. Zhang, W.W. Li, C.C. Qi. Data mining study and system design of traditional Chinese medicine based on strategy model (In Chinese). *World Science Technology Modern Journal of Traditional Chinese Medicine* **2015**, *5*, 929-933.
3. L.J. Liu. Research and application of improved Apriori algorithm (in Chinese). Computer & Digital Engineering 2017, 38, 3324-3328.
4. L. Chen, S. Feng. Two-level confidence threshold setting method for positive and negative association rules (in Chinese). *Journal of Computer Applications* **2018**, *38*, 1315-1319.
5. R. Jin, Q. Lin, B. Zhang, X. Liu, S.M. Liu, Q. Zhao, et al. A study of association rules in three-dimensional property-taste-effect data of Chinese herbal medicines based on Apriori algorithm (in Chinese). *Journal of Chinese Integrative Medicine* **2011**, *9*, 794-803.

Key CMM Combinations in Prescriptions for Treating Mastitis and Working Mechanism Analysis Based on Network Pharmacology

Diyao Wu,[1] Xinyou Zhang,[1*] Liping Liu,[2] and Yongkun Guo[2]
[1]*School of Pharmacy, Jiangxi University of Traditional Chinese Medicine, Nanchang, Jiangxi Province 330004, China*
[2]*College of Computer Science, Jiangxi University of Traditional Chinese Medicine, Nanchang, Jiangxi Province 330004, China*
Correspondence should be addressed to Xinyou Zhang; xinyouzhang@jxutcm.edu.cn

Aims Using both data mining and network pharmacology methods, this paper aims to construct a molecule-target-disease network for medicines used for treating mastitis, mine out targets, and signaling pathways related to mastitis and explore the mechanism of Chinese materia medica (CMM) prescriptions in treating mastitis. Methods. A total of 131 CMM prescriptions for treating mastitis were collected from clinical practice and related literatures. A database of prescriptions for treating mastitis (DPTM) was then constructed. Based on data mining method, Traditional Chinese Medicine Inheritance Support System (TCMISS) was employed to mine out high-frequency CMM and key CMM combinations in DPTM. Subsequently, TCM Systems Pharmacology Database and Analysis Platform (TCMSP) and Traditional Chinese Medicine Information Database (TCM-ID) were searched for the targets of ingredients of high-frequency CMM. Then, Bioinformatics Analysis Tool for Molecular Mechanism of TCM (BATMAN-TCM) was searched for diseases and signaling pathways corresponding to the targets of key CMM combinations. The obtained results were denoted as results 1. In addition, human disease database MalaCards was searched for targets and signaling pathways related to mastitis. The obtained results were denoted as results 2. Results 1 and 2 were compared to obtain targets and signaling pathways included in both results, namely, mastitis-related targets of TCMs and mastitis-related signaling pathways that CMM involves in. Then, the biological functions of these targets and signaling pathways were investigated, on which basis the mechanism of CMM prescriptions in treating mastitis was explored. Results. A total of 12 key TCM combinations were identified. Taraxaci Herba, Glycyrrhizae Radix et Rhizoma, Paeoniae Radix Alba, semen citri reticulatae, etc. were CMM with the highest frequency of use for treating mastitis. The potential targets of these high-frequency CMM in treating mastitis were intercellular adhesion molecule 1 (ICAM-1), interleukin-6 (IL-6), lipopolysaccharide binding protein (LBP), and lactotransferrin. The potential signaling pathways that key CMM combinations may involve in during mastitis treatment were NF-κB signaling pathway, immune system, PI3K/Akt signaling pathway, and TNF signaling pathway. Conclusions. From a perspective of network pharmacology, molecule-target-disease analysis may serve as an entry point for the research of

mechanism of CMM. On this basis, we studied the mechanism of CMM prescriptions in treating mastitis by data mining and comparison of results. Our work thus provides a new idea and method for studying the working mechanism of CMM prescriptions.

1 Introduction

Mastitis is a disease commonly occurring in lactating women. The most frequent type of mastitis is acute suppurative mastitis with clinical symptoms of lumps in breast, swelling, pain, fever, and pus oozing out. In traditional Chinese medicine (TCM), acute mastitis is called "breast carbuncle". This name was first seen in the book Zhenjiu Jiay (women's miscellaneous disease ten, volume ten) written by Huangfu Mi of the Jin Dynasty [1]. Since then, this disease has been studied by TCM doctors of successive dynasties. Therefore mastitis has been identified in ancient times in TCM history. Rich experience has been accumulated for its treatment and many classic prescriptions have remained in use until today. However, due to the difference in clinical experience among TCM doctors and the complexity of TCM, the prescriptions for treating mastitis vary greatly from each other. Moreover, related research mainly focuses on the causes of mastitis and the summary of experience. These are a lack of in-depth research on medication rules in prescriptions and their working mechanism. In this paper, we collected prescriptions for treating mastitis from clinical research literatures and clinical practice in recent decade. The key Chinese materia medica (CMM) combinations in the prescriptions for treating mastitis as well as their potential targets and signaling pathways were analyzed. The results may provide useful information for the treatment of mastitis and the study of working mechanism of CMMs.

For the compatibility of medicines in CMM prescriptions, a "monarch-minister-assistant-messenger" rule should be followed. Various CMMs are used in combination to treating imbalance and disorders in the body. This is because the use of single CMM can hardly achieve high therapeutic efficacy, which indeed illustrates the idea of "multicomponents, mul-titargets, and systematic regulation" in TCM theory. Previous researches mainly attempt to explain the pharmacology of CMM son the basis of the drug activity of single molecule and the effect of single target, which often fail to completely explain the working mechanism of CMMs.

With the introduction of systems biology and the application of bioinformatics, network pharmacology is also proposed. Based on the interaction among diseases, genes, targets, and medicines, network pharmacology enables to comprehensively investigate the effects of medicines on diseases. If key CMM combinations (namely, high-frequency CMM combinations) for treating mastitis are mined out, then a "key CMM combination-target-disease" network can be constructed. Subsequently, signaling pathway enrichment analysis of targets can be performed. Then, the mechanism of multiple compounds in the cooperative treatment of mastitis can be explained from the perspective of network pharmacology. This method agrees with the idea of holistic medicine and intuitively illustrates the mechanism of multi-system regulation in TCM. It also constructs a bridge between traditional Chinese medicine and western medicine since it enables investigating CMM prescriptions from a perspective of target-disease relationship, which is highlighted in western medicine.

2. Methods and search tools

TCM prescriptions for treating mastitis were collected from clinical practice and related literatures and then a database of prescriptions for treating mastitis (DPTM) was con-structed. On the basis of data mining method, Traditional Chinese Medicine Inheritance Support System (TCMISS) was employed to mine high-frequency CMMs and key CMM combinations in DPTM. Then, Traditional Chinese Medicine Systems Pharmacology Database and Analysis Platform (TCMSP) and Traditional Chinese Medicine Information Database (TCM-ID) were searched for the targets of high-frequency CMM. Later, Bioinformatics Analysis Tool for Molecular Mechanism of TCM (BATMAN-TCM) was searched for diseases and signaling pathways corresponding to the targets of key CMM combinations. The obtained results were denoted as results 1. Furthermore, human dis-ease database MalaCards was searched for the targets and signaling pathways related to mastitis. The obtained results were denoted as results 2. Results 1 and 2 were compared to obtain the targets and signaling pathways included in both results, namely, mastitis-related targets of CMM and mastitis-related signaling pathways that CMM involve in. Then, the biological functions of these targets and signaling pathways were identified, on which basis the mechanism of CMM prescriptions in treating mastitis was explored. A flow chart is shown in Figure 1.

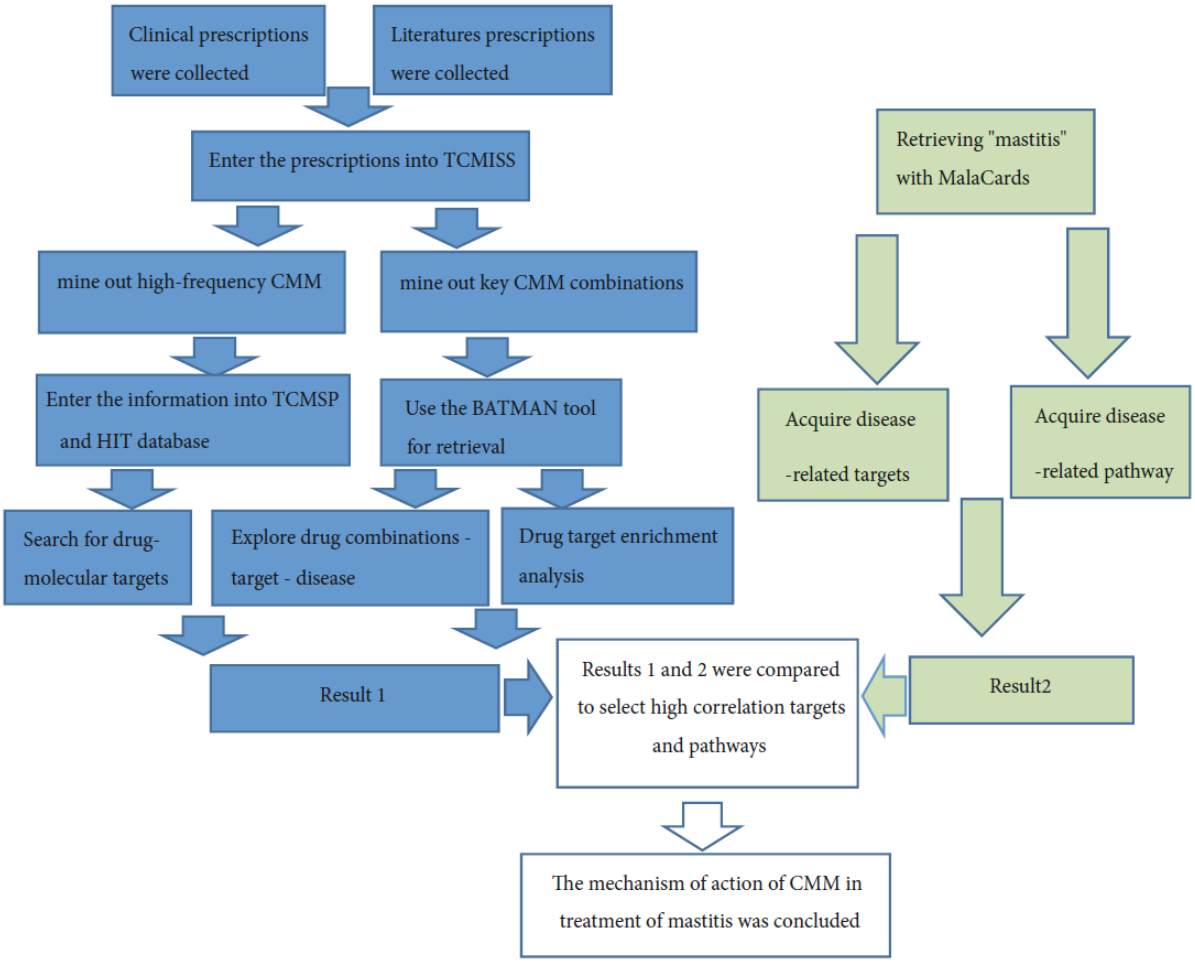


Figure 1. The flow chart of mining Key CMM combinations in prescriptions for treating mastitis and analyzing working mechanism based on network pharmacology.

2.1 Data collection

DPTM consists of two kinds of prescriptions: prescriptions used in clinical practice for treating mastitis and prescriptions from related literatures for treating mastitis. First, from January 2018 to May 2018, all prescriptions for treating diseases in both outpatients and inpatients in a provincial-level breast specialist hospital were collected. Among them, prescriptions for treating mastitis were screened out. After the same prescriptions were excluded, a total of 98 prescriptions for treating mastitis were collected. Second, we searched for papers in PubMed (http://www.ncbi.nlm.nih.gov) and CNKI (http://cnki.net/) with "mastitis" and "Chinese materia medica" as two key-words. Then, a total of 45 prescriptions for treating mastitis were collected from the papers. The above prescriptions were combined and after the same prescriptions were excluded, a total of 131 prescriptions were collected, on which basis DPTM was constructed.

2.2 TCMISS

TCMISS is a platform focusing on analysis of CMM data, which integrates general statistics, text mining, association rules, and complex system entropy clustering method. It has already been widely applied to prescription compatibility investigation and prescription analysis [2,3]. Prescriptions from DPTM were input one by one into TCMISS. Then, the frequencies of CMM were statistically analyzed and CMM were ordered according to their frequencies. Subsequently, association rules method was used to mine out high-frequency combinations of CMM to obtain the key CMM combinations.

2.3 MalaCards

MalaCards is an integrated database of human maladies and their annotations. It is modeled on the architecture and richness of the popular GeneCards database of human genes [4]. MalaCards was searched with "mastitis" as the keyword and then genes, signaling pathways, and other pieces of information related to mastitis were shown.

2.4 TCMSP and TCM-ID

TCMSP includes 499 CMM described in the Pharmacopoeia of the People's Republic of China. It involves 29,384 ingredients, 3311 targets, 837 associated diseases, and pharmacokinetic characteristics of CMM. This platform allows users to check and analyze the drug molecule-target network and drug-target-disease network, which can help reveal the working mechanism of CMM [5]. The obtained 11 CMM with the highest use frequency were input one by one in TCMSP and with additional information from TCM-ID, the ingredients of high-frequency CMM and their targets were obtained.

2.5 BATMAN-TCM

BATMAN-TCM is a Bioinformatics Analysis Tool for Molecular Mechanism of Traditional Chinese Medicine and is the first online bioinformatics analysis tool specially

designed for the study of molecular mechanism of TCM. It mainly performs TCM ingredients' target prediction and the subsequent network pharmacology analyses of the potential targets, aiming to improve the understanding of the "multicomponent, multitargets, and multipathway" combinational therapeutic mechanism of CMM. For each ingredient of CMM, BATMAN-TCM ranks its predicted candidate targets according to the order of decreasing score given by the target prediction algorithm for the drug-target interaction prediction. It uses a similarity-based method to predict the potential targets of CMM ingredients. The core idea of this method is to rank potential drug-target interactions based on their similarity to the known drug-target interactions. If the score of a candidate target ⩾ "score cutoff ", then this target will be taken as the potential target of the ingredient investigated [6].

In BATMAN-TCM, herb list was selected and the 12 key CMM combinations mined out from DPTM were input into the platform. Score cutoff was set at 80 and adjusted P-value was set at 0.05. BATMAN-TCM first predicted the potential targets of each ingredient of CMM investigated and then performed functional analyses of these targets including Gene Ontology (GO), KEGG pathway, and OMIM/TTD disease enrichment analyses. CMM ingredient-target-pathway/disease association network and biological pathways in which CMM's targets are significantly enriched were also shown.

2.6 Target and signaling pathway screening

The composition of CMM is very complicated. The number of targets of CMM ingredients and the number of signaling pathways that they involve in are very large. In order to screen out highly associated targets and signaling pathways, their association with mastitis should be considered. In this paper, MalaCards was searched for targets and signaling pathways related to mastitis. If CMM can also act on the same targets or signaling pathways, then these targets or signaling pathways are taken as highly associated targets and signaling pathways. In this way, the slightly relevant and irrelevant targets and signaling pathways can be excluded (Figures 2 and 3). For example, the activation of intercellular adhesion molecule 1 (ICAM-1) is highly related to mastitis, and heartleaf houttuynia herb can also actonICAM-1. Then, ICAM-1isconsidered as a highly associated target. Therefore, the potential working mechanism of herb houttuynia may be that it inhibits ICAM-1 and thereby exerts an anti-inflammatory effect.

3 Results

3.1 Results of DPTM mining by TCMISS

3.1.1 Frequencies of CMMs

TCMISS was employed to statistically analyze the frequencies of all CMM in DPTM. DPTM includes 131 CMMs and their frequencies are shown in Table 1.

3.1.2 Association rules mining results

Association rules mining of DPTM was performed when support was set to ⩾ 26 and

confidence was set to ⩾ 0.9 in TCMISS. The aim of association rules mining was to find frequent item sets, namely, key CMM combinations frequently appearing in the datasets. After repeated items were removed, 12 key CMM combinations for treating mastitis were obtained (Table 2).

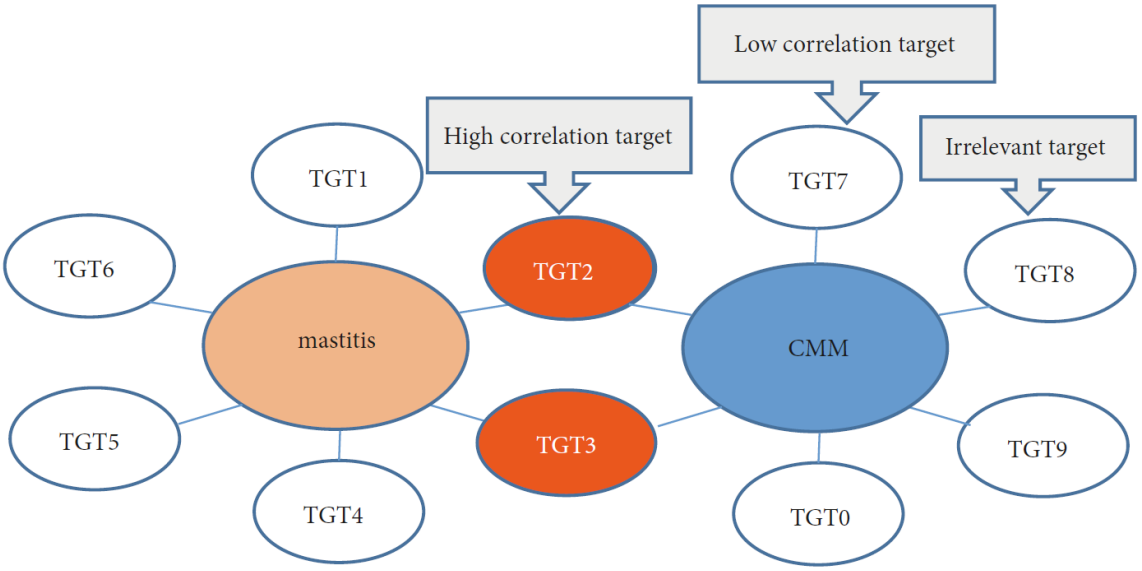


Figure 2. Screening of high related targets of CMM. Annotation: in this research, the orange ovals are high correlation targets, whereas the white ovals may be either low correlation targets or irrelevant targets.

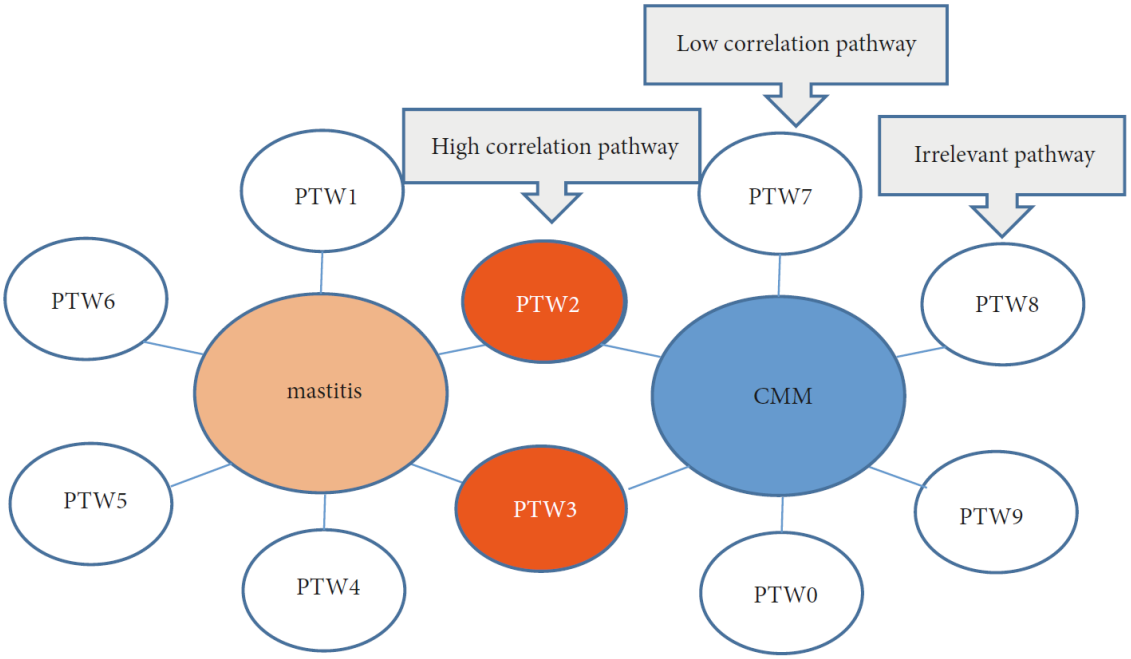


Figure 3. Screening of high related pathways of CMM. Annotation: in this research, the orange ovals are high correlation pathways, whereas the white ovals may be either low correlation pathways or irrelevant pathways.

Table 1. All CMMs in DPTM and their frequencies (Fre).

CMM	Fre	CMM	Fre	CMM	Fre	CMM	Fre
Taraxaci Herba	99	Scutellaria barbata D. Don	16	semen euryalis ferocis	4	Saposhnikoviae Radix	1
Glycyrrhizae Radix et Rhizoma	98	Dendranthema indicum	15	caulis akebiae	4	rhizoma phragmitis communis	1
Paeoniae Radix Alba	64	Asarum sagittarioides C. F. Liang.	15	radix notoginseng	4	Cluster Mallow Fruit Fructus Malvae	1
semen citri reticulatae	60	Penthorum chinense Pursh	12	semen sinapis albae	4	Schizonepetae Herba	1
Bupleuri Radix	59	Forsythia suspensa	10	Ranunculus ternatus	4	rhizoma sparganii	1
pericarpium citri reticulatae viride	53	Fructus Camptothecae Acuminatae	10	nidis vespae	4	Fructus Toosendan	1
pericarpium trichosanthis kirilowii et multilobae	48	Trichosanthes kirilowii Maxim.	10	Carthamus tinctorius L.	3	semen pruni persicae	1
Semen Coicis	47	Fructus hordei Germinatus	10	caulis milletiae seu longae	3	Cuscuta japonica Choisy	1
herba houttuyniae	45	Gardenia jasminoides Ellis	9	Mahonia fortunei ïijĹLindl. ïijĽFedde	3	semen litchi chinensis	1
Semen Vaccariae	42	Manis pentadactyla	9	Anemarrhena asphodeloides Bunge	3	Amomi Fructus	1
Salvia miltiorrhiza Bunge	41	Prunella vulgaris L	9	fructus rosae laevigatae	3	cacumen bioyae orientalis	1
Bulbus Fritillariae Thunbergii	36	Platycodon grandiflorus	9	Manyinflorescenced sweetvetch root	3	Draconis Sanguis	1
Lamiophlomis rotata Kudo	35	Atractylodes Lancea (Thunb.) DC.	8	Herba Thlaspis	3	Hematite Haematitum	1
Astragalus membranaceus	33	Ligusticum chuanxiong hort	8	Massa Fermentata	3	Rhizoma smilacis glabrae	1
Poria	31	fructus liguidambaris taiwaniae	7	Leonurus Linn.	3	Glabrous greenbrier rhizome	1
Retinervus luffae Fructus	30	Bombyx Batryticatus	6	Codonopsis Radix	3	Corydalis yanhusuo W.T.Wang	1
Vigna angularis	30	radix paeoniae rubra	6	Fritillaria cirrhosa D. Don	2	Scrophularia ningpoensis Hemsl	1
Viola yedoensis Makino	30	thallus algae	6	Rehmannia glutinosa (Gaetn.) Libosch. ex Fisch. et Mey.	2	AIpiniatonkinensis Gagnep	1
Trichosanthis Radix	29	Rehmanniae Radix	6	Curcumae Radix	2	Notopterygium incisum Ting ex H. T. Chang	1
Sarcandraglabra	25	Alismatis Rhizoma	6	Rhizoma Pinelliae	2	Phellodendron amurense Rupr	1
Angelicae Sinensis Radix	25	Rhizoma Cyperi	6	Cornu Cervi Degelatinatum	2	Kadsura longipedunculata Finet et Gagnep	1
Curcumae Rhizoma	24	Sargassum	6	Itoa orientalis	2	Dendranthema morifoliumïijĹRamat. ïijĽTzvel	1
Zea mays L	24	Rhodiola rosea L.	6	Polygoni Multiflori Radix	2	Rubia cordifolia L	1
Atractylodes macrocephala Koidz	24	Radix Actinidiae Chinensis	6	Semen Juglandis	2	Curcumae Longae Rhizoma	1
Lonicera japonica Thunb	22	Hemsleya amabilis Diels	6	Hedyotis diffusa	2	Ginseng Radix et Rhizoma	1
Pericarpium Citri Reticulatae	21	Stemmacanthauniflora (L.) Dittrich	5	semen cannabis sativae	2	Radix Aucklandiae	1
Scutellaria baicalensis Georgi	21	Rhizoma Paridis	5	Solanum nigrum Linn	2	radix linderae strychnifoliae	1
Crataegus pinnatifida Bunge	20	Arctium lappa L.	5	Rhei Radix et Rhizoma	2	radix angelicae pubescentis	1
Tetrapanax papyriferus	17	Myrrh	5	Folium Perillae	2	rhizoma arisaematis	1
Spina Gleditsiae	17	Gentiana scabra Bunge	5	Amygdalus Communis Vas	2	fructus germinatus oryzae sativae	1
ostrea gigas tnunb	17	Boswellia sacra	5	Gypsum Fibrosum	2	Arc Shell Concha Arcae	1
Angelicae Dahuricae Radix	16	fructus citri aurantii	4	cortex albizziae julibrissinis	1	Zingiber officinale Roscoe	1
Corni Cervi	16	Plantago asiatica L.	4	Semen Cassiae	1		

Table 2. Key CMM combinations in DPTM.

Number	Key CMM combinations
1	Taraxaci herba, Salvia miltiorrhiza Bunge, Paeoniae Radix Alba and Glycyrrhizae Radix et Rhizoma
2	Taraxaci herba, semen citri reticulatae, Paeoniae Radix Alba and Glycyrrhizae Radix et Rhizoma
3	Taraxaci herba, herba houttuyniae and Glycyrrhizae Radix et Rhizoma
4	Taraxaci herba, Semen Coicis and Vigna angularis
5	Taraxaci herba, herba houttuyniae and Semen Coicis
6	Taraxaci herba,herba houttuyniae and pericarpium trichosanthis kirilowii et multilobae
7	Taraxaci herba, trichosanthes kirilowii peel and Semen Coicis
8	Paeoniae Radix Alba, herba houttuyniae and Glycyrrhizae Radix et Rhizoma
9	Paeoniae Radix Alba, Bupleuri Radix and Glycyrrhizae Radix et Rhizoma
10	semen citri reticulatae, pericarpium citri reticulatae viride and Bupleuri Radix
11	Taraxaci herba and Viola philippica
12	Taraxaci herba and Lamiophlomis rotata Kudo

3.2 Results of searching in MalaCards

3.2.1 Gene targets related to mastitis.

MalaCards was searched with “mastitis” as the keyword for targets and signaling pathways related to mastitis. The results are shown in Tables 3 and 4.

Table 3. Targets related to mastitis disease.

Symbol	**Description**	**Score**
LTF	Lactotransferrin	27.2
CXCL8	C-X-C Motif Chemokine Ligand 8	23.43
TLR2	Toll Like Receptor 2	23.13
IL6	Interleukin 6	22.28
LBP	Lipopolysaccharide Binding Protein	22.18
ALB	Albumin	22.16
CCL5	C-C Motif Chemokine Ligand 5	21.49
NOD2	Nucleotide Binding Oligomerization Domain Containing 2	21.14
IL17A	Interleukin 17A	20.81
HP	Haptoglobin	20.74
CP	Ceruloplasmin	20.69
CSF2	Colony Stimulating Factor 2	20.01
ICAM1	Intercellular Adhesion Molecule 1	19.99
CSN2	Casein Beta	13.66
LALBA	Lactalbumin Alpha	12.71
OXT	Oxytocin/Neurophysin I Prepropeptide	12.49
CXCL6	C-X-C Motif Chemokine Ligand 6	11.6
CSN3	Casein Kappa	11.3
SLPI	Secretory Leukocyte Peptidase Inhibitor	11.2
STAT5A	Signal Transducer And Activator Of Transcription 5A	10.48

3.3 Results of searching in TCMSP and HIT

After the 11CMMswith the highest frequencies were input one by one into TCMSP and HIT (it is used to supplement the information of CMM which is unrecorded in TCMSP), the ingredients and their targets of 11 CMMs were obtained. The obtained targets were compared with those related to mastitis found in MalaCards. The same targets were screened out and taken as the mastitis-associated targets of CMM for treating mastitis (Table 5). Notably, the targets of

Taraxaci Herba, semen citri reticulatae, and Semen Coicis are all not related to mastitis.

Table 4. Signal pathways related to mastitis disease and top affiliating genes of pathways.

Super pathways	Top affiliating genes
Innate Immune System	CCL5,CSF2,CXCL8,HP,ICAM1,IL17A
Akt Signaling	CCL5,CSF2,CXCL6,CXCL8,IL17A,IL6
Cytokine Signaling in Immune system	CCL5,CSF2,CXCL8,ICAM1,IL17A,IL6
Toll-Like receptor Signaling Pathways	CCL5,CP,CXCL8,IL6,NOD2,TLR2
Influenza A	CCL5,CXCL8,ICAM1,IL6,TLR2
Kaposi's sarcoma-associated herpesvirus infection	CSF2,CXCL8,ICAM1,IL6,TLR2
Toll-like receptor signaling pathway	CCL5,CXCL8,IL6,LBP,TLR2
Selenium Micronutrient Network	ALB,CCL5,ICAM1,IL6
IL-17 Family Signaling Pathways	CSF2,CXCL6,CXCL8,IL17A,IL6,TLR2
Tuberculosis	IL6,LBP,NOD2,TLR2
Bacterial infections in CF airways	CXCL8,IL6,LBP,TLR2
Interleukin-4 and 13 signaling	CXCL8,ICAM1,IL17A,IL6,LBP
IL27-mediated signaling events	IL17A,IL6,TLR2
TNF signaling pathway	CCL5,CSF2,ICAM1,IL6,NOD2
AGE-RAGE signaling pathway in diabetic complications	CXCL8,ICAM1,IL6
Amoebiasis	CSF2,CXCL8,IL6,TLR2
NF-kappa B signaling pathway	CXCL8,ICAM1,LBP
Th17 Differentiation Pathway	IL17A,IL6,TLR2
Salmonella infection	CSF2,CXCL8,IL6,LBP
Pertussis	CXCL6,CXCL8,IL6
IgA-Producing B Cells in the Intestine	ICAM1,IL6,TLR2
Lung fibrosis	CCL5,CSF2,CXCL8,IL6
Photodynamic therapy-induced NF-kB survival signaling	CSF2,CXCL8,ICAM1,IL6
Glucocorticoid receptor regulatory network	CSF2,CSN2,CXCL8,ICAM1,IL6
Legionellosis	CXCL8,IL6,TLR2
Cytokine production by Th17 cells in CF	CSF2,CXCL6,CXCL8,ICAM1,IL17A,IL6
Malaria	CXCL8,ICAM1,IL6,TLR2
amb2 Integrin signaling	HP,ICAM1,IL6
Interleukin-10 signaling	CCL5,CSF2,CXCL8,ICAM1,IL6
Rheumatoid arthritis	CCL5,CSF2,CXCL6,CXCL8,ICAM1,IL17A
G-protein signaling RhoB regulation pathway	CCL5,CSF2,CXCL6,CXCL8,IL17A,IL6
G-protein signaling Rap2B regulation pathway	CCL5,CSF2,CXCL6,CXCL8,IL17A,IL6

Table 5. High correlation target of high frequency Chinese medicine in the treatment of mastitis diseases.

CMM	Pharmaceutical molecule	Target
Pericarpium citri reticulatae viride	hesperidin	ICAM-1
Paeoniae Radix Alba	paeoniflorin	IL-6
Bupleuri Radix	methyl palmitate, lauric acid	IL-6
Salvia miltiorrhiza Bunge	oleanolic acid, ursolic acid, luteolin, Tanshinone I, apigenin	ICAM-1
	glycine	lactotransferrin
Radix Salviae miltiorrhizae	glycyrrhizic acid	IL-6
Semen Vaccariae	quercetin	IL-6
Herba houttuyniae	Caryophyllene,rutin,quercetin	IL-6
	Kaempferol, quercetin	ICAM-1
Pericarpium trichosanthis kirilowii et multilobae	lauric acid, methyl palmitate	IL-6
	glycine	Lactotransferrin

3.3.1 Intercellular Adhesion Molecule 1 (ICAM-1)

ICAM-1 is a kind of membrane glycoprotein that participates in the interaction between cells or between cells and extracellular matrices [7]. During the development of inflammation, ICAM-1 has important effects on the directed migration of neutrophils and lymphocytes and their infiltration into surrounding tissues [8]. It is thus closely related to the development of inflammation. pericarpium citri reticulatae viride, Salvia miltiorrhiza Bunge, Paeoniae Radix Alba, and herba houttuyniae can all act on ICAM-1. The mechanism of them in treating mastitis might be that they inhibit ICAM-1 and thereby exert certain anti-inflammatory effects.

3.3.2 Interleukin-6 (IL-6)

Interleukin (IL) is a kind of cytokine that is secreted by certain cells and has an effect on other cells. IL plays an important role in information transfer, activation and regulation of immune cells, activation, proliferation and differentiation of T and B cells, as well as inflammatory response. IL-6, as a member of interleukin family, mainly plays a role in the proliferation of B cells and antibody secretion, proliferation of T cells and CTL activation, formation of acute phase proteins by liver cells, inflammatory response, etc. [9]. The mechanism of Paeoniae Radix Alba, Glycyrrhizae Radix et Rhizoma, Bupleuri Radix, Semen Vaccariae, herba houttuyniae, and pericarpium trichosanthis kirilowii et multilobae in treating mastitis might be that they can decrease IL-6 level.

3.3.3 Lipopolysaccharide Binding Protein (LBP)

LBP is a kind of glycoprotein existing in human and animal serum. LBP has a high affinity with lipoid A in lipopolysaccharide (LPS). It can function as a LPS carrier protein, catalyze the binding of LPS to CD14, stimulate monocytes and endothelial cells, and promote the release of inflammatory mediators such as TNF. LBP can also function as an opsonin, promoting monocytes to engulf conditioned LPS and gram-negative bacteria; thus LBP can regulate inflammatory response induced by LPS [10]. The possible mechanism of Paeoniae Radix Alba in treating mastitis might be that paeoniflorin inhibits the expression of LBP and antagonize LBP-mediated LPS inflammatory response.

3.3.4 Lactotransferrin

Lactotransferrin is a natural glyco-protein with immune functions that exists in breast milk. Its physiological functions include iron absorption promotion, immunomodulation, antibacterial, and antiviral effects, etc.[11]. The mechanism of Glycyrrhizae Radix et Rhizoma in treating mastitis might be that the glycine it contains can regulate lactotransferrin leve1, enhance the body's immunity and exert anti-inflammatory effects.

3.4 Results of searching in BATMAN-TCM

3.4.1 Key TCM combination-target-disease analysis.

BATMAN-TCM was employed to perform molecule-target-disease analysis of the 12 key CMM combinations. Diseases related to mammary gland and mastitis were screened out.

Results show that each key CMM combination is related to two to four of the diseases including inflammation, inflammatory disease, breast cancer, and hormone-dependent breast cancer (Table 6). Hormone-dependent breast cancer refers to that when tumor cells show positive expression of estrogen receptor (ER)/progesterone receptor (PR) and the growth and proliferation of tumor cells are regulated by estrogen and progesterone, antiestrogenic drugs must be used for treatment [12]. As can be seen, the key CMM combinations studied here are closely related to the treatment of mastitis and breast cancer.

3.4.2 Key TCM combination-target-signaling pathway analysis

BATMAN-TCM was used to perform signaling pathway enrichment analysis of the targets of 12 CMM combinations. Then the obtained signaling pathways were compared with those KEGG signaling pathways related to mastitis found in MalaCards. Four common signaling pathways were identified: NF-κB signaling pathway, immune system, PI3K/Akt signaling pathway, and TNF signaling pathway. These four signaling pathways were the signaling pathways related to mastitis that CMMs involve in.

(1) NF-κB Signaling Pathway. Nuclear factor-κB(NF-κB) is a transcription factor widely existing in eukaryotic cells. NF-κB normally exists in nonactivated state in cells. When cells are stimulated by stimulating factors such as inflammatory mediators, viral infection, oxidative stress, etc., NF-κB will be activated and transfer to cell nucleus. It will then bind to the enhancer sites of target genes such as cytokines, growth factors, intercellular adhesion molecule, acute phase protein, etc. and enhance their transcription. Therefore, NF-κB signaling pathway plays a key role in regulating immune response, inflammatory response, cell proliferation/differentiation/apoptosis, etc. If the activation of NF-κB signaling pathway cannot be timely inhibited, various pathological responses may occur [13]. In recent years, the relationship between NF-κB signaling pathway and human diseases has received more and more attention. Research has shown that many CMMs which have significant efficacy in treating NF-κB-related diseases can inhibit the activity of NF-κB. By analyzing the working mechanism of CMM at cellular and molecular levels, it is found that CMM contain some active ingredients, which can regulate the activity of NF-κB at cellular or molecular levels and thus exert therapeutic effects. Among the CMM extracts that can significantly inhibit the activity of NF-κB, many are glycosides, including flavonoids, nonflavonoid polyphenols, and other glycosides [14]. Therefore, the mechanism of key CMM combinations in treating mastitis may be that they inhibit the activity of NF-κB signaling pathway, block the NF-κB-mediated expression of various cytokines, and thus exert therapeutic effects on mastitis.

(2) Immune System. The inappropriate activation of NF-κB signaling pathway can not only cause inflammatory response, but also decrease the body's immunity [15]. When the body's immunity is low, infection may become severer. Therefore, in terms of mastitis, regulation of immune system is also a mechanism of drug treatment. Many CMMs show effects of immune enhancement. For example, astragalus root, radix ginseng, and tangshen can replenish qi and strengthen body resistance. Baikal skullcap root and amur cork-tree can remove heat and eliminate toxicity. Since NF-κB signaling pathway is closely related to immunomodulation, the key CMM combinations studied here may exert immunomodulatory

effects by regulating NF-κB signaling pathways and immune system signaling pathways.

(3) PI3K/Akt Signaling Pathways. Phosphoinositide 3-kinase (PI3K) and its downstream target Akt are important signaling molecules and key survival factors that control cell proliferation, apoptosis, and tumorigenesis [16]. Research has shown that the enhancement of PI3K/Akt signaling pathway is one of the causes of hormonal therapy resistance in breast cancer. The inhibitors of many molecules in this signaling pathway can inhibit the growth of breast cancer cells and induce the apoptosis of cancer cells; thus they are often used as important drugs for treating breast cancer [17]. In addition, PI3K/Akt may affect the expression of proinflammatory cytokines and participate in inflammatory response by regulating TLR4 and its downstream molecules in macrophage. The phosphorylation of Akt can promote the phosphorylation of inhibitory subunit alpha of NF-κB(IκB-α). In this way, IκB-α is separated from NF-κB, which is thus activated and enter cell nucleus. Then, it can induce the expression of many inflammatory factors such as IL-6, tumor necrosis factor-α (TNF-α), etc. and cause inflammatory response [18]. Many CMMs have regulatory effects on PI3K/Akt signaling pathway. For example, hyperoside can regulate PI3K/Akt signaling pathway, decrease the activity of TNF-α and IL-6, and mitigate inflammatory response [19]. Astragalus polysaccharide can significantly decrease the expression of ρ-Akt and PI3K in colonic mucosa and exert therapeutic effects on colitis [20]. The key CMM combinations studied here may reduce the expression of proinflammatory cytokines by regulating PI3K/Akt signaling pathways and then exert anti-inflammatory effects.

(4) TNF Signaling Pathway. The activation of NF-κB signaling pathway can promote the release of proinflammatory factors such as TNF-α and IL-6. This can enhance inflammatory response and cause pro-/anti-inflammation imbalance, finally leading to further enhancement of inflammatory response and immune disorders [21]. Inducible transcription factors NF-κB family are activated in response to various stimuli. The most characteristic inducers are TNF cytokine family [22]. TNF is a major mediator of apoptosis, inflammation, and immunity. The activation of TNF signaling pathway is related to a wide range of human diseases, including septicemia, diabetes, cancer, osteoporosis, multiple sclerosis, rheumatoid arthritis, and inflammatory bowel disease [21]. CMMs which can remove heat and promote blood circulation contain active ingredients, which can inhibit the secretion of TNF, decrease the activity of NF-κBand block the development of inflammation. Red sage, Chinese angelica, moutan bark, peony root, sanqi, etc. are frequently used CMMs that can promote blood circulation. These CMMs not only can promote blood circulation and resolve stasis, but also have anti-inflammatory effects [23,24]. The mechanism of key CMM combinations studied here in treating mastitis might be that they reduce the secretion of TNF by inhibiting TNF signaling pathway and block the development of inflammation.

Table 6. “Drug combination-target-disease” related to “breast” and “mastitis”.

No	key CMM combination	Target related diseases	Drug target
1	Taraxaci herba, semen citri reticulatae, Paeoniae Radix Alba and Glycyrrhizae Radix et Rhizoma	Inflammation Breast Cancer Inflammatory Diseases Breast Cancer (Hormone-Sensitive)	ADORA1;ADORA2A;PLA2G1B;PLD1; CYP19A1;ESR1;PGR;VDR; ADORA1;CNR1; HSD17B1;
2	Taraxaci herba, Salvia miltiorrhiza Bunge, Paeoniae Radix Alba and Glycyrrhizae Radix et Rhizoma	Breast Cancer Inflammation Inflammatory Diseases Breast Cancer (Hormone-Sensitive)	CYP19A1;ESR1;PGR;VDR; ADORA1;ADORA2A;PLA2G1B;PLD1; PIK3CD; HSD17B1;
3	Taraxaci herba, Radix Salviae miltiorrhizae, Paeoniae Radix Alba and Glycyrrhizae Radix et Rhizoma	Inflammation Breast Cancer Breast Cancer (Hormone-Sensitive)	ADK;ADORA1;ADORA2A;PLA2G1B;PLD1;PTGS2; CYP19A1;ESR1;PGR;PTGS2;VDR; HSD17B1;
4	Taraxaci herba, Semen Coicis and Vigna angularis	Breast cancer Inflammation	ESR1;PGR;VDR; PLD1;
5	Taraxaci herba, Semen Coicis and Vigna angularis	Inflammation Inflammatory diseases Breast cancer	ADK;PLA2G1B;PLD1;PTGS2; PTGS2; ESR1;PGR;PTGS2;VDR;
6	Taraxaci herba,herba houttuyniae and pericarpium trichosanthis kirilowii et multilobae	Inflammation Breast cancer Inflammatory diseases	ADK;ADORA1;PLA2G1B;PLD1;PTGS2; ESR1;PGR;PTGS2;VDR; PTGS2;
7	Taraxaci herba, trichosanthes kirilowii peel and Semen Coicis	Inflammation Breast cancer	ADORA1;PLD1; ESR1;PGR;VDR;
8	Taraxaci herba, trichosanthes kirilowii peel and Semen Coicis	Inflammation Breast cancer Breast cancer (hormone-sensitive) Inflammatory diseases	ADK;ADORA1;ADORA2A;PLA2G1B;PTGS2; CYP19A1;ESR1;PGR;PTGS2;VDR; HSD17B1; PIK3CD;PTGS2;
9	Paeoniae Radix Alba, Bupleuri Radix and Glycyrrhizae Radix et Rhizoma	Inflammation Breast cancer Breast cancer (hormone-sensitive) Inflammatory diseases	ADORA1;ADORA2A;PLA2G1B;PPARD;PPARG;PTGS2; AKT1;CYP19A1;ESR1;PGR;PTGS2;VDR; HSD17B1; PIK3CD;PTGS2;
10	semen citri reticulatae, pericarpium citri reticulatae viride and Bupleuri Radix	Inflammation Breast cancer Inflammatory diseases	ADORA1;ADORA2A;PLA2G1B;PPARD;PPARG;PTGS2; AKT1;CYP19A1;ESR1;PGR;PTGS2;VDR; PIK3CD;PTGS2;
11	Taraxaci herba and Viola philippica	Breast cancer Inflammation	ESR1;PGR;VDR; PLA2G1B;PLD1;
12	Taraxaci herba and Lamiophlomis rotata Kudo	Breast cancer Inflammation	ESR1;PGR;VDR; PLD1;

4 Discussion

The treatment of mastitis in western medicine mainly adopts beta-lactam antibiotics to sterilize and prevent infection. The mechanism of drug action is to kill bacteria by destroying the cell wall. In comparison, there are many kinds of Chinese materia medica for treating mastitis, and the efficacy of CMM prescriptions also includes clearing heat and detoxification, removing swelling from breast, soothing liver and regulating qi, etc. Therefore, the study on the mechanism of action of CMM prescriptions in the treatment of mastitis is more complicated. Thus there are few related reports, most of which are about the mechanism of action of single Chinese materia medica. Gao Ruifeng [25] reported that one of the main components of honeysuckle, chlorogenic acid, acts as an antimastitis mechanism by inhibiting the activation of TLR4 and NF-κB signaling pathways. In addition, it can bind and activate PPAR- γ so that TLR4 can downregulate expression and inhibit the activation of downstream NF-κB signaling pathway. Finally, the expression levels of genes and proteins of inflammatory factors such as TNF-α, IL-1β, and IL-6 were decreased. Zhao Yongwang [26] reported that the main ingredient of scutellaria baicalensis can stabilize the mast cell membrane and inhibit its degranulation to reduce the release of inflammatory mediators. In addition, it can regulate the secretion of TNF-α and IFN-γ compounds to maintain a certain level. It can not only participate in antibacterial immunity and prevent excessive inflammation of tissues, but also regulate cellular immunity and improve breast immunity.

In this paper, data mining method was used to statistically analyze the frequencies of CMMs in prescriptions for treating mastitis and find association rules. Key CMM combinations for treating mastitis were obtained and can provide useful information for clinical therapy of mastitis. Network pharmacology was employed to obtain the potential targets of high-frequency CMMs and the potential signaling pathways that key CMM combinations involve in. This provides a new method for the research of the mechanism of CMMs in treating mastitis. However, the results are only based on already-known chemical composition of CMMs, related targets, and signaling pathways. With the development of technology, new ingredients and targets will be found in CMMs and there will also be more disease-related information. This will help enrich the results of this paper. In addition, according to CMM combination-target-disease analysis results, many CMMs for treating mastitis are also related to the treatment of breast cancer. Therefore, our future research will focus on the difference between CMMs used for treating mastitis and breast cancer.

Due to the large number of ingredients in CMMs and the complicated interaction between CMM and human body, it remains difficult to elaborate the working mechanism of CMM. In fact, figuring out the working mechanism of CMM has become a bottleneck in the modernization and inter-nationalization of CMM. From the perspective of network pharmacology, medicine-target-disease analysis may provide an entry point and a new strategy for further investigation into the working mechanism of CMM prescriptions.

Data Availability

The data that support the findings of this study are openly available in https://www.malacards.org/pages/info, http://lsp .nwu.edu.cn/tcmsp.php, http://bidd.nus.edu.sg/group/tcm-site/default.aspx, and http://bionet.ncpsb.org/batman-tcm/.

Conflicts of Interest

There are no conflicts of interest.

Authors' Contributions

Diyao Wu wrote the manuscript and finished data mining research. Xinyou Zhang took charge of guiding the experiments and paper writing. Liping Liu and Yongkun Guo collected the information and preprocessed the data of research.

Acknowledgments

This study was supported by National Natural Science Foundation of China (Grant No. 81660727).

References

1. Y.J. Bao, H.F. C. A review of risk factors of mastitis during lactation. *Traditional Chinese Medicine* **2017**, *6*, 224–231.
2. L. Peng, L. Jian, T. Shihuan, *et al*. Development and application of traditional Chinese medicine inheritance support system. *Chinese Journal of Experimental Traditional Medical Formulae* **2012**, *18*, 1–4.
3. S.L. Yang, C.L. Han, S.Z. Li, A.P. Chen, Y. Liu, Y.L. Zhang. Research on Traditional Chinese Medicine Syndrome Factors in Cognitive Disorder After Apoplexy based on TCMISS. Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine 2016; pp. 1403–1405.
4. N. Rappaport, N. Nativ, G. Stelzer, et al. MalaCards: An integrated compendium for diseases and their annotation. *Database* **2013**, ArticleID bat018.
5. J. Ru, P. Li, J. Wangetal. TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. *Journal of Cheminformatics* **2014**, *6*, article13.
6. Z. Liu, F. Guo, Y. Wang, et al. BATMAN-TCM: a bioinformatics analysis tool for molecular mechanism of traditional Chinese medicine. *Scientific Reports* **2016**, *6*.
7. Y. Feng, C.X. Li. Review of intercellular adhesion molecule-1 and its clinical significance. *International Journal of Immunology* **2012**, *35*, 107–112.
8. A.M. Vignola, G. Bonsignore, L. Sienaetal. ICAM-1and alpha3beta1 expression by bronchial epithelial cells and their in vitro modulation by inflammatory and anti-inflammatory mediators. *Allergy* **2000**, *55*, 931–939.
9. Z.W. Zhao, J. Tao, Z.N. Zhang, et al. The Relationships of IL-1, IL-6 and IL-8 with gastric

cancer. *Progress in Modern Biomedicine* **2015**, *15*, 1589–1591.
10. H. Xia. Effect of LBP on the expression of fractalkine and signal transduction mechanism in ARDS. Jinan University, Shangdong, China, 2017.
11. H.X. Tang, Z.H. Zhang, Z.Y. Zhao, H.X. Lu. Research progress of lactoferrin as drug carriers. *Acta Pharmaceutica Sinica* **2015**, *50*, 675–681.
12. D.M. Hyams, A. Chan, C. DeOliveira, et al. Cediranibin combination with fulvestrant in hormone-sensitive metastatic breast cancer: A randomized Phase II study. *Investigational New Drugs* **2013**, *31*, 1345–1354.
13. X.M. Wang, J. Aiguo. NF-κB signaling pathway and inflammatory response. *Progress in Physiological Sciences* **2014**, *45*, 68–71.
14. W. Xiuyan, W. Wei, F. Si et al. Effects of Hua tan Fang on expressions of nuclear gene-κB and epoxy synthase-2 mRNA in rats with chronic bronchitisc. *China Journal of Traditional Chinese Medicine and Pharmacy* **2005**, *20*, 315-316.
15. S.C. Gupta, B.B. Aggarwal. Preface: NF-κB, immune system and chronic diseases: how are they linked? *Forum on Immunopathological Diseases and Therapeutics* **2013**, *4*, 5-6.
16. H.Q. Zhu, W.H. Zhou. Research progress of Akt in drug addiction mechanism. *Chinese Pharmacological Bulletin* **2017**, *33*, 1046–1050.
17. S. Shukla, G.T. MacLennan, D.J. Hartman, P. Fu, M.I. Resnick, S. Gupta. Activation of PI3K-Akt signaling pathway promotes prostate cancer cell invasion. *International Journal of Cancer* **2007**, *121*, 1424–1432.
18. Z. Huiyan, Y. Xiaoping. IKK/NF-κB signaling pathway and central nervous system diseases. *Journal of Apoplexy and Nervous Disease* **2016**, *33*, 1143–1145.
19. J. Han, J.L. Xuan, H.R. Hu, Z.W. Chen. Protective effect against myocardial ischemia reperfusion injuries induced by hyperoside preconditioning and its relationship with PI3K/Akt signaling pathway in rats. *China Journal of Chinese Materia Medica* **2015**, *40*, 118–123.
20. Z. Haimei, H. Minfang, L. Duanyong, et al. Regulatory effects of astragalus polysaccharide on PI3K/Akt signals in colonic mucosa of rats with acute ulcerative colitis. *Chinese Traditional Patent Medicine* **2015**, *37*, 2029–2031.
21. M.S. Hayden, S. Ghosh. Regulation of NF-κBby TNF family cytokines. *Seminars in Immunology* **2014**, *26*, 253–266.
22. G. Chenand, D.V. Goeddel. TNF-R1 signaling: a beautiful pathway. *Science* **2002**, *296*, 1634-1635.
23. P. Huaxin, W. Peixun, Z. Lian, et al. Influence of effective ingredient of clearing heat and activating blood circulation kinds of traditional Chinese medicine on formation of THP-1 macrophages foam cells and secreting IL-1β and TNF-α. *Modern Journal of Integrated Traditional Chinese and Western Medicine* **2009**, *18*, 4325–4327.
24. Z. Hong, L. Cuilan. Traditional Chinese medicines that can promote blood circulation, remove blood stasis and have anti-inflammatory effect. *Journal of Modern Clinical Medicine* **2006**, 47-48.
25. G. Ruifeng. Study on the Anti-Mastitis Effect and Mechanism of Chlorogenic Acid. Jilin University, Jilin, China, 2014.
26. Z. Yongwang. Study on the Pathogenesis of Mastitis and the Therapeutic Effect of Baicalin. Hebei Agricultural University, Hebei, China, 2006.

System Design and Application of Data Mining in Chinese Materia Medica Based on Strategy Pattern

WU Di-yao, ZHANG Xin-you*
College of Pharmacy, Jiangxi University of Traditional Chinese Medicine, Nanchang 330004, China

Abstract: Objective: The aim is to design a data mining system based on strategy pattern and apply it to the data mining of Chinese materia medica. Method: Firstly, the system development was carried out on the platform of Matlab2014a. The operating system was Windows10.And it's integrated with graphical user interface (GUI) design. Association rules, entropy clustering of complex system and frequency analysis were encapsulated in the system. Secondly, three data sets were collected for obesity prescriptions, hyperlipidemia prescriptions and mastitis prescriptions respectively. Then data analysis was conducted on the three data sets by using the data mining system based on strategy pattern. Lastly, we compared the results and the characteristics of each algorithm. Results: Although the text content of obesity prescriptions data set and hyperlipidemia prescriptions data set were similar, the core drug combinations were different. The core drug combinations obtained by mining the mastitis prescription data set with different algorithms are also significantly different. The association rules are more suitable for explaining the principles of dialectical therapy with syndrome differentiation in traditional Chinese medicine while entropy clustering of complex system is more suitable for new prescriptions discovering. Conclusion: The data mining system based on strategy pattern can realize the data mining and analysis of Chinese materia medica more intelligently and user-friendly. Which algorithm the system performs depends on the characteristics of the database or user requirement. Additionally, user can also compare all the results calculated by different algorithms.
Keywords: Strategy pattern; Chinese Materia medica; data mining; system design

This study was supported by National Natural Science Foundation of China (Grant No. 81660727)

1 Background

After thousands of years of development, traditional Chinese medicine (TCM) has accumulated a wealth of theoretical knowledge and a lot of books and records, more and more Chinese medicine databases are created. Using data mining technology to mine the unknown knowledge and rules and put forward assumptions for experiment and theory can be a good auxiliary research of Chinese Materia medica (CMM)[1]. CMM is becoming increasingly important in modern health care, with the potential for new or improved clinical protocols

and reduction in treatment costs [2]. Clinical and research practices as well as ancient books and literatures provide a large amount of CMM data including theories, methods, formulas, drugs, etc. However, questions emerge such as: How to use these CMM data effectively? How to find underlying relationships among them? How to discover the potential knowledge hidden behind them? These are important questions that need to be answered in order to improve the use value of CMM.

In this paper, we developed a new way of mining CMM data which was based on strategy pattern. Strategy pattern defines a family of algorithms, encapsulates each one and makes them interchangeable. This pattern enables the use and interchange of different algorithms or implementations for a certain policy[3]. Data mining based on strategy pattern technique in CMM is different from onefold data mining method which is adopted conventionally in the field. This new method can solve problems that onefold data mining method cannot solve. Thus, it significantly improves the outcome of data mining in CMM. Strategy pattern enables one-to-many, many-to-one and many-to-many resolutions between different CMM problems and data mining methods, which expands the ways of mining data in CMM. Though several related researches have been done for data mining platform based on strategy pattern in Chinese medicine, all of them focus on the design of platform and construction of modules [4-5]. Furthermore, CMM data mining is hardly involved in these existing studies. This paper gives a full account of the latest achievements of studies on CMM data mining based on strategy pattern, which built on the basis of system design and research of CMM data mining[6]. The study might provide insights into the development and application of CMM.

2 Theoretical basis

2.1 CMM data mining method

With the technology of pattern recognition and statistical and mathematical techniques sieved across the stored information, data mining helps researchers recognize important facts, relationships, trends an patterns [7]. It aims to help decision makers to find the correlation between data and other information that cannot be easily found by humans themselves [8-9]. Different from traditional information processing method, most of the data mining methods are based on machine learning, pattern recognition, statistics, etc[10]. Besides, they can be used to analyze data in linear and non-linear manners and further discover potentially useful knowledge. Data mining techniques find applications in almost all the fields of CMM research, especially prescription compatibility and quality assessment of CMM. The frequently used data mining methods include association rules, regression analysis, decision tree method, cluster analysis, frequency analysis, artificial neural networks, etc. Data mining methods have evident advantages over traditional information processing method in processing vague and non-linear CMM data [11]. By using them, we can effectively carry out in-depth study on CMM. For example, using data mining technique, we can deeply explore the theories developed by many of the pharmacists, which will promote a unified understanding of CMM theories and practices. Using Bayesian neural networks, we can predict the relationship

between drug and adverse reaction. That would be significant to detect the signal of adverse reaction in early stage[12].Using cluster analysis, we can classify Chinese herb and predict their properties according to the four natures (i.e., cold, hot, warm and cool) and the five flavors (i.e., pungent, sweet, sour, bitter and salty) of them. Moreover, some of the unclassified Chinese herb can be classified by using association rules to recognize their medicinal characteristics.

In this paper, we investigated the application of data mining methods in CMM research. Using "data mining method" and "Chinese Materia medica" as search terms, we searched papers published during the past ten years in China National Knowledge Infrastructure (CNKI), which acts as the database. Papers that are not related to the topic were eliminated and a total of 103 papers were obtained as the final research samples. During the whole process, we tried to make sure that the papers were chosen as randomly as possible so that the data would be more reliable. Finally, the papers were statistically analyzed. Li et al [13] used decision tree to analyze the compatibility of herbal formulae in CMM. Yu et al [14] used association rules to study the medicinal characteristics and channel tropism of Chinese herbs, and found close associations between warm and pungent, cold and bitter, neutral and sweet, cool and sweet, etc. Shang et al[15] used association rules to investigate what was prohibited for clinical formula of Chinese herbs, and they found that some combinations of properties including hot-lung, hot-bitter, hot-stomach, etc. often have low frequency of occurrence and are usually prohibited for clinical formula of Chinese herbs. Data mining methods and literatures related to various attributes of Chinese herbs are shown in Table 1. Numbers represent the amount of associated literatures.

Table 1. Data mining methods and literatures related to various attributes of Chinese herb.

Data mining	Medicinal characteristics	Channel tropism	Efficacy	Quality evaluation	Toxicity	Compatibility	Pharmacological effect	Formula
Association	26	14	15	1	3	12	3	2
Frequency	6	9	12	3	2	13	4	13
Cluster	4	7	8	0	3	4	11	9
Decision	7	7	8	0	1	3	4	3
Principal	3	2	7	9	1	0	6	2
Artificial	9	7	3	11	2	1	12	3
Rough set method	2	6	7	0	5	5	1	3

On the basis of the results in Table 1, we summarized data mining methods and the fields of CMM research that they can be applied in (Table 2).

Although we mainly discussed the advantages of data mining technique, it still has some limitations. Generally, each type of data mining method is effective only in solving one or several problems. Moreover, the results given by only one data mining method are not very reliable. Also, these results are presented in fixed forms, some of which cannot be easily understood. Therefore, it is necessary to employ different data mining methods to solve one

problem. By doing this, the analysis results can be presented in an understandable form and are more reliable. In most cases, however, a problem is often solved by only onefold data mining method, so there is often a lack of comparison among different methods for solving the same problem and a deep analysis of it as well. Moreover, a data mining method is often not suitable to solve various problems, so the onefold data mining method has limited application. Besides, using onefold data mining method is relatively inefficient. Considering these facts, we proposed data mining method based on strategy pattern.

Table 2. Data mining methods and the fields of CMM research that they can be applied in.

Data mining methods	Association rules	Frequency analysis	Cluster analysis	Decisiontree	Principal component analysis	Artificial neural network	Rough set method
Fields of TCM	Medicinal characteristics, channel tropism, efficacy, compatibility	Formula, compatibility, efficacy, channel tropism	Pharmacological effect, channel tropism, efficacy, formula	Efficacy, medicinal characteristics, channel tropism	Efficacy, quality evaluation, pharmacological effect	Pharmacological effect, quality evaluation, channel tropism, medicinal characteristics	Efficacy, channel tropism, toxicity, compatibility

2.2 Strategy pattern

In China, many data mining methods have been applied to CMM research, but they are limited to specific problems. In other words, there is one-to-one relationship between data mining method and specific problem in CMM research (Figure 1). Although these data mining methods are effective in solving specific problems in CMM research, CMM data are not mined and used fully, and more valuable CMM data have not yet been obtained.

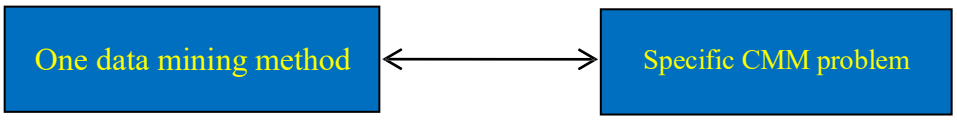


Figure 1. One-to-one relationship between data mining method and CMM problem.

Therefore, we applied strategy pattern to data mining method for CMM research,and then we constructed and achieved one-to-many and many-to-many relationship between different problems in CMM research and data mining methods on the basis of strategy pattern(Figure 2). We encapsulated CMM problems (data or information) and data mining methods (algorithm), respectively. Then, different methods can be used to solve one problem and we can compare among corresponding results by using different methods for solving the same problem. Furthermore, various problems can be solved by one method.so its application is expanded greatly.

Strategy pattern is a kind of behavioral pattern and its core idea is the encapsulation of algorithm. During the encapsulation process, the responsibility of algorithm is separated from the algorithm itself, and they are managed by different objects[16]. The strategy pattern often encapsulates a series of algorithms in a series of strategy classes, which acts as a subclass of abstract strategy class. In other words, the strategy pattern defines a family of algorithms,

encapsulates each algorithm, and makes the algorithms interchangeable within that family[17].

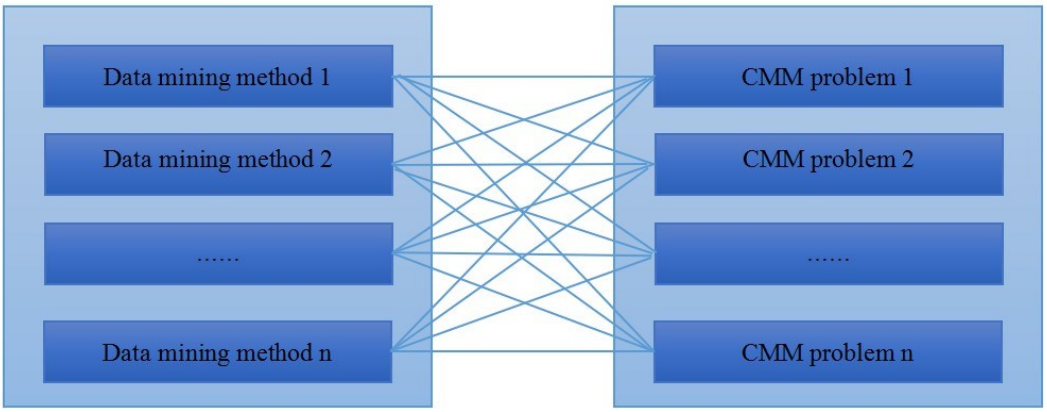


Figure 2. one-to many and many-to-many relationships between data mining methods and CMM problems.

On the basis of strategy pattern, we investigated the application of a combination of data mining methods. Specifically, we employed various data mining methods at the same time to discover potential knowledge hidden behind the data of Chinese herbal compound. Bayesian network and support vector machine were used to identify the common features of Chinese herbal compound. Rough set method was adopted to extract the characteristics of Chinese herbal compound. Association rules and decision tree method were employed to analyze the compatibility pattern of Chinese medicinal formulae. Cluster analysis and association rules can be used together to classify the efficacy of Chinese herbs and explore the association pattern between the efficacy and medicinal characteristics of Chinese herbs. Rough set method can be used to simplify the medicinal characteristics of Chinese herbs based on CMM classification. Association rules can be used to find rules that cannot be otherwise found by statistical methods and traditional artificial intelligence, so they are very important in CMM research. In fact, research mode based on association rules is the first choice for analyzing the underlying relationship among Chinese medicinal formulae. On this basis, decision tree can be used to represent results graphically. To some extent, these are advantages that onefold data mining method does not have.

3 Data mining system based on strategy pattern (DMSSP)

3.1 Software used for CMM data mining

The development of DMSSP was carried out on the platform of Matlab2014a. The operating system was Windows10.And it's integrated with graphical user interface (GUI) design.

3.2 Data mining methods

3.2.1 Association rules

Association rule is mainly used to discover the meaningful relationships hidden in large data sets through association rule analysis. It contains two stages: the first stage is to find all frequent item sets from the data set, and the second stage is to generate association rules from frequent item sets.

3.2.2 Entropy clustering of complex systems

Entropy clustering of complex system are based on the theory of information entropy which was brought up by Claude Shannon. The entropy clustering would determine whether they are positive correlation by calculating correlation coefficient between the variables and the others. If the three variables are positive correlation between any two variables, then these three variables are gathered into a heap. By that analogy, we could obtain a series of related core CMM combination [18].

3.2.3 Frequency analysis

Frequency analysis means that the total data are grouped according to some standards and the number of individuals in each group is counted. The drug that appears most frequently in the data set can be considered as the core drug of the data set

3.3 Design of DMSSP

Then data mining methods were sorted and classified. Association rules, entropy clustering of complex system and frequency analysis were encapsulated in the system.A mode was designed so that we can add, delete or modify CMM data mining methods/algorithms and achieve one-to-many or many-to-many resolutions between data mining methods and CMM problems. The interface can be switched to different methods or problems.

4 System application

Data mining technology is one of the most important means to conduct TCM research[19].We use DMSSP to analyze the data of CMM prescriptions to reveal the compatibility rules and explore the core drugs. As we mentioned above, we can construct one-to-many and many-to-many relationship between different CMM problems and data mining methods on the basis of strategy pattern. Therefore, we proceeded two experiments in this research. The one was to adopt the many-to-many mode which means using different algorithms to solve different CMM problems. The other one was to adopt the one-to-many mode which means using different algorithms to solve one CMM problem.

4.1 Data collection

We used "CMM" and "obesity" as retrieval words to search the related literatures from Pubmed (http://www.ncbi.nlm.nih.gov) and CNKI (http://cnki.net/).Then we collected 21 CMM prescriptions which are used in obesity treatment from these literatures. Likewise, we used "CMM" and "hyperlipidemia" as retrieval words to search literatures by the same

method mentioned above and finally collected 29 CMM prescriptions which are used in hyperlipidemia treatment. The two data sets are named as DPTO (database of prescriptions for treating obesity) and DPTH (database of prescriptions for treating hyperlipemia). They were prepared to mine the core CMM combinations based on many-to-many mode.

DPTM consists of two kinds of prescriptions: prescriptions used in clinical practice for treating mastitis and prescriptions from related literatures for treating mastitis. First, from January 2018 to May 2018, all prescriptions for treating diseases in both outpatients and inpatients in a provincial-level breast specialist hospital were collected. Among them, prescriptions for treating mastitis were screened out. After the same prescriptions were excluded, a total of 98 prescriptions for treating mastitis were collected. Second, we searched for papers in Pubmed (http://www.ncbi.nlm.nih.gov) and CNKI (http://cnki.net/) with "mastitis" and "Chinese materia medica" as two keywords. Then, a total of 45 prescriptions for treating mastitis were collected from the papers. The above prescriptions were combined and after the same prescriptions were excluded, a total of 131 prescriptions were collected, on which basis DPTM was constructed. It's prepared to mine the core CMM combinations based one-to-many mode.

4.2 System running

After imported 3 databases, we set about mining the core CMM combinations in each database. We chose association rule and frequency analysis to achieve the data mining of DPTO and DPTH. Then we chose association rule and entropy clustering of complex system to achieve the data mining of DPTM

4.3 The results and analysis

4.3.1 The core CMM combinations in DPTO and DPTH

The core CMM combinations in DPTO and DPTH by using association rule and frequency analysis methods are shown in Table 3.

In clinical practice, we found the composition of Chinese medicine prescription which is used for the treatment of hyperlipidemia is much similar to the prescription which is used for the treatment of obesity. Then we adopted data mining technique to analyze the medication rules of two diseases and excavate the core CMM combinations in the formula. The results of that can reveal the significance of prescriptions and explore the optimal composition of the medical effect of Chinese medicine prescription. The prescriptions which contain in Radix polygoni multiflori, fructus crataegi and salvia miltiorrhiza has the effects of promoting blood circulation and removing blood stasis, nourishing the kidney and reducing fat to lower blood lipids. The prescriptions which contain poria cocos, alisma orientalis, and atractylodes can strengthen the spleen and reduce swelling, strengthen the spleen and stomach to treat.Depending on the patient's disease type, the prescription can be increased or decreased based on the core CMM combination. This result is of guiding significance for clinical treatment of these two diseases.

Table 3. The core CMM combinations in DPTO and DPTH by using different mining methods.

Data mining method	DPTO	DPTH
Association rule	Polygonum multiflorum, the root of red-rooted salvia, fructus crataegi, rhizoma alismatis	Poria cocos, Rhizoma Atractylodis Macrocephalae, rhizoma alismatis, fructus crataegi
Frequency analysis	Astragalus membranaceus, the root of red-rooted salvia, fructus crataegi, rhizoma alismatis	rheum officinale, fructus crataegi, rhizoma alismatis

4.3.2 The core CMM combinations in DPTM

Firstly, the entropy clustering algorithm of complex system was selected. The correlation degree was set to 6 and the punishment degree was set to 4. The core CMM combination in DPTM was shown in Table 4.

Secondly, the association rule method was proceeded under the support which was greater than or equal to 23 and the confidence which was greater than or equal to 0.9.The core CMM combination in DPTM was shown in Table 5.

Table 4. The core CMM combination in DPTM by using entropy clustering of complex system.

Prescription No.	Core CMM combinations in DPTM by using entropy clustering of complex system
1	Green Tangerine Peel, stigma of corn, Tetrapanax papyriferus, Poria cocos, Astragalus, membranaceus, Rhizoma Atractylodis Macrocephalae
2	curcuma zedoary, lamiophlomis rotate, Chinese violet, loofah sponge
3	oyster, liquorice, fructus rosae laevigatae, fructus crataegi
4	Angelica sinensis, dandelion, Ligusticum wallichii, radix angelicae

Table 5. The core CMM combination in DPTM by using association rule.

Prescription No.	Core CMM combinations in DPTM by using association rule
1	dandelion, red sage, debark peony root, liquorice root
2	dandelion, tangerine seed, debark peony root, liquorice root
3	dandelion, cordate houttuynia, debark peony root, liquorice root
4	radix bupleuri, tangerine seed, debark peony root, liquorice root
5	radix bupleuri, Green Tangerine Peel, debark peony root, liquorice root

The core CMM combinations for the treatment of mastitis based on the two methods are quite different.The CMM combination obtained by association rule algorithm is a juxtaposition combination repeatedly occurring in the prescription database, which can directly reflect the overall principle of dialectical therapy with syndrome differentiation in traditional Chinese medicine.While,the entropy clustering algorithm of complex system tends to discover the potential connection between items. The core CMM combinations mined out is not shown repeatedly in the database, but they are the closest related CMM combination which is inferred by calculating the entropy value.It is more suitable for the research and development of new drugs or new prescriptions.

5 Discussion

With increasing research on CMM information, traditional methods already cannot meet the requirements of the development of modern Chinese traditional medicine. Many researchers are attempting to find a new way of studying CMM information. Data mining method based on strategy pattern can also find applications in research on Chinese medicinal formulae. Different formulae problems can be solved by one kind of data mining method and different data mining methods can be used to solve one formulae problem. Thus, we are able to comprehensively study Chinese medicinal formulae and find underlying rules. On this basis, we can reveal the compatibility of Chinese medicinal formulae, the relationship between the flavour and medicinal characteristics of CMM and the relationship between formulae, which deepen our understanding of disease and its treatment. In addition, CMM data mining based on strategy pattern can find important applications in the study and development of new drugs, research on Chinese herbs fingerprints, clinical fields, pharmacological studies, research on the medicinal characteristics of CMM, etc., so it provides a new way of developing and using CMM.

In this paper, we proposed to apply strategy pattern in CMM data mining and put this proposal into practice. Data mining based on strategy pattern is an effective way of managing and using CMM data. Besides, it maximizes the using of data mining algorithm in related area. Strategy pattern enables deep and extensive CMM data mining, which can provide more information for the industrialization and modernization of CMM.

References

1. Z. Wang, K.K. She. The application of data mining technology in the data analysis of traditional Chinese medicine. *Applied Mechanics & Materials* **2014**, 687-691, 1266-1269.
2. H.Y. Tse, V.W. Li, Hui M.N., et al. Data mining for Chinese materia medica and pharmacological research. *Journal of Biomolecular Screening* **2008**, *13*, 390
3. A. Christopoulou, E.A. Giakoumakis, V.E. Zafeiris, et al. Automated refactoring to the Strategy design pattern. *Information and Software Technology* **2012**, *54*, 1202-1214
4. L.Y. Zhang, L. Lv. Traditional Chinese medicine data mining platform based on strategy pattern. *Journal of Computer System Application* **2010**, *19*, 5-7.
5. Zhang L Y, Lv L. Platform Design and Research for data mining in traditional Chinese medicine based on strategy pattern. *Journal of Small Microcomputer System* **2011**, *32*, 1406-1411.
6. S.S. Luo, X.Y. Zhang, C.Q. Zhang. Research and system design of Chinese medicine data mining based on strategy pattern. *Journal of World Science and Technology: Modernization of Traditional Chinese Medicine* **2015**, *5*, 929-933.
7. E. Pastuchová, Š. Václavíková. Cluster analysis–data mining technique for discovering natural groupings in the data. *Journal of Electrical Engineering* doi:10.2478/jee-2013-0019:128-131.
8. J.W. Han, K. Micheline. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2001.

9. D. Cheung, H.Y. Hwang. Efficient rule-based attribute-oriented induction for data mining. *Journal of Intelligent Information System* **2000**, *15*, 175-200.
10. T.T. Chen, Y.F. Liu. Analysis of medication regularity of ancient Chinese medical records for treating depression based on data mining technology (in Chinese). *Journal of Acta Chinese Medicine and Pharmacology* **2016**, *3*, 16-20.
11. X. Wang. Research on medical information integration based on data mining technique (in Chinese). *Journal of Qiqihar Medical College* **2010**, *6*, 911-912.
12. A. Bate, M. Lindquist, I.R. Edwards, et al. A Bayesian neural network for adverse reaction signal generation. *European Journal of Clinical Pharmacology* **1998**, *54*, 315.
13. J.H. Li, Y. Feng, T. Yu. A data mining system for herbal formula compatibility analysis. *Lecture Notes in Electrical Engineering* **2014**, *269*, 1821-1828
14. H.Y. Yu, C.G. Xu. Application of association rule mining in exploration of relationship between medicinal properties and other attributes of traditional Chinese medicine (in Chinese). *China Journal of Experimental Traditional Medical Formulae* **2013**, *19*, 343-346.
15. E.X. Shang, X.S. Fan, J.A. Duan, et al. Data mining study on incompatibility characters of Chinese herbal medicine in accordance with association rules (in Chinese). *Journal of Nanjing University of traditional Chinese medicine* **2010**, *26*, 421-424.
16. E. Gamma, R. Helm, R.E. Johnson, et al. Design Patterns: Elements of Reusable Object-Oriented Software. Addison Wesley Longman: New Jersey, USA, 1995.
17. Y. Ma, Z.B. Zhao. Design and implementation of data roughening algorithm library based on strategy pattern. *Journal of Software Guide* **2009**, *12*, 4-46.
18. S.H. Tang, J.X. Chen, H.J. Yang, et al. New Chinese medicine prescription discovery based on complex systematic entropy clustering method. *World Science and Technology: Modernization of Traditional Chinese Medicine* **2009**, *11*, 225-228.
19. J.Y. Ren, M. Xu, Y.H. Lu. Research progress of clinical medication law and syndrome type of traditional Chinese medicine based on data mining. *Chinese Journal of Traditional Chinese Medicine and Pharmacy* **2017**, *10*, 4579-4582.

Analysis of Compatibility Law in The Prescriptions for Treating Kidney Deficiency by Using the WD-Get Rules Algorithm

Di-yao Wu, Xin-you Zhang*, Xiao-ling Zhou, Yong-kun Guo, Si-ting Yu
(*School of Pharmacy in Jiangxi University of Traditional Chinese Medicine, Nanchang, China, 430000*)

Abstract Objective The objective is to study frequently used traditional Chinese medicines (TCMs) in the prescriptions for treating kidney deficiency, analyze their medicinal properties and reveal their compatibility rules, thus providing useful information for the treatment of kidney deficiency. **Methods** "Traditional Chinese Medical Database System", "CNKI" and "Medicine Intelligence Database" were searched for prescriptions for treating kidney deficiency and then a prescription database was constructed. Excel was used to statistically analyze the main sources of the prescriptions, the frequency of each TCM appearing in the prescriptions and key TCM pairs. Subsequently, the prescription database was simulated by Matlab and WD-Get Rules algorithm integrating both width search and depth search was used to mine association rules. The confidence was set to ≥ 0.5 and the support was set to ≥ 0.1. **Results** All 192 prescriptions in the prescription database came from 88 medical books, such as *Sheng Ji Zong Lu* and *Pu Ji Fang*. The 192 prescriptions involved 343 TCMs. Among them, 123 TCMs each appeared no less than 3 times in the prescription database. The TCMs mainly have warm, neutral and cold natures, and sweet and pungent flavors. Channel tropism analysis reveals that they mainly act on kidney and spleen. There were 28 pairs of TCMs each of which appeared more than 20 times in the prescription database and 15 strong association rules were mined out. **Conclusions** Poria cocos, prepared rehmannia root, common yam rhizome, eucommia bark, *etc.* are commonly used TCMs in the treatment of kidney deficiency. Since poria cocos can nourish and coordinate yin and yang, invigorate spleen and eliminate dampness, it can be used in pair with many other TCMs. Through the combined use of TCMs, it is possible to achieve the combination of their efficacies such as nourishing kidney yin, warming kidney yang, replenishing kidney qi, enriching kidney essence, tonifying spleen and kidney, nourishing yin to tonify yang, replenishing blood and promoting blood circulation in the treatment of kidney deficiency.
Keywords: kidney deficiency; prescription compatibility law; association rules; data mining.

1 Introduction

Traditional Chinese medicine (TCM) information involves massive and complex data systems [1]. Data mining is a powerful tool for processing TCM information as it can analyze and sort massive data and achieve the rational utilization of effective information [2]. This

technology can be applied to almost all aspects of TCM research, in which prescription compatibility law is an especially hot topic. From the perspectives of frequency of medication use, combination of efficacies, *etc.*, data mining provides a reliable method for scientifically and efficiently analyzing the underlying compatibility law in target prescription data. The results can provide important guidance for clinical medication and drug discovery.

In the field of TCM, kidney deficiency is defined as kidney essence-qi-yin-yang deficiency and can be classified into four types: kidney essence deficiency, kidney qi deficiency, kidney yin deficiency and kidney yang deficiency [3]. In TCM theory, kidney is considered as the most fundamental organ in human body and the root of twelve meridians, and it is the "inborn foundation". According to the theory of state of viscera in kidney, kidney is thought to store essence, govern water metabolism, govern bones, govern inspiration, and govern urination and defecation [4]. If a person suffers from kidney deficiency, all above-mentioned functions will be affected and yin-yang imbalance may occur in his/her five zang viscera and six fu viscera, thus threatening his/her health. Modern medical research has proved that kidney deficiency can lead to neuroendocrine dysfunction, reproductive dysfunction, hearing dysfunction and bone metabolism dysfunction, as well as can reduce human immunity. In recent years, more and more people suffer from kidney deficiency due to unhealthy lifestyles and some environmental factors [5]. TCM treatment of kidney deficiency has a long history. Rich experiences have been accumulated and good results have been achieved. In this paper, data mining technology was used to analyze the TCM prescriptions described in medical books that can be used for treating kidney deficiency. We aimed to identify the prescription compatibility law and provide useful information for clinical treatment of kidney deficiency.

2 Data sources and methods

2.1 Prescription collection

"Traditional Chinese Medical Database System", "CNKI" and "Medicine Intelligence Database" were searched for prescriptions for treating kidney deficiency. The prescriptions that involve only one medicine were excluded. A total of 192 effective prescriptions were collected. The names of the involved TCMs were standardized according to the *Pharmacopoeia of the People's Republic of China* (2015). The names of the medicines that were not included into the *Pharmacopoeia of the People's Republic of China* (2015), such as pig kidney, green salt, *etc.,* were kept. Subsequently, a database of prescriptions for treating kidney deficiency was constructed and then Matlab was used for simulation.

2.2 Improved association rule algorithm

2.2.1 Association rule mining

Association rule mining aims to find frequent itemsets in the database whose support is no lower than the user-specified minimum support and then use the frequent itemsets to generate association rules, among which strong association rules can be identified according to

the user-specified minimum confidence [6]. As an important branch of data mining algorithm, association rule algorithm has been used for more than ten years in TCM research since it was first proposed for use in this field in 2002 [7]. Because TCM data are massive, nonlinear and nonstructural [8], traditional data mining techniques can hardly be used for TCM data analysis. It is necessary to further improve the efficiency and rule interestingness of classic association rule mining algorithms and thus improve their speed and ability to process TCM data. In this paper, we designed an improved algorithm (called WD-Get Rules) integrating both width search and depth search. This algorithm has higher mining efficiency than traditional algorithms in processing relatively large-scale datasets. Therefore, the newly designed WD-Get Rules was used here for mining analysis of 192 prescriptions for treating kidney deficiency.

2.2.2 Improved algorithm integrating both width search and depth search (WD-Get Rules)

The idea of WD-Get Rules is that: width-first strategy is used first to find the candidate consequent set (H) of strong association rules. In this set, each consequent is a one-itemset. On the basis of H, association rule generation algorithm based on set enumeration tree (GRSET) was adopted to perform depth search. Notably, during depth search, the consequents of rules include only the elements in H, thus the elements that cannot become the consequents of rules are avoided. Besides, depth search can be performed only if itemset is frequent k-itemset (k > 2) and there are more than one element in H. Otherwise, the next frequent itemset is used.

Suppose that there is a frequent itemset $L = L_1 \cup L_2 \cup \quad \cup L_k$, where L_k is the collection of frequent k-itemsets and each frequent itemset is ordered. The WD-Get Rules algorithm is shown in Table 1.

Table 1. WD-Get Rules algorithm

```
Algorithm: WD-GetRules(L,min_conf,TIDcount,supData)
Input: frequent itemset (L), the minimum confidence threshold (min_conf), the total number of affairs (TIDcount), support count (supData)
Output: all association rules generated from L (F)

F is initially {'rule', 'support', 'confidence'}
Obtain the number of rows in L (m)
For i=2:m start from frequent 2-itemset
Obtain all frequent i-itemsets (L1)
Obtain the number of rows in L1 (m1)
For j=1:m1
    Obtain the jth subset in L1 (freqSet)
    %perform width search
    After initialization, the set of consequents that can be the consequents of strong association rules (H) is empty.
    Combine all items in freqSet into string (S)
    For r=1:i
        Obtain the rth element in freqSet (a)
        Use behind to save a in string format
        Save the difference set between freqSet and a in C
        Combine the sequence of elements in C to front
        Calculate rule confidence conf=sup(S)/sup(front)
        If conf>=min_conf
```

```
                Obtain rules f1= front=>behind
                Rule support sup= supData(S)/TIDcount
                Add f1, sup, conf to F F=[F;f1,sup,conf]
                Add a to H
            End
        End
        %width search is finished
        Obtain the number of elements in H (num)
        If i>2 && num>1
            %perform depth search
            For jj=1:num
                Obtain the jjth element in H (h)
                s=jj+1;
                If s<=num
                    F=getall_new(h, s, freqSet, min_conf, TIDcount, supData, i, F, S, H)
                End
            End
        Else
            Continue
        End
    End
    End
```

For getall_new algorithm, see Table 2.

Table 2. getall_new algorithm.

```
    Algorithm: getall_new(h, s, freqSet, min_conf, TIDcount, supData, k, F, S)
    Input: consequent set (h), the initial positions of items that need to be combined in freqSet (s, the position of the last item in h +1≤s≤k), freqSet, minimum confidence threshold (min_conf), the total number of affairs (TIDcount), support count (supData), the number of items in freqSet (k), rule set (F), string generated by combining all items in freqSet (S)
    Output: all association rules generated from freqSet and whose consequents include h.

    Obtain the number of elements in h (n)
    If n<k-1
        For j=s:k
            Obtain the jth element in freqSet (a)
            Calculate rule antecedent C=freqSet-h-a (- represent subtraction operation)
            Combine all items in C into rule antecedent string front
     Calculate rule confidence conf= supData(S)/supData(front)
            If conf>=min_conf
                Combine h+a into rule consequent string behind (+ represents union operation)
                Obtain rule f1= front=>behind
                Rule support sup= supData(S)/TIDcount
                Add f1, sup, conf to F F=[F;f1,sup,conf]
                If j<k
                    getall(h+a, j+1, freqSet, min_conf, TIDcount, supData, k, F, S)
                End
            End
        End
    End
```

3 Results

3.1 Analysis of the sources of prescriptions

The collected 192 prescriptions for treating kidney deficiency were obtained from 88 medical books. Among them, 18 books that each provide more than 3 prescriptions for treating kidney deficiency were selected as the important prescription sources (Table 3). Notably, *Sheng Ji Zong Lu (圣济总录)* and *Pu Ji Fang（普济方）* each provides 10 prescriptions for treating kidney deficiency. *Sheng Ji Zong Lu* is the earliest existing comprehensive medical book compiled by a team of scholars organized by the government in ancient China [9] and it has important historical values. *Pu Ji Fang* is the largest and most voluminous medicine prescription book in ancient China and listed as one of the important ancient medical books in China [10]. In this study, data mining technology was used to fully mine the TCM database and systematically explore the prescriptions formulated by TCM doctors in ancient China. The results may provide important guidance for clinical treatment of kidney deficiency.

Table 3. Important sources of prescriptions for treating kidney deficiency

Prescription source	Number of prescriptions
Sheng Ji Zong Lu	10
Pu Ji Fang	10
Yi Lve Liu Shu	9
Sheng Hui	8
National TCM prescriptions	8
Shou Shi Bao Yuan	6
Jing Yue Quan Shu	6
Yi Xue Ru Men	5
Yi Tong	4
Tai Ping Hui Min He Ji Ju Fang	4
San Yin	4
Bian Zheng Lu	4
Yin Shan Zheng Yao	3
Wai Tai	3
San Yin Ji Yi Bing Zheng Fang Lun	3
Min Jian Fang	3
Ji Feng	3
Gu Jin Yi Jian	3

3.2 Analysis of the frequencies of TCMs appearing in prescription database

The 192 prescriptions involved a total of 343 TCMs. The frequency of each TCM appearing in the prescriptions (use frequency) was analyzed. There were a total of 123 TCMs with use frequency ≥ 3 (Table 4).

Table 4. TCMs with use frequency ≥ 3.

TCMs	Frequency	TCMs	Frequency	TCMs	Frequency	TCMs	Frequency	TCMs	Frequency
Poria cocos	68	Debark peony root	18	Plantain seed	7	Platycodonroot	5	Tiger bone	3
Prepared rehmannia root	61	Pilose antler	18	Turtle shell	7	Sulfur	5	Amber	3
Commonyam rhizome	52	Fennel	18	Medicine terminaliafruit	7	Frankincense	5	Talc	3
Eucommiabark	51	Dwarf lilyturf tuber	17	Green salt	7	Villous amomum fruit	5	Fineleaf schizonepe ta herb	3
Two-toothe d achyranthesroot	49	Dendrobiu m	17	Largetrifol iolious bugbane rhizome	7	Fresh ginger	5	Red sage	3
Prepared common monkshood branched root	48	Amur cork-tree	15	Orangefruit	7	Medicinal evodia fruit	5	Largeleaf gentian root	3
Chinese angelica	48	Tree peony root bark	15	Gordoneuryale seed	7	Pinellia tuber	5	Feather cockscombseed	3
Radix ginseng	48	Sichuan lovage rhizome	14	Ass hide glue	6	Polygonum multiflorum	5	Scorpion	3
Liquoriceroot	46	Magnetite	14	Dahurian angelica root	6	Atractylodes rhizome	5	Chinese taxillus herb	3
Dodder seed	45	Incised notopterygium rhizome and root	12	Areca seed	6	Mountain spicy fruit	4	Flatstem milkvetchseed	3
Chinese magnoliavine fruit	42	Hot pepper	12	Chinese thorowax root	6	Chinese star anise	4	Commoncnidium fruit	3
Atractylode s macrocephala	41	Chinese eaglewood	11	Common fenugreek seed	6	Clove	4	Musk	3
Astragalusroot	37	Bone fossil of big mammals	11	Peach seed	6	Cibot rhizome	4	Combined spicebush root	3

Table 4. *Cont*.

TCMs	Frequency	TCMs	Frequency	TCMs	Frequency	TCMs	Frequency	TCMs	Frequency
Asiatic cornelian cherry fruit	37	Dried tangerine peel	10	Nutgrass galingale rhizome	6	Cablin patchouli herb	4	Poppy capsule	3
Malaytea scurfpea fruit	37	Common anemarrhe na rhizome	10	Sharp-leaf glangal fruit	6	Lotus seed	4	Cinnabar	3
Desertliving cistanche	30	Nutmeg	10	Chrysanthe mum flower	6	Degelatined deer-horn	4	Tatarian aster root	3
Cassia bark	29	Peony root	10	Common floweringq ince fruit	6	Myrrh	4	Cassia seed	3
Morindaroot	26	Cochinchinese asparagus root	9	Red tangerine peel	6	Mantis egg-case	4	Chinese honey locust	3
Barbary wolfberryfruit	25	Oystershell	9	Rhizoma acori calami	5	Tall gastrodia tuber	4	Pig kidney	3
Milkwortroot	21	Officinal magnolia bark	8	Green onion stalk	5	Deer-horn glue	4	-	-
Common aucklandiaroot	21	Puncturevine caltrop fruit	8	Tangshen	5	Fine salt	4	-	-
Unprocessed rehmannia root	21	Grassleaf sweetflagrhizome	8	Rehmannia Root	5	White salt	3	-	-
Himalayanteasel root	20	Manchurian wildginger	8	Doubleteeth Pubescent angelica root	5	Halloysite	3	-	-
Divaricate saposhnikovia root	19	Spine date seed	8	Baical skullcap root	5	English walnut Seed	3	-	-

Mining results of use frequency of TCMs show that the most commonly used TCMs are poria cocos, prepared rehmannia root, common yam rhizome, eucommia bark, prepared common monkshood branched root, *etc*. Poria cocos and prepared rehmannia root have neutral nature and sweet flavor. They can strengthen and coordinate yin and yang, replenish qi and enrich essence. Channel tropism reveals that they have effects on lung, spleen, kidney, *etc*. Therefore, these two medicines can not only be used alone to tonify spleen and strengthen kidney yang, but also be used in combination with other TCMs to treat various types of kidney deficiency. Since ancient times, prepared rehmannia root has been used to nourish yin and tonify kidney. Because blood and body fluid are the material basis of kidney yang and kidney qi, prepared rehmannia root is frequently used in combination with other TCMs by patients with kidney qi deficiency and kidney yang deficiency to nourish yin, replenish qi and strengthen yang [11]. Therefore, prepared rehmannia root occurrs in many prescriptions. Eucommia bark is a warm herb and prepared common monkshood branched root is a hot herb. They can both warm kidney and strengthen yang and are important herbs used for tonifying kidney and treating kidney yang deficiency.

In addition, some ancient medical books also suggest diet therapy, for example, eating sheep or pig kidney to tonify human kidney. In the field of TCM, “analogy” is used in diet therapy, namely, benefiting human organs by eating similarly-shaped food. Some examples can be found in ancient medical books. For instance, *Tiao Ji Yin Shi Bian*（*调疾饮食辩*）written by Mu Zhang of the Qing dynasty suggests that eating animal heart benefits human heart and eating animal kidney benefit human kidney [12]. Simiao Sun, a famous TCM doctor of the Tang dynasty, also proposed that eating animal organs can benefit corresponding human organs [13]. Although such analogy in diet therapy has not yet been proven clinically for its validity, it should still receive certain attention and be used in a scientific way in treating disease. Modern medical research demonstrates that kidney deficiency patients with severe clinical symptoms should not eat animal organs with high content of cholesterol [13]. The above-mentioned diet therapy has indeed been gradually abandoned in the treatment of kidney deficiency, which is reflected from the decreased frequency of sheep kidney and pig kidney used in kidney deficiency treatment. Therefore, the idea of "eating animal kidney benefits human kidney" should be applied dialectically to clinical practice.

3.3 Analysis of the natures, flavors and channel tropisms of TCMs in the prescription database

The information about the natures, flavors and channel tropisms of TCMs was obtained from the *Pharmacopoeia of the People's Republic of China* (2015). Since there are too many TCMs used in the treatment of kidney deficiency, we selected 20 TCMs with the highest use frequency for investigation into their natures, flavors and channel tropism (Table 5). In addition, 85 TCMs with high use frequency (≥5) were selected and statistically analyzed according to their natures, flavors and channel tropisms (Tables 6–8).

The results reveal that most of the TCMs used for treating kidney deficiency have warm, neutral and cold natures, and sweet and pungent flavors, and they mainly have effects on kidney and spleen meridians. According to the relationship between combination of

properties of TCMs and efficacy [14], warm-pungent-kidney (nature-flavor-channel tropism) TCMs can tonify kidney, strengthen yang, as well as strengthen bones and muscles; neutral-sweet-kidney TCMs can enrich essence, tonify kidney and tonify liver; cold-sweet-kidney TCMs can remove heat, promote the secretion of body fluid and moisten dryness; warm-sweet-spleen TCMs can invigorate spleen and replenish qi; neutral-liver-spleen TCMs can nourish spleen and replenish qi. According to the results of channel tropisms of TCMs, the majority of the TCMs used for treating kidney deficiency directly act on kidney or on both kidney and spleen. According to the results of natures and flavors of TCMs, pungent TCMs are used to nourish kidney, and sweet TCMs are used to replenish qi and benefit kidney. The combined use of these two kinds of TCMs can strengthen yang. Warm TCMs are used to promote the movement, warmth and qi transformation of the kidney. Cold TCMs can nourish kidney yin, remove ministerial fire, nourish the body, and reduce yang heat. Neutral TCMs can nourish both yin and yang and thus can be used to treat all types of kidney deficiency.

Table 5. The natures, flavors and channel tropisms of 20 TCMs with the highest use frequency.

TCMs	Use frequency	Nature	Flavor	Channel tropism
Poria cocos	68	Neutral	Sweet	Heart, lung, spleen and kidney meridians
Prepared rehmannia root	61	Slightlywarm	Sweet	Liver and kidney meridians
Common yam rhizome	52	Neutral	Sweet	Lung, spleen and kidney meridians
Eucommia bark	51	Warm	Sweet and slightly pungent	Liver and kidney meridians
Two-toothed achyranthes root	49	Neutral	Bitter, sweet and sour	Liver and kidney meridians
Radix ginseng	48	Warm and neutral	Sweet and slightly bitter	Spleen, lung and heart meridians
Chinese angelica	48	Warm	Sweet and pungent	Liver, heart and spleen meridians
Liquorice root	46	Warm	Sweet	Heart, lung, spleen and stomach
Dodder seed	45	Neutral	Pungent and sweet	Liver, kidney and spleen meridians
Prepared common monkshood branched root	45	Hot	Pungent and sweet	Heart, kidney and spleen meridians
Chinese magnoliavine fruit	42	Warm	Sour and sweet	Lung, heart and spleen meridians
Atractylodes macrocephala	41	Warm	Bitter and sweet	Spleen and stomach meridians
Astragalus root	37	Warm	Sweet	Lung, spleen, liver and kidney meridians
Asiatic cornelian cherry fruit	37	Slightlywarm	Sour	Liver and kidney meridians
Malaytea scurfpea fruit	37	Warm	Bitter, pungent and warm	Kidney and spleen meridians
Desertlivingcistanche	30	Warm	Sweet, sour and salty	Kidney and the large intestine
Cassia bark	29	Hot	Pungent and sweet	Kidney, liver, heart and liver meridian
Morinda root	26	Slightlywarm	Sweet and pungent	Kidney and liver meridians
Barbary wolfberry fruit	25	Neutral	Sweet	Liver and kidney meridians
Rhizoma alismatis	24	Cold	Sweet	Kidney and bladder meridians

Table 6. The proportions of TCMs with different natures.

Nature	TCMs	Use frequency	Proportion
Warm	44	796	54.48%
Cold	20	222	15.20%
Neutral	17	315	21.56%
Hot	5	112	7.67%
Cool	1	9	0.62%

Note that each TCM may have more than one nature and such TCM was repeatedly counted in terms of its different natures.

Table 7. The proportions of TCMs with different flavors.

Flavor	TCMs	Use frequency	Proportion
Sweet	44	1011	69.20%
Pungent	42	613	41.96%
Bitter	36	486	33.26%
Sour	11	217	14.85%
Salty	5	78	5.34%

Note that each TCM may have more than one flavor and such TCM was repeatedly counted in terms of its different flavors.

Table 8. The proportions of TCMs with different channel tropisms.

Channel tropism	TCMs	Use frequency	Percentage
Kidney	56	1316	78.20%
Spleen	39	992	59%
Liver	35	816	48.50%
Heart	28	651	38.70%
Lung	22	573	34.10%
Stomach	13	167	9.90%
Bladder	8	62	3.70%
The large intestine	7	59	3.50%

Note that each TCM may involve in more than one channel tropisms and such TCM was repeatedly counted in terms of its different channel tropisms.

3.4 Analysis of the composition principles of prescriptions for treating kidney deficiency

3.4.1 Analysis of key TCM pairs

In the database of prescriptions for treating kidney deficiency, a total of 28 TCM pairs with use frequency ≥ 20 were found. The efficacy of each TCM pair was summarized according to the *Pharmacopoeia of the People's Republic of China* (Table 9).

Table 9. Key TCM pairs for treating kidney deficiency, their use frequency and efficacy.

Key TCM pairs	Use frequency	Efficacy
Common yam rhizome, poria cocos	34	Replenish kidney qi
Prepared rehmannia root, poria cocos	32	Nourish kidney yin
Radix ginseng, poriacocos	27	Replenish kidney qi
Prepared rehmannia root, common yam rhizome	27	Nourish kidney yin, enrich kidney essence, and replenish kidney qi
Liquorice root, radix ginseng	26	Replenish qi and invigorate spleen
Two-toothed achyranthes root, poria cocos	25	Nourish kidney yin and clear deficiency fire
Prepared rehmannia root, Chinese magnoliavine fruit	25	Nourish kidney yin and replenish kidney qi
Two-toothed achyranthes root, prepared rehmannia root	24	Nourish kidney yin
Common yam rhizome, asiatic cornelian cherry fruit	24	Enrich kidney essence, replenish kidney qi and warm kidney yang
Asiatic cornelian cherry fruit, prepared rehmannia root	23	Enrich kidney essence, as well as nourish yin and yang
Chinese angelica, radix ginseng	23	Replenish qi and promote blood circulation
Chinese angelica, two-toothed achyranthes root	23	Nourish kidney yin, replenish blood, and promote blood circulation
Asiatic cornelian cherry fruit, poria cocos	22	Enrich kidney essence and warm kidney yang
Chinese magnoliavine fruit, poria cocos	22	Replenish kidney qi
Eucommia bark, prepared rehmannia root	22	Nourish yin and strengthen yang
Liquorice root, poria cocos	22	Replenish qi and invigorate spleen
Largehead atractylodes rhizome, poria cocos	21	Replenish qi, invigorate spleen and reduce swelling
Radix ginseng, prepared rehmannia root	21	Nourish kidney yin and replenish kidney qi
Eucommia bark, two-toothed achyranthes root	21	Replenish kidney qi, clear deficiency fire and strengthen bones and muscles
Chinese angelica, prepared rehmannia root	21	Nourish kidney yin, replenish blood and promote blood circulation
Chinese angelica, liquorice root	21	Invigorate spleen-stomach, replenish qi, replenish blood and promote blood circulation
Liquorice root, prepared rehmannia root	21	Nourish kidney yin and replenish kidney qi
Malaytea scurfpea fruit, eucommia bark	20	Warm kidney yang
Astragalus root, radix ginseng	20	Warm kidney yang and replenish kidney qi
Largehead atractylodes rhizome, liquorice root	20	Replenish qi and invigorate spleen
Chinese magnoliavine fruit, common yam rhizome	20	Enrich kidney essence and replenish kidney qi
Dodder seed, poria cocos	20	Warm kidney yang and enrich kidney essence
Two-toothed achyranthes root, common yam rhizome	20	Nourish kidney yin and replenish kidney qi

These key TCM pairs mainly have the following efficacies: nourishing kidney yin, warming kidney yang, replenishing kidney qi, enriching kidney essence, invigorating spleen and tonifying kidney, nourishing yin to strengthen yang, as well as replenishing qi and promoting blood circulation. The first four efficacies are corresponded to kidney yin deficiency, kidney yang deficiency, kidney qi deficiency and kidney essence deficiency patients, respectively. "Invigorating spleen and tonifying kidney" is based on the theory of correlation between spleen and kidney [15]. This theory suggests that there is a very close correlation between spleen and kidney, which is far closer than that between other organs in the body. In the theory of TCM, kidney is considered as the "inborn foundation" and mainly store essence, while spleen is considered as the "acquired foundation" and the source of qi and blood. TCM doctors also propose that inborn foundation generates acquired foundation and the latter is conducive to the former, which illustrates a close relationship between spleen and kidney from both physiological and pathological perspectives. Therefore, in clinical practice, patients are suggested to tonify their spleen in order to strengthen the kidney-tonifying efficacy or suggested to tonify both spleen and kidney so that both inborn and acquired foundations receive equal attention [16]

"Nourishing yin to strengthen yang" is an example of application of Yin and Yang theory. Jingyue Zhang, TCM doctor of the Ming dynasty, put forward that yin is rooted in yang and yang is also rooted in yin. He thought that yang qi and yin essence are one and come from the same source, and they cannot be separated. Yang qi provides impetus for life activities and it must take yin essence and blood as material basis [17]. In the treatment of kidney deficiency, if only sweet and cold TCMs are used to nourish yin and enrich essence, the transportation and transformation governed by spleen will sometimes be inhibited. Under this circumstance, essential qi can hardly be generated in kidney, and yin and cold pathogen may injury the body's primordial qi. Similarly, when only warm and hot TCMs are used to warm and tonify kidney yang, certain effects can be achieved, but long-term use can result in the loss of essence and blood and no source for kidney qi generation. Therefore, there are many prescriptions in the database for strengthening yang and at the same time nourishing yin so as to ensure the warmth but no dryness of yang TCMs and the nourishing effects but no greasy effects of yin TCMs.

"Replenishing qi and promoting blood circulation" is consistent with the idea of "tonifying kidney and promoting blood circulation" proposed by Prof. Daning Zhang [18]. Prof. Zhang put forward that kidney deficiency and blood stasis are not two independent events. He thought that kidney deficiency must lead to blood stasis, which in turn leads to severe kidney deficiency, thus a vicious circle occurs [19]. There are many key TCM pairs in the database that can replenish blood, promote blood circulation, remove kidney meridian stasis and strengthen vital qi to eliminate pathogenic factor.

3.4.2 Analysis of association rules

WD-Get Rules algorithm was used to perform association rules analysis of the database of prescriptions for treating kidney deficiency. Support was set to ≥ 0.1 and confidence was set

to ≥ 0.5. The mining results are shown in Table 10. Cytoscape was used to visualize the association rules of TCMs (Fig. 1).

Table 10. Results of association rules analysis of TCMs for treating kidney deficiency.

Rules	Support	Confidence
Asiatic cornelian cherry fruit=>common yam rhizome	0.125	0.64865
Asiatic cornelian cherry fruit=>prepared rehmannia root	0.11979	0.62162
Common yam rhizome=>poria cocos	0.17708	0.61818
Chinese magnoliavine fruit=>prepared rehmannia root	0.13021	0.59524
Asiatic cornelian cherry fruit=>poria cocos	0.11458	0.59459
Radix ginseng=>poria cocos	0.14063	0.5625
Prepared rehmannia root=>poria cocos	0.16667	0.54237
Radix ginseng=>liquorice root	0.13542	0.54167
Malaytea scurfpea fruit=>eucommia bark	0.10417	0.54054
Astragalus root=>radix ginseng	0.10417	0.52632
Chinese magnoliavine fruit=>poria cocos	0.11458	0.52381
Liquorice root=>radix ginseng	0.13542	0.52
Largehead atractylodes rhizome=>poria cocos	0.10938	0.5122
Poria cocos=>common yam rhizome	0.17708	0.50746
Two-toothed achyranthes root=>poria cocos	0.13021	0.5

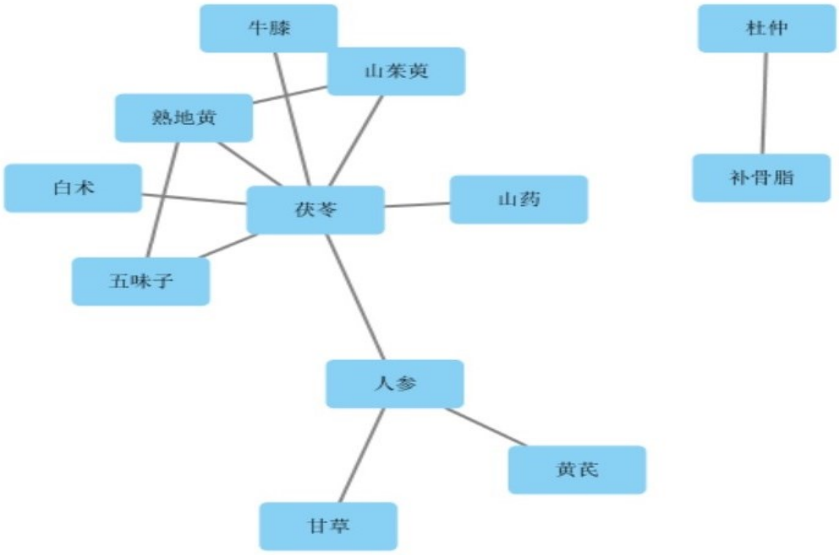


Figure 1. Results of association rules analysis of TCMs for treating kidney deficiency.

Poria cocos is at the core position of the database of prescriptions for treating kidney deficiency. Poria cocos itself does not tonify the kidney, but it can eliminate dampness, replenish qi and coordinate both yin and yang due to its plain, sweet flavor, and neutral nature [20]. It can be used in combination with many TCMs such as common yam rhizome, prepared rehmannia root, asiatic cornelian cherry fruit, Chinese magnoliavine fruit, two-toothed achyranthes root, radix ginseng, *etc.* to form key TCM pairs to tonify the kidney.

The combined use of poria cocos and common yam rhizome can invigorate spleen, nourish

stomach, tonify kidney and arrest seminal emission. This TCM pair is especially suitable for patients with spleen and kidney qi deficiency. The combined use of poria cocos and asiatic cornelian cherry fruit can warm and nourish liver and kidney, arrest seminal emission and relieve prostration. This TCM pair is suitable for patients with kidney essence deficiency and kidney yang deficiency. Two-toothed achyranthes root can tonify liver and kidney, strengthen bones and muscles, and induce diuresis for treating stranguria. It can be used in combination with poria cocos to treat aching pain in waist and knees caused by kidney deficiency-fire. Prepared rehmannia root can nourish yin and replenish blood. It can be used in combination with poria cocos to treat kidney yin deficiency. The combined use of poria cocos and largehead atractylodes rhizome can invigorate spleen and replenish qi. This TCM pair is suitable for patients with kidney qi deficiency. The poria cocos-Chinese magnoliavine fruit pair can tonify kidney, calm the mind, replenish qi and promote the secretion of body fluid. Poria cocos and radix ginseng can be used together to benefit the kidney and replenish qi. This TCM pair is suitable for patients with kidney qi deficiency.

In addition, eucommia bark and malaytea scurfpea fruit are also a frequent TCM pair. According to *Su Wen·Mai Yao Jing Wei Lun (素问·脉要精微论)*, kidney essence deficiency can result in aching pain in waist and knees if tendons and vessels in the waist are not strengthened. Kidney deficiency patients often show the symptoms of aching pain in waist and knees. Eucommia bark and malaytea scurfpea fruit can be used together to warm and nourish kidney yang, as well as strengthen bones and muscles.

4 Discussion

This paper studied TCMs frequently used for treating kidney deficiency. Their natures, flavors, channel tropism were analyzed, then key TCM pairs were identified and association rules were revealed, on which basis the prescription compatibility law was explored. In combination with the mining results, several classic methods for treating kidney deficiency were highlighted. Note, however, that this study only discussed various types of TCMs, but did not take into account the dosages of these TCMs since the measures adopted in different dynasties may be different. Further research will take the dosage of TCM as object and deeply explore the composition principles of prescriptions for treating kidney deficiency.

The WD-Get Rules used here is a new algorithm designed specially for prescription database. Compared with width-first algorithm and depth-first algorithm, WD-Get Rules has higher efficiency in mining association rules between TCMs in the prescription database. This new algorithm has good applicability and can be applied to mining of prescriptions for treating other diseases. Moreover, this algorithm should be further optimized, which can provide more technical support for the interactive integration of data mining and TCM research.

References

1. X. Tong, Q.Y. Xie, Q.G. Meng. Scientific value of TCM integrative data analysis in big data era. *Chinese Journal of Information on Traditional Chinese Medicine* **2015**, *22*, 1–3.

2. B. Zhang. Research on the application of data-mining technology based on association rules in traditional Chinese prescription compatibility. *Journal of Gansu Lianhe University (Natural Sciences)* **2011**, *25*, 82–86.
3. F.S. Chen. Correctly understand kidney deficiency and scientifically protect kidney. *China Consumer News* **2010**, 8th March.
4. A.F. Zhou. Theory of state of viscera in kidney and its clinical application. *Journal of Tianjin University of Traditional Chinese Medicine* **2014**, *33*, 1-5.
5. H. Li, J. Xiong, Q.R. Zhou. Advances in research on kidney deficiency. *Chinese Journal of Integrated Traditional and Western Nephrology* **2005**, *6*, 246-248.
6. Y.Y. Tong. Research on the application of association rules in the field of traditional Chinese medicine. *2008 Academic Annual Meeting of the Institute of Chinese Medicine Information, China Academy of Chinese Medical Sciences* **2009**.
7. J.W. Han, M. Kamber, J. Pei. Data mining: Concepts and Techniques (Third edition). China Machine Press: Beijing, China, 2017; pp. 10-320.
8. J. Yi, L. Nian, J.Y. Zhang. An overview of the study of *Sheng Ji Zong Lu*. *Liaoning Journal of Traditional Chinese Medicine* **2015**, 2024-2026.
9. Y. Zeng, J. Zhang. The application of data mining technology in the field of traditional Chinese medicine. *Chinese Journal of Information on Traditional Chinese Medicine* **2012**, *19*, 99-100.
10. An overview of the research on the masterpiece book of prescriptions *Pu Ji Fang* since the 1980s. *Journal of Pingdingshan University* **2012**, *27*, 48-55.
11. W.M. Li W.H. Li. Research on the compatibility law of prepared rehmannia root in prescriptions. *China Journal of Traditional Chinese Medicine and Pharmacy* **2011**, 2810-2812.
12. Y. Lin, Y.C. Wang. Is it feasible to benefit human organs by eating similarly-shaped food? *TCM Healthy Life-Nurturing* **2017**, 31–34.
13. J.M. Li. Is eating similarly-shaped food to benefit human organs equal to eating animal organs to benefit corresponding human organs? *China Drug Store* **2014**, *12*, 80-81.
14. B. Xiao, Y. Wang, W.J. Guo, *et al*. Relationship between Chinese herbal nature combination and functions. *World Science and Technology-Modernization of Traditional Chinese Medicine* **2010**, *12,* 902-908.
15. C.L. Liu, S.J. Qiu, X.B. Liu. Discussion on the connotation of the theory of association between spleen and kidney. *Journal of Guangzhou University of Traditional Chinese Medicine* **2009**, *26*, 491-494.
16. H. Zhao, X.L. Rong, Z.X. Chen. Objective research on splenic asthenia, nephrasthenia syndrome and correlation between spleen and kidney. *Chinese Journal of Tissue Engineering Research* **2006**, *10*, 130-135.
17. Y.B. Li, Z.M. Gao. Analysis of the characteristics of Zhang Jingyue's Yang strengthening method–strengthen Yang and at the same time nourish yin. *China Modern Medicine* **2009**, *16*, 77-78.
18. M.Z. Zhang, D.N. Zhang, M.Y. Zhang, *et al*. Clinical study on treatment of 160 cases of IgA nephropathy by method of tonifying the kidney and activating blood circulation. *Journal of Traditional Chinese Medicine* **2006**, *47*, 38-40.

19. D.N. Zhang. Research on the TCM Method of Tonifying Kidney and Promoting Blood Circulation. China Medical Science Press: Beijing, China, 1997.
20. Z. Li. Classic and famous traditional Chinese medicine–Poria cocos. *Cancer Frontier* **2011**, *3*, 52-53.

Association Rule-Generation Algorithm Integrating Width-First and Depth-First Search and Its Application

Xiaoling Zhou[1] Xinyou Zhang[1*] Diyao Wu[2] Liping Liu[1] Yongkun Guo[1]
(1. School of Computer, Jiangxi University of Traditional Chinese Medicine, Nanchang, Jiangxi, China 330004;
2. School of Pharmacy, Jiangxi University of Traditional Chinese Medicine, Nanchang, Jiangxi, China 330004) *E-mail: 1084198181@qq.com

Abstract: When width-first strategy is used to generate strong association rules from frequent itemsets, it often requires a long time to generate candidate consequent set. Therefore, some research proposed a depth-first algorithm called GRSET. This algorithm can effectively shorten the time required to generate candidate consequent set. However, GRSET has the disadvantage of repetitively evaluating the consequents of weak association rules in the consequent set. As a result, GRSET is not always more efficient than the width-first algorithm. In this paper, a new algorithm integrating both width-first and depth-first algorithms was developed for association rule generation. This algorithm (called WD-Get Rules), width-first algorithm and GRSET were applied to T40I10D100K, T10I4D100K, Mushroom, Connect and text dataset of single traditional Chinese medicines. Results indicate that WD-Get Rules always show higher efficiency than the other two algorithms at different combinations of minimum support thresholds and minimum confidence thresholds. This work may help improve association rule-generation algorithms.
Keywords: frequent itemsets; association rule generation, width first; depth first.
CLC number: TP311 Document code: A Article number:

1. Introduction

The generation of strong association rules from frequent itemsets is often divided into two steps. In the first step, non-empty subsets corresponding to each frequent itemset are generated. In the second step, for the rule generated by taking each non-empty subset as rule consequent, confidence is calculated. If it satisfies the minimum confidence threshold, then the rule is considered as a strong rule; otherwise, the rule is not a strong rule [1–3]. When width-first algorithm is used to generate strong association rules from frequent itemsets, it often requires a long time to generate candidate consequent set. Some research thus proposed a depth-first algorithm called GRSET for association rule generation [4]. This algorithm uses set enumeration tree [5] to reduce dimensions and uses depth-first strategy to generate rule consequents. However, GRSET necessitates repetitive evaluation of the consequents of weak association rules in the consequent set, thus its efficiency is not always higher than that of width-first algorithm.

In this paper, a new association rule-generation algorithm integrating both width-first and

depth-first strategies was developed. Results indicate that this algorithm (called WD-Get Rules) shows higher efficiency than width-first algorithm and GRSET at different combinations of minimum support thresholds and minimum confidence thresholds.

2. Association rule generation

2.1 Concepts of association rules

Suppose that $\mathrm{I}=\{I_1, I_2, \cdots, I_n\}$ is a collection of n items and $\mathrm{T}=\{T_1, T_2, \cdots, T_m\}$ is a collection of m task-related database affairs. It is required that each affair is a non-empty itemset. Itemset is a collection of items and itemset consisting of k items is called k-itemset. Note that $T_k \subseteq \mathrm{I}(1 \le k \le \mathrm{m})$. Suppose that $A(A \subseteq \mathrm{T})$ and $B(B \subseteq \mathrm{T})$ are two itemsets and $A \subset \mathrm{I}$, $B \subset \mathrm{I}$, $A \neq \varnothing$, $B \neq \varnothing$, $A \bigcap B = \varnothing$. Association rule has the form of "$A \Rightarrow B$" [6-8], where A is rule antecedent and B is rule consequent. The support and confidence [9,10] of the rule are calculated according to equations (1) and (2), respectively.

$$\begin{gathered} support(A \Longrightarrow B) \\ = P(A \cup B) \\ -the\ probability\ that\ both\ A\ and\ B\ occur\ in\ T \\ -\frac{the\ number\ of\ occurences\ of\ both\ A\ and}{the\ total\ number\ of\ affairs\ in\ T} \end{gathered} \tag{1}$$

$$\begin{gathered} confidence\ (A \Longrightarrow B) \\ = P\left(\frac{B}{A}\right) \\ -\frac{the\ probability\ that\ both\ A\ and\ B\ occur\ in\ T}{the\ probability\ that\ A\ occurs\ in\ T} \\ -\frac{the\ number\ of\ occurrences\ of\ both\ A\ and\ B\ in\ T}{the\ number\ of\ occurrences\ of\ A\ in\ T} \end{gathered} \tag{2}$$

If the support of "$A \Rightarrow B$" is no lower than the minimum support threshold (*min-sup*) [11], then "$A \Rightarrow B$" is a frequent pattern. If the support and confidence of "$A \Rightarrow B$" are no lower than the minimum support threshold and minimum confidence threshold (*min-conf*) [2,11], respectively, then "$A \Rightarrow B$" is a strong association rule.

Equation (2) suggests that the confidence of "$A \Rightarrow B$" can be calculated as the ratio of "the number of occurrences of both A and B in T" to "the number of occurrences of A in T". If the number of occurrence is defined as support count, then the minimum support threshold can be transformed to minimum support count (minimum support) during recognition of frequent itemsets and only support count of each frequent itemset is obtained. Then, the confidence of rule can be easily calculated. This paper focuses on the generation of strong association rules from frequent itemsets, thus the following discussions are based on the assumption that frequent itemsets have already been obtained.

Assume that there is a frequent itemset $L = L_1 \cup L_2 \cup \cdots \cup L_k$, where L_k denotes the collection of frequent k-itemsets and each of them is ordered (*i.e.*, certain frequent itemset

$f=\left(\left[Ii\right],\cdots,\left[Ij\right],\cdots,\left[Ik\right],\cdots,\left[Ip\right],\cdots,\left[Iq\right]\right)(i<j<k<p<q)$). All frequent *j*-itemsets are stored in the *jth* row of L. The support count (supData) of each frequent itemset is obtained. supData is a Map object and key-value pairs are stored. For example, suppose that there is a frequent itemset $\{I1,I2,I3,I5\}$. The support count of the frequent itemset is count and then supData (' $I1,I2,I3,I5$ ')=count. Then one can directly obtain the support count (supCount) of $\{I1,I2,I3,I5\}$ according to the equation supCount = supData(' $I1,I2,I3,I5$ '). The confidence of each rule can be easily calculated according to the support count. For the finally output rule, its support should be given. Therefore, the total number of affairs (TIDcount) should be transmitted to the rule generation function. Then, rule support can be calculated as the ratio of corresponding support count to TIDcount.

A lot of rules can be generated from one frequent itemset. For example, the association rules that can be generated from $\{I1,I2,I3,I5\}$ are shown in Fig. 1. Except that the rule in the first layer is not an association rule, all other rules including "I2,I3,I5=>I1", "I1,I3,I5=>I2", "I1,I2,I5=>I3", "I1,I2,I3=>I5", "I3,I5=>I1,I2", "I2,I5=>I1,I3", "I2,I3=>I1,I5", "I1,I5=>I2,I3", "I1,I3=>I2,I5", "I1,I2=>I3,I5", "I5=>I1,I2,I3", "I3=>I1,I2,I5", "I2=>I1,I3,I5" and "I1=>I2,I3,I5" are possibly strong association rules.

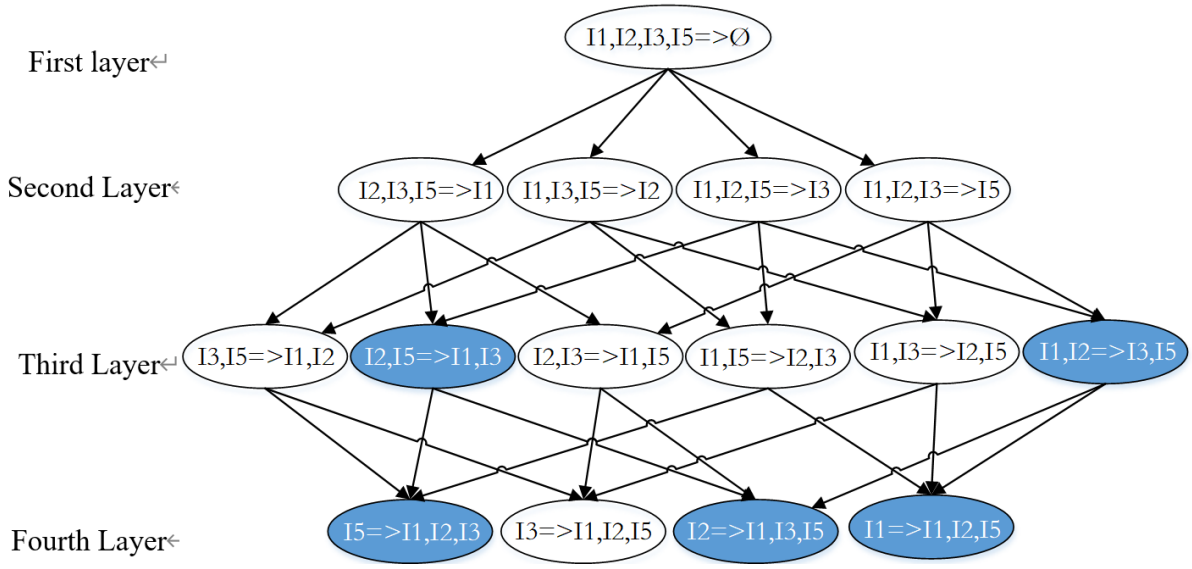


Figure 1. Association rule network for frequent itemset $\{I1,I2,I3,I5\}$.

2.2 Width-first algorithm for strong association rule generation

The width-first algorithm for association rule generation [12] works as follows. Each rule is searched and if one rule does not satisfy the minimum confidence, then the set including the consequent of this rule will also not satisfy the minimum confidence. For example, as shown in Fig. 1, if "I2,I5=>I1,I3" and "I1,I2=>I3,I5" do not satisfy the minimum confidence, then "I5=>I1,I2,I3", "I2=>I1,I3,I5" and "I1=>I2,I3,I5" are not strong association rules since their consequents "I1,I2,I3", "I1,I3,I5" and "I2,I3,I5" include "I1,I3" or "I3,I5". When all the strong association rules in the *kth* (*k*>1) layer are found, the set of their corresponding consequents (I)

can be obtained. If I contain more than one items, then "self-join" is adopted to generate a set of consequents in the *(k+1)th* layer for I. In order to adopt self-join, the first (*k*-1) items should be the same. For example, among <I1,I2,I3>, <I1,I2,I4> and <I1,I4,I5>, the first two items in <I1,I2,I3> and <I1,I2,I4> are both <I1,I2>, thus <I1,I2,I3> and <I1,I2,I4> can be combined to <I1,I2,I3,I4>. The first two items in <I1,I2,I3> and <I1,I4,I5> are different, thus they cannot be combined. For the convenience of description, the width-first algorithm for strong association rule generation is denoted as W-Get Rules.

2.3 Depth-first algorithm for strong association rule generation

Depth-first algorithm for strong association rule generation (GRSET) is based on set enumeration tree. It is proposed to address the problem that width-first algorithm requires a long time to generate candidate consequent set. GRSET is a typical depth-first algorithm [4].

Suppose that there is a frequent itemset $L = L_1 \cup L_2 \cup \cdots \cup L_k$, where L_k denotes the collection of *k*-itemsets and each of them is ordered. We also have the following:

Suppose $f=([Ii],\cdots,[Ij],\cdots,[Ik],\cdots,[Ip],\cdots,[Iq])(i<j<k<p<q)$ is a frequent itemset. $pos(f, Ir)$ denotes the position of Ir in f. $f[j]$ denotes the j*th* element in *f*. Suppose that *h* and *a* are two item subsets. $\|a\|$ denotes the number of items in *a*. If *h* consists of elements that are the first *j* (0<=j<||*a*||) elements of *a*, then *h* is called the head of *a*.

Literature [4] has shown that if the rule, whose consequent is $h + f[j]$, is not a strong association rule, then the consequents of association rules do not include $h + f[j]$. Therefore, we only need to verify whether $h + f[j+1]$ is the consequent of strong association rule. Table 1 shows the GRSET algorithm for generating all strong association rules from *L*.

Table 1. GRSET algorithm.

```
Algorithm: GRSET (L, min_conf, TIDcount, supData)
Input: frequent itemset (L), the minimum confidence threshold (min_conf), the total number of affairs
(TIDcount), support count (supData).
Output: all association rules (F) generated from L

F is initially {'rule', 'support', 'confidence'}
Obtain the number of rows in L (m)
For i=2:m start from frequent 2-itemset
    Obtain all frequent i-itemsets and save them as temporary variable L1
    Obtain the number of rows in L1 (m1)
    For j=1:m1
        Obtain the jth subset in L1 (freqSet)
        Combine all items in freqSet as string S
        F=getall(Ø, 1, freqSet, min_conf, TIDcount, supData, i, F, S)
    end
end
```

Note that Ø is an empty set. For getall algorithm, see Table 2.

Table 2. getall algorithm.

```
Algorithm: getall(h, s, freqSet, min_conf, TIDcount, supData, k, F, S)
Input: consequent set (h), the initial positions of items in freqSet that need to be combined (s, the position
of the last item in h +1≤s≤k), freqSet, the minimum confidence threshold (min_conf), the total number of
affairs (TIDcount), support count (supData), the number of items in freqSet (k), rule set (F), string
generated by combining all items in freqSet (S).
Output: all association rules generated from freqSet and whose consequents have h as head

Obtain the number of elements in h (n)
If n<k-1
    For j=s:k
       Obtain the jth element in freqSet (a)
       Obtain rule antecedent C=freqSet-h-a (- represents subtraction operation)
       Combine all items in C into rule antecedent string front
       Obtain rule confidence conf= supData(S)/supData(front)
       If conf>=min_conf
            Combine h+a into rule consequent string behind (+ represents union operation)
            Obtain rule f1= front=>behind
            Rule support sup= supData(S)/TIDcount
            Add f1, sup, conf to F F = [F; f1, sup, conf]
            If j<k
                getall(h+a, j+1, freqSet, min_conf, TIDcount, supData, k, F, S)
            end
       end
   end
end
```

Taken Fig. 1 as an example, at first, h=Ø, a={I1} and consequent {I1} is obtained. Then, we obtain that "I2,I3,I5=>I1" is a strong association rule and it is added to F. Based on depth-first strategy, h={I1}, a={I2} and consequent "I1,I2" is obtained. Then, we obtain that "I3,I5=>I1,I2" is a strong association rule and it is added to F. At this time, h={I1,I2}, a=I3, and consequent "I1,I2,I3" is obtained. Then, we obtain that "I5=>I1,I2,I3" is not a strong association rule. Subsequently, h={I1,I2}, a={I5}, and consequent "I1,I2,I5" is obtained. It is then obtained that "I3=>I1,I2,I5" is a strong association rule and is added to F. h={I1},a={I3}, and consequent"I1,I3" is obtained. We then obtain that "I2,I5=>I1,I3" is not a strong association rule. h={I1}, a={I5} and consequent "I1,I5" is obtained. It is further obtained that "I2,I3=>I1,I5" is a strong association rule and is added to F. In this way, all strong association rules from {I1,I2,I3,I5} can be generated.

3 WD-Get Rules algorithm

In W-Get Rules, much time is spent on generating candidate consequent set. Although candidate consequent set is not generated in GRSET, the rules whose consequents include the consequent of certain weak association rule might be repetitively evaluated in this method. As shown in Fig. 2, the rules in blue ellipses are weak association rules. In GRSET, "I1,I5", "I2,I5" and "I3,I5" will be evaluated subsequently. When frequent itemset includes more elements, such case will occur more frequently.

In fact, if width-first search is adopted, it can be easily found that rules with I5 included in

the consequents are not strong association rules. Therefore, time will not be spent on evaluating "I1,I5", "I2,I5" and "I3,I5". By careful analysis, we have found the following:

(1) If an element Ii in frequent itemset f cannot become the consequent of a strong association rule, then all the rules with Ii included in the consequents are not strong association rules.

(2) Suppose $f=\left([Ii],\cdots,[Ij],\cdots,[Ik],\cdots,[Ip],\cdots,[Iq]\right)(i<j<k<p<q)$ is a frequent itemset. After width-first search for one layer, a set of consequents (H) that can become the consequents of strong association rules is obtained. The itemsets in H are all frequent 1-itemsets. Then, consequents consisting of more than two elements that can be formed can only be obtained from H.

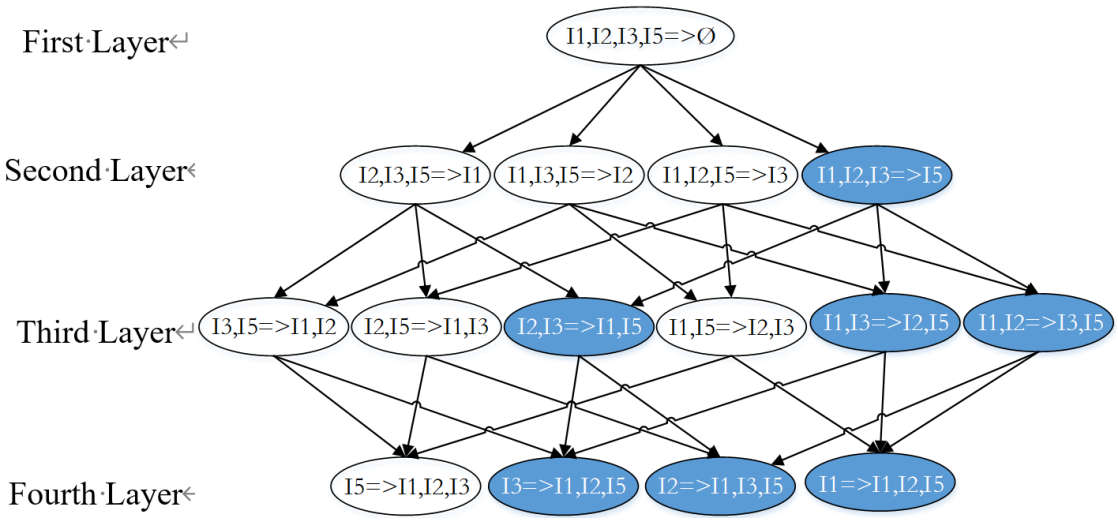


Figure 2. Association rule network for frequent itemset $\{I1,I2,I3,I5\}$.

On the basis of above findings, this paper proposed a new association rule-generation algorithm integrating both width and depth search (WD-Get Rules). The basic idea is that: Width-first strategy is used first to find a set of consequents (H) than can become the consequents of strong association rules. In this set, each consequent is a frequent 1-itemset. On the basis of H, GRSET is used for depth search. During depth search, rule consequents only consist of elements from H. In this way, elements that cannot become the consequents of strong association rules are avoided. If the itemset is frequent k-itemset (k>2) and there are more than one elements in H, then depth search can be performed; otherwise, next frequent itemset is analyzed.

Suppose that there is a frequent itemset $L=L_1\cup L_2\cup\cdots\cup L_k$, where L_k is the collection of frequent k-itemsets and each of them is ordered. Table 3 shows the WD-Get Rules.

Table 3. WD-Get Rules algorithm.

```
Algorithm: WD-Get Rules(L, min_conf, TIDcount, supData)
Input: frequent itemset (L), the minimum confidence threshold (min_conf), the total number of affairs
(TIDcount), support count (supData)
Output: all association rules generated from L (F)

F is initially {'rule', 'support', 'confidence'}
Obtain the number of rows in L (m)
For i=2:m    start from frequent 2-itemset
    Obtain all frequent i-itemsets and save them as temporary variable L1
    Obtain the number of rows in L1 (m1)
    For j=1:m1
        Obtain the jth subset in L1 (freqSet)
        %perform width search
         After initialization, the set of consequents that can be the consequents of strong association
         rules (H) is empty.
        Combine all items in freqSet into string (S)
        For r=1:i
            Obtain the rth element in freqSet (a)
            Use behind to save a in string format
            Save the difference set between freqSet and a in C
            Combine the sequence of elements in C to front
            Calculate rule confidence conf=sup(S)/sup(front)
            If conf>=min_conf
                Obtain rules f1= front=>behind
                Rule support sup= supData(S)/TIDcount
                Add f1, sup, conf to F F=[F;f1,sup,conf]
                Add a to H
            end
        end
        %width search is finished
        Obtain the number of elements in H (num)
        If i>2 && num>1
            %perform depth search
            For jj=1:num
                Obtain the jjth element in H (h)
                s=jj+1
                If s<=num
                        F=getall_new(h, s, freqSet, min_conf, TIDcount, supData, i, F, S, H)
                end
            end
        else
            continue
         end
    end
end
```

For getall_new algorithm, see Table 4.

Table 4. Getall_new algorithm.

```
Algorithm: getall_new(h, s, freqSet, min_conf, TIDcount, supData, k, F, S)
Input: consequent set (h), the initial positions of items that need to be combined in freqSet (s, the position
of the last item in h +1≤s≤k), freqSet, minimum confidence threshold (min_conf), the total number of
affairs (TIDcount), support count (supData), the number of items in freqSet (k), rule set (F), string
generated by combining all items in freqSet (S)
Output: all association rules generated from freqSet and whose consequents include h.

Obtain the number of elements in h (n)
If n<k-1
    For j=s:k
       Obtain the jth element in freqSet (a)
       Calculate rule antecedent C=freqSet-h-a (- represent subtraction operation)
       Combine all items in C into rule antecedent string front
       Calculate rule confidence conf= supData(S)/supData(front)
       If conf>=min_conf
            Combine h+a into rule consequent string behind (+ represents union operation)
            Obtain rule f1= front=>behind
            Rule support sup= supData(S)/TIDcount
            Add f1, sup, conf to F F=[F;f1,sup,conf]
            If j<k
                getall(h+a, j+1, freqSet, min_conf, TIDcount, supData, k, F, S)
            end
       end
   end
end
```

4 Experiment and analysis

Simulation was performed on Matlab [13–15]. Classical Apriori algorithm [16-18] was used to recognize frequent itemsets. W-Get Rules, GRSET and WD-Get Rules were used to generate association rules from T40I10D100K, T10I4D100K, Mushroom, Connect [19,20] and text dataset of single traditional Chinese medicines (TCMs). Since these datasets have different properties, the adopted minimum support thresholds (*min-sup*) were also different. For frequent patterns obtained at different minimum support thresholds, the minimum confidence threshold (*min-conf*) was changed from 0 to 1 with a step of 0.1.

For T40I10D100K dataset, the *min-sup* was set at 0.03, 0.025 and 0.02, respectively, and the *min-conf* was changed from 0 to 1 with a step of 0.1. The running time of the three algorithms are shown in Figs. 3–5. Clearly, at the same *min-sup*, the running time of the three algorithms decrease with increase in *min-conf*. Notably, the running time of WD-Get Rules is always shorter than the running time of W-Get Rules and GRSET.

For T10I4D100K dataset, the *min-sup* was set at 0.008, 0.007 and 0.005, respectively, and the *min-conf* was changed from 0 to 1 with a step of 0.1. The running time of the three algorithms are shown in Figs. 6–8. When the *min-sup* $\geq$ 0.007 and the *min-conf* $\leq$ 0.1, the running time of GRSET is shorter than that of W-Get Rules. However, when the *min-sup* $<$ 0.007 and the *min-conf* $>$ 0.1, the running time of GRSET is longer than that of W-Get Rules. This illustrates that GRSET is not always more efficient than W-Get Rules. Note however that WD-Get Rules always enables a shorter running time than the other two algorithms.

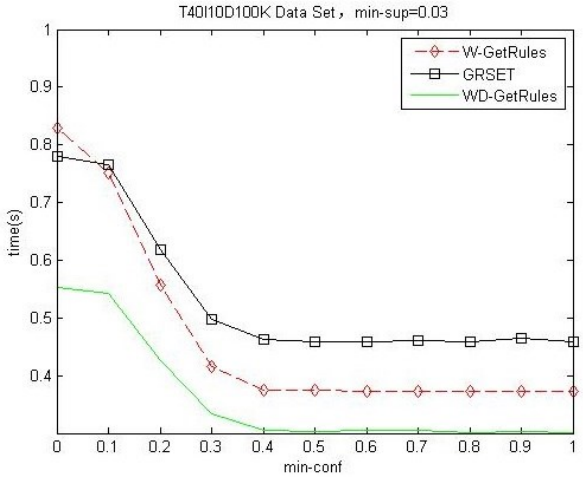


Figure 3. The running time of three algorithms at different min-conf (T40I10D100K dataset, *min-sup*=0.03).

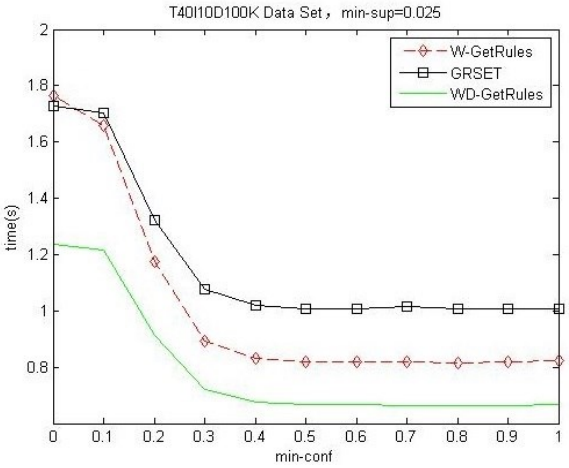


Figure 4. The running time of three algorithms at different min-conf (T40I10D100K dataset, *min-sup*=0.025).

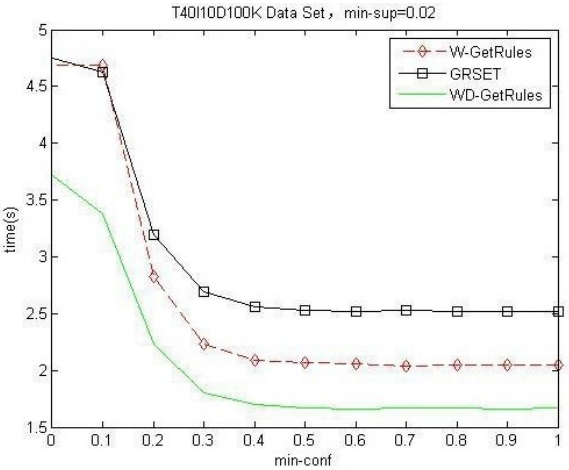


Figure 5. The running time of three algorithms at different min-conf (T40I10D100K dataset, *min-sup*=0.02).

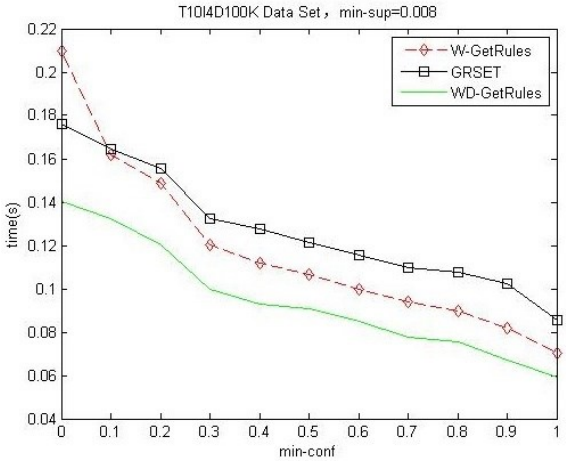


Figure 6. The running time of three algorithms at different min-conf (T10I4D100K dataset, *min-sup*=0.008).

For Mushroom dataset, the *min-sup* was set at 0.8, 0.5 and 0.3, respectively, and the *min-conf* was changed from 0 to 1 with a step of 0.1. The running time of the three algorithms are shown in Figs. 9–11. When *min-sup* ≥ 0.5 and *min-conf* = 0 (or 1), the difference in running time between any two of the three algorithms is less than 2 s. When *min-sup* = 0.3 and *min-conf* = 0 (or 1), the difference in running time between any two of the three algorithms is more than 700 s. We can see that GRSET is not always more efficient than W-Get Rules, but in no case is the efficiency of WD-Get Rules lower than those of the other two algorithms.

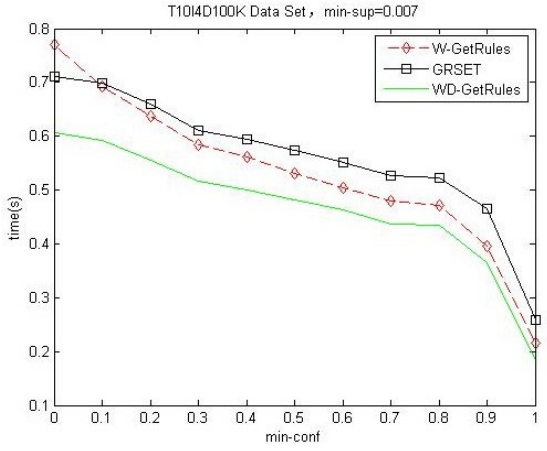


Figure 7. The running time of three algorithms at different min-conf (T10I4D100K dataset, *min-sup*=0.007).

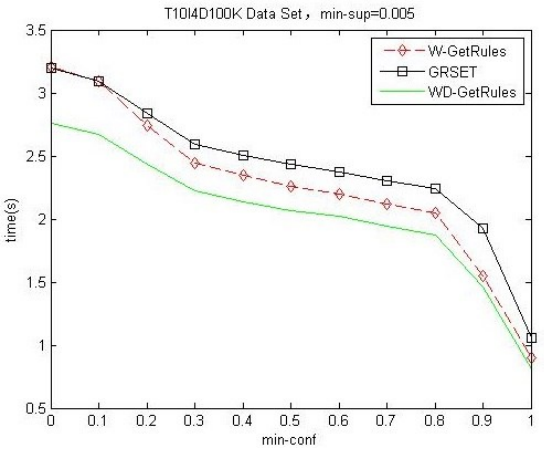


Figure 8. The running time of three algorithms at different min-conf (T10I4D100K dataset, *min-sup*=0.005).

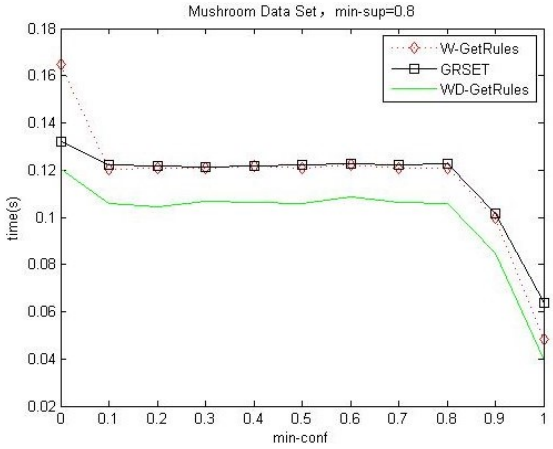


Figure 9. The running time of three algorithms at different min-conf (Mushroom dataset, *min-sup*=0.8).

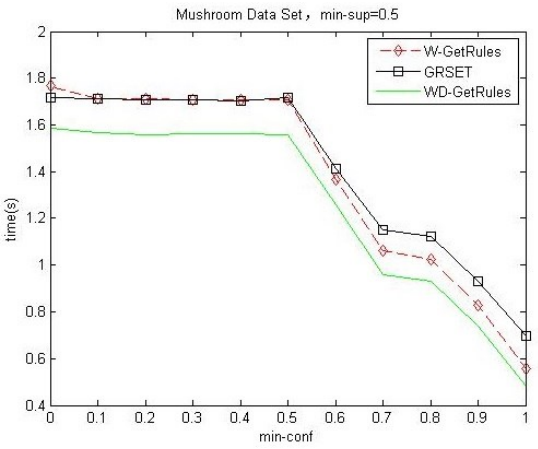


Figure 10. The running time of three algorithms at different min-conf (Mushroom dataset, *min-sup*=0.5).

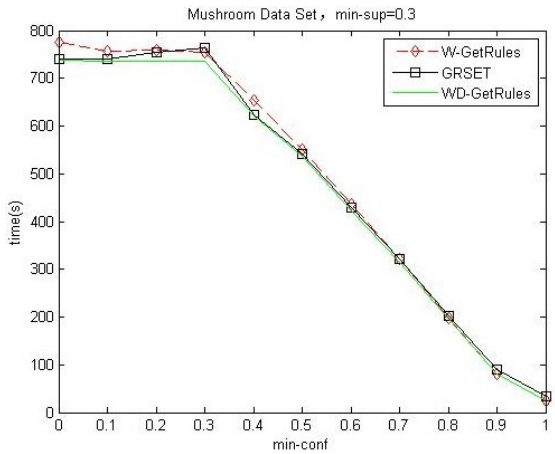


Figure 11. The running time of three algorithms at different min-conf (Mushroom dataset, *min-sup*=0.3).

For Connect dataset, the *min-sup* was set at 0.98, 0.97 and 0.95, respectively, and the *min-conf* was changed from 0 to 1 with a step of 0.1. The running time of the three algorithms are shown in Figs. 12–14. When *min-conf* ≤ 0.9, WD-Get Rules enables a significantly shorter running time than the other two algorithms. When *min-conf* > 0.9, the difference in efficiency between the three algorithms seems insignificant.

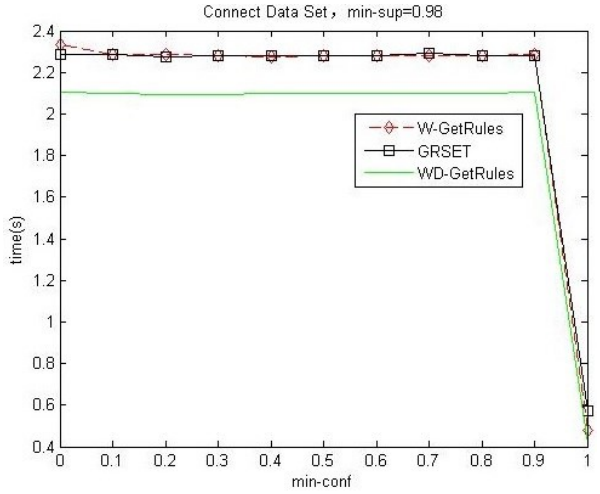


Figure 12. The running time of three algorithms at different min-conf (Connect dataset, *min-sup*=0.98).

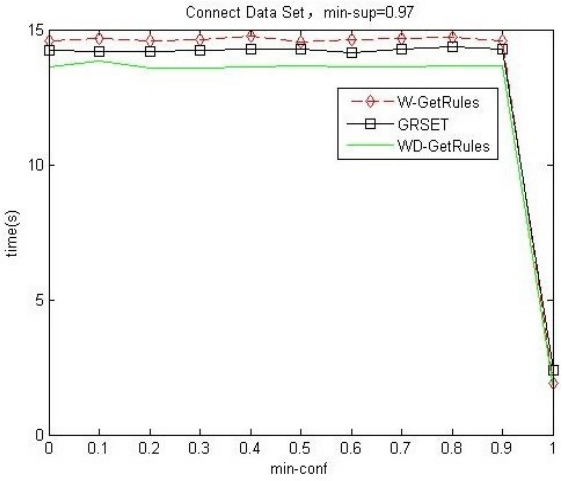


Figure 13. The running time of three algorithms at different min-conf (Connect dataset, *min-sup*=0.97).

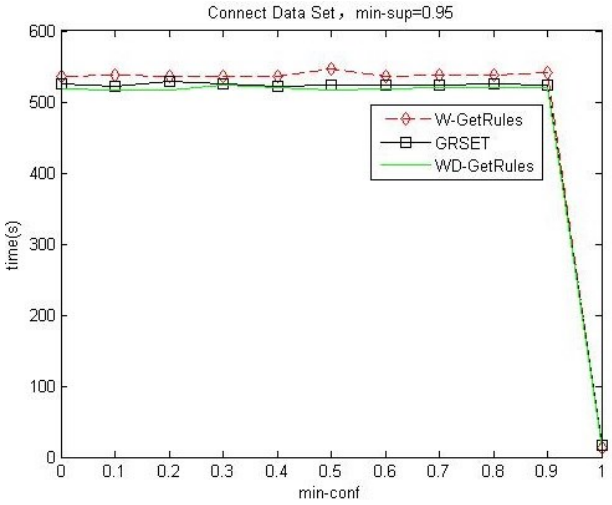


Figure 14. The running time of three algorithms at different min-conf (Connect dataset, *min-sup*=0.95).

For TCM dataset, the *min-sup* was set at 0.2, 0.08 and 0.06, respectively, and the *min-conf* was changed from 0 to 1 with a step of 0.1. The running time of the three algorithms are shown in Figs. 15–17. At the same *min-conf*, the lower is the *min-sup*, the more efficient is WD-Get Rules than W-Get Rules and GRSET.

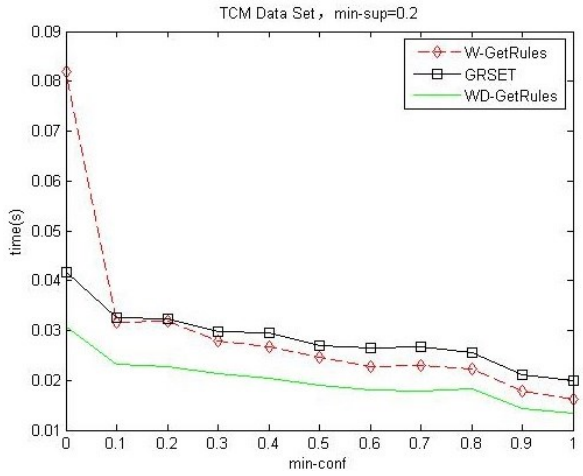


Figure 15. The running time of three algorithms at different min-conf (TCM dataset, *min-sup*=0.2).

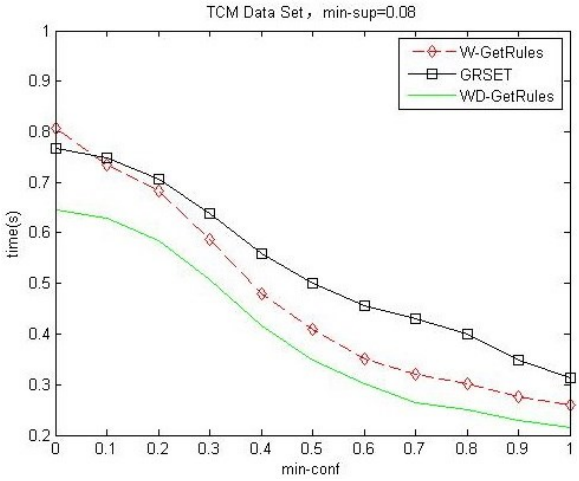


Figure 16. The running time of three algorithms at different min-conf (TCM dataset, *min-sup*=0.08).

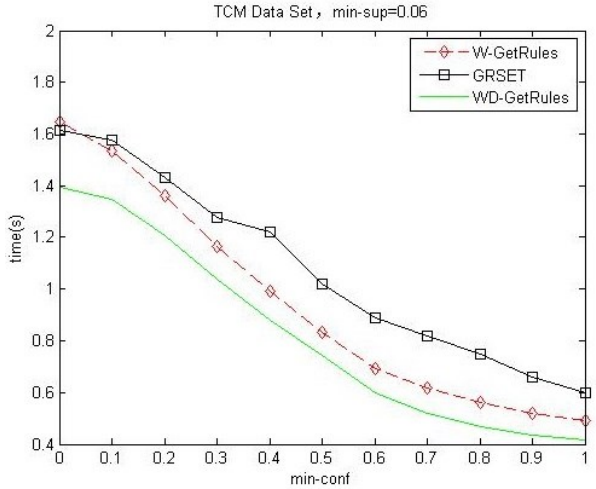


Figure 17. The running time of three algorithms at different min-conf (TCM dataset, *min-sup*=0.06)

5 Conclusions

Width-first and depth-first algorithms for association rule generation were analyzed, on which basis a new association rule-generation algorithm integrating both width-first and depth-first search was developed. In this algorithm (called WD-Get Rules), depth-first search was performed first, then frequent 1-itemsets of consequents that can become the consequents of strong association rules were stored in variable H, and finally depth-first search was performed based on H. WD-Get Rules, W-Get Rules and GRSET were then applied to five datasets including T40I10D100K, T10I4D100K, Mushroom, Connect and TCM. Results show that WD-Get Rules is more efficient than the other two algorithms at different minimum support thresholds and minimum confidence thresholds. WD-Get Rules not only overcomes the problem that width-first algorithm requires a long time to generate candidate consequent set,

but it also solves the problem that depth-first algorithm repetitively evaluate consequents of weak association rules in the consequent set. This study may provide new insights into the improvement of association rule-generation algorithms.

References

1. K. Sumathi, S. Kannan, K. Nagarajan. A search space reduction algorithm for mining maximal frequent itemset. *International Journal of Computer Applications* **2013**, *82*, 32-36.
2. J.W. Han, M. Kamber, J. Pei, *et al*. Data Mining: Concepts and Techniques. Machinery Industry Press: Beijing, China, 2012.
3. T. Mai, B. Vo, L.T.T. Nguyen. A lattice-based approach for mining high utility association rules. *Information Sciences* **2017**, *399*, 81-97.
4. K. Wu, N.X. Li, Q. Wei, *et al*. Association rules generating algorithm based on set-enumeration tree. *Computer Engineering and Applications* **2006**, *42*, 152-155.
5. J. Gao, B.L. Shi. Research on fast association rule mining algorithm. *Computer Science* **2005**, *3*, 200–201.
6. D. Martin, A. Rosete, J. Alcala-Fdez, *et al*. A new multiobjective evolutionary algorithm for mining a reduced set of interesting positive and negative quantitative association rules. *IEEE Transactions on Evolutionary Computation* **2014**, *18*, 54-69.
7. L.T.T. Nguyen, B. Vo, *et al*. ETARM: an efficient top-k association rule mining algorithm. *Applied Intelligence* **2017**, *5*, 1-13.
8. L. Galárraga, C. Teflioudi. Amie: association rule mining under incomplete evidence in ontological knowledge bases. **2013**, 413-422.
9. Liao S.H., Chen Y J. A rough set-based association rule approach implemented on exploring beverages product spectrum. *Applied Intelligence* **2014**, *40*, 464-478.
10. R. Zhang, W. Chen, T.C. Hsu, *et al*. ANG: a combination of Apriori and graph computing techniques for frequent itemsets mining. *Journal of Supercomputing* **2017**, *3*, 1-16.
11. V.D. Hai, T.C. Truong. An efficient method for mining association rules based on minimum single constraints. *Vietnam Journal of Computer Science* 2015, 2, 67-83.
12. P. Harrington. Machine Learning in Action. Posts & Telecom press: Beijing, China, 2013.
13. M. Grant, S. Boyd. CVX: Matlab Software for Disciplined Convex Programming, Version 1.21. *Global Optimization* **2008**, 155-210.
14. J. Lofberg. YALMIP: a toolbox for modeling and optimization in MATLAB. Optimization **2004**, *3*, 284-289.
15. C. Solomon, T. Breckon. Fundamentals of digital image processing: A practical approach with examples in Matlab **2018**, *16*, 1420-1424.
16. R. Agrawal. Mining association rule between sets of items in large databases. *ACM SIGMOD Conference on Management of Data* **1993**.

17. J. Rauch, M. Šimůnek. Apriori and GUHA–Comparing two approaches to data mining with association rules. *Intelligent Data Analysis* **2017**, *21*, 981-1013.
18. C.H. Chee, J. Jaafar, I.A. Aziz, *et al*. Algorithms for frequent itemset mining: a literature review. *Artificial Intelligence Review* **2018**, *3*, 1-19.
19. G. Bart. Frequent Itemset Mining Dataset Repository. Access online: http://fimi.ua.ac.be/data/.
20. X.H. Fu, D.J. Chen, Z.Q. Wang. Depth first frequent itemset mining based on bitable and inverted index. *Journal of Chinese Computer Systems* **2012**, *33*, 1747-1751.

ISBN 978-1-78900-105-1

£100