2018 International Conference on Chemical, Physical and Biological (ICCPB2018)

# Application of Data Mining in Data Preprocessing of Traditional Chinese Medicine

Rongchuan Guo
*Xinglin Medicine Research Centre*
*Jiangxi University of Traditional Chinese Medicine*
Jiangxi, Nanchang 330025, China
rongchuan613@163.com

***Abstract*—The data pretreatment research on traditional Chinese medicine prescriptions is to convert the information of the four flavors and five flavors, efficacy, returning, and toxicity of the medicine into digital information, so that it is more accurate in the data mining of Chinese medicine prescriptions, and it is the research and clinical practice of traditional Chinese medicine prescriptions. Practical research provides a theoretical reference. The regional differences in traditional Chinese medicine culture have brought many uncertainties in the data of traditional Chinese medicine. To solve the data problem of the new drug development decision support system based on data mining, a set of methods for regulating the original Chinese medicine data is proposed. The application of data reduction technology, clustering method and fuzzy set theory improved the quality of TCM data, and made important rules in the pre-treated Chinese medicine prescription database, which provided powerful decision support for the development of new Chinese medicine.**

***Keywords—Data Preprocessing, Data Mining, Data Reduction***

## I. Data Mining in the Field of Traditional Chinese Medicine

Data mining, also known as knowledge discovery, is the task of discovering patterns or patterns hidden in large amounts of data, that is, identifying effective, novel, and potentially useful data from large, incomplete, noisy, fuzzy, and random data sets. And the non-trivial process of the pattern that can ultimately be understood [1]. Data mining is mainly used for prediction and description. Its research includes classification, clustering, regression, association, rules and deviations. Common methods include statistical analysis, association rules, decision trees, fuzzy theory, rough sets, artificial neural networks and genetics. Algorithms, etc. [2]. In the field of traditional Chinese medicine, data mining applications mainly have the following aspects. The characteristics of Chinese medicine data can be further summarized as (1) data diversity: Chinese medicine data has a large time span, wide sources, and different types; (2) data complexity: Chinese medicine data is incomplete or redundant due to collection and processing techniques, due to history And human factors cause the lack of uniform standards and norms of information, and may involve personal privacy;(3) data non-quantitative: Chinese medicine data are mostly qualitative descriptions, lack of scientific unified quantitative expression; (4) data timeliness: due to different syndromes and pharmacological effects, Make some Chinese medicine data have certain timeliness [3]. Due to the above characteristics of traditional Chinese medicine data, the simple application of a certain mining method in the data mining process can only obtain the one-sided expression for the research object.[4] The fusion of different methods can help to fully explore the internal law; and the data mining through statistical induction and machine Learning to deal with the object, prefers to explore the statistical law of the surface layer, lack of in-depth discussion of the internal mechanism or connotation of the system.

## II. Chinese Medicine Data Preprocessing

Data preprocessing is mainly to normalize the data. Before formal data mining, especially when using a distance-based mining algorithm, such as neural network, k-nearest neighbor classifier, etc., data normalization must be performed. That is, converting Chinese medicine data information into actionable information and shrink to a specific range. Traditional Chinese medicine itself has a wealth of information, including the four flavors and five flavors of the drug, efficacy, menstruation, and toxicity[5]. The four-sex and five-flavors not only reflect the medicinal characteristics, but also further determine the properties of the prescription through odor and combination, so it is an important prescription information. However, Chinese medicine's understanding of the four sexes is rather vague. In order to make the data mining analysis of the drugs in the other agent more precise, the above information is converted into a digital form[6]. In the Chinese medicine prescription database, for the prescription of the prescription, the history of the symptom table and the manual recording of the prescription, there are blank fields, duplicate data, the name of the Chinese medicine and the description of the symptoms are not standardized, and the correction of the pretreatment process and Filtering can establish standards that conform to data mining standards.

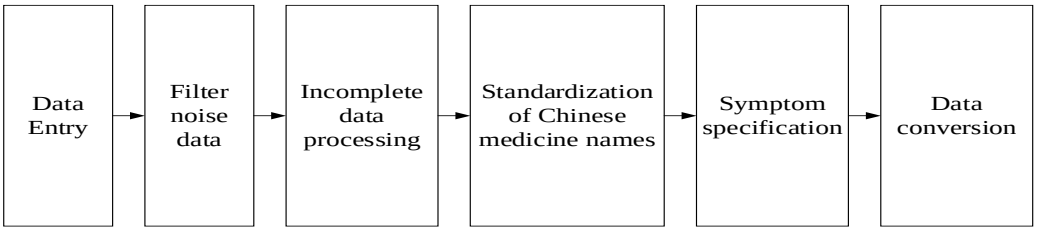


Figure 1 data preprocessing process

  DOI: 10.25236/iccpb.2018.046

*A. Processing of Noise Data*

Due to the historical span of traditional Chinese medicine prescriptions, some drug naming and symptom representations are inconsistent in different dynasties, and most of them are manual records, which results in a lot of noise data. If the processing of these noise data is not good, it will directly affect the later data mining effect. First, the defect data and duplicate data in the database should be filtered and filtered, and the wrong data should be modified. In the description of traditional Chinese medicine prescriptions, there are often phenomena such as polysemy, ambiguity, and semantic overlap. For example, the term "dizziness" is explained in the "Chinese Traditional Chinese Medicine Thesaurus" as "Dizziness is dazzling, dizziness is dizziness, collectively referred to as dizziness." However, it is not appropriate to use this term to describe the symptoms that appear alone, and it is not convenient to further analyze the symptoms later. Therefore, for "dizziness" you can use the words "dizziness" and "dizziness" instead, so that the description of the symptoms is more reasonable.

*B. Incomplete Data*

Incomplete data means that the object's properties have no value. The main reasons for incomplete data are: some data are not retained for historical reasons; there is no record due to equipment failure or misunderstanding; some attribute data of some prescriptions are considered to be unnecessary and artificially deleted during the inheritance process. Wait. Many of the data on traditional Chinese medicine prescriptions have incomplete data, such as the name of the drug, the dose of the drug, and the symptoms. Among them, because the prescription is mostly for manual recording, the defect of the dose data is the most prominent. For the solution to these problems, regression analysis and Bayesian algorithm can be used to infer the maximum possible value of the attribute. This is the solution adopted.

*C. Standardization of Chinese Medicine Names*

After the founding of the People's Republic of China, there have been many achievements in the standardization of Chinese medicine terminology, but there is still a considerable gap with the standard of modern terminology. In practice, irregular terminology is often seen in books and books, causing confusion. For example, chest pain, heartache, heartache, chest pain, heartache, heartache, violent heartache, clinical confusion. As for the name of the drug, it is more messy. Such as honeysuckle, also known as silver flower, double flower. With the increasing use of computers in the field of traditional Chinese medicine, some problems have arisen. Especially in the application of the old expert diagnosis and treatment system, the prescription database, etc., the name of the Chinese medicine is not standardized is one of the important reasons for the problem. The history of traditional Chinese medicine has a long history. In addition, China has a vast territory, many dialects, strong traditional Chinese culture, and the influence of minority medicine and foreign medicine, making the task of standardizing Chinese medicine terminology particularly difficult.

*D. Data Conversion*

Data conversion is the normalization of data, the conversion of data into actionable information, and the control of its value within the specified range. In the expression of traditional Chinese medicine, information such as the four characteristics of the drug, the efficacy, the presence or absence of toxicity, etc., are mostly expressed in text form. In order to mine the data in the other party, the results are more accurate, and the text information needs to be transformed. Data in digital form. For example, the cold, hot and cool four-symbols are represented by a coded form, and the succinic acid and saponin are coded by 1, 2, 3, 4, and 5, respectively, and the toxicity of the drug is expressed by floating-point data. In the expression of traditional Chinese medicine prescriptions, the expression of the dose is very irregular. Due to historical reasons, the dose is often expressed in terms of “Jin, liang, and Qian”. Because of the different historical dynasties, the unit of measurement for weight is inconsistent, so according to the age of the prescription, the unified weight data is converted, and finally it is converted into a modern "gram".

## III. Chinese Medicine Data Preprocessing Application

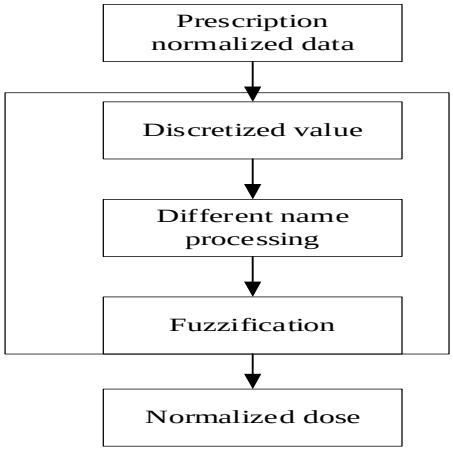


Figure 2 Standardization of the dose

In order to solve the above problems and smoothly classify data mining of Chinese medicine data, we propose a

corresponding set of preprocessing methods. The standardization of prescriptions mainly completes the process of orthogonal decomposition of components and elimination of redundancy. The attribute "prescription composition" is a long text type field, which is composed of some irregular data, which is not conducive to the data mining. Therefore, it must be orthogonally decomposed, and each Chinese medicine and its dose are separately stored, which will be " The elimination of the alias in ()" and the removal of the unimportant drug name message. In order to realize numerical data mining, we have designed a series of treatment methods to continue to process the “prescription composition” information, as shown in Figure 2. The order of each step is determined, and the change will affect the outcome of the process.

*A. Discretized Values*

This step is designed to discretize the processing table. Extract the numerical information in the dose and calculate it as follows:

Dosage = {Dose: If the dose is discrete data

Min+max/2: If the dose is a continuous value, it is expressed as [min,max]}

*B. Different Name Processing*

For the same kind of traditional Chinese medicine, with the changes of the times, different geographical environments, customs, etc., there may be multiple aliases. Therefore, the same Chinese medicine may appear under different names in the unilaterality of different medical workers. Our patent database was entered manually, so there is data on the "same drug name". In order to solve this problem, a mapping table of "Chinese medicine dictionary" was introduced. [2]The Chinese medicine dictionary contains the existing national standard Chinese herbal medicines, lists the unique index code of Chinese medicine, the Chinese proper name and the alias and sub-drug collection of the Chinese medicine. Multiple aliases of traditional Chinese medicine are listed in the alias attribute; the sub-category (SUB) attribute makes the "multi-drug-source" phenomenon available in the dictionary.

That is, for the same record, if the traditional Chinese medicine a∈SUB, then a and N are the same source and the drug is equivalent. Through the retrieval of the Chinese medicine dictionary, the drug names in the patent data are respectively matched with the proper names and aliases in the Chinese medicine dictionary, and the drug name is identified by the code of the drug, so that the "same drug name" can be solved. The Merger algorithm is following.

1) Read a Formula message;

2) for each Mi∈Formula, separate Mi and Chinese medicine

Match Chinese Chinese names, aliases, and subclasses in the dictionary;

If Mi=NJ then Mi=IDJ; if Mi∈BK then Mi=IDK;

If Mi∈SUBL then Mi=IDL;

3) the dose of the same herb in the combined formula,

If Mi=Mj then Wi=Wi+Wj, clear Mj and Wj;

Formula' is quickly sorted in the encoding order of Mi;

*C. Generalized Reduction*

The problem of multi-drug-source in the composition of prescriptions will undoubtedly increase the burden of data mining. [2]Therefore, we adopt data reduction technology to obtain a compressed representation of the data set, which is much smaller than the source data set, but still retains the original data. Integrity, such that mining on a reduced data set will be more efficient and produce the same or nearly identical analysis results.

*D. Dosing Weight*

In our data, the dose is measured by quality, and our results are ultimately reflected in the efficacy. However, the medicinal properties of traditional Chinese medicine are more complicated. The comprehensive consideration of four gas and five flavors, returning to the menstruation, lifting and sinking, toxic and non-toxic, all have an impact on the efficacy. [2]In order to obtain a more reasonable result, we used the result of multiplying the effective ingredient content coefficient of each drug with the drug quality as a dose, and corrected the dose to more accurately reflect the effect of the drug. Finally, in order to meet the needs of data mining, the percentage of prescription Chinese medicine is used instead of the dose.

*E. Fuzzy Processing*

After the above treatment, the names of the traditional Chinese medicines are represented by codes in sequence, and the dosage has a suitable value, but the dosage of each medicine is different, which leads to a problem: the dosage forms a continuous numerical variable. . On the one hand, it is not conducive to the treatment of the dose of the classified data mining; on the other hand, the formation of the dose of a certain potion in a certain prescription is too one-sided, which affects the quality of the rule. To this end, we borrow the fuzzy set theory, and then use our membership function to turn the dose of each drug into a fuzzy set, which not only meets the requirements of classified data mining but also does not lose its universality.

## IV. Conclusion

Although data mining technology still faces many problems and challenges, it is undeniable that data mining technology is a young and promising research field. Every year, new data mining methods and models are available. People are studying it more and more extensively. . It is necessary to establish a Chinese medicine data warehouse and use data mining technology for Chinese medicine research. This is a research prospect with promising prospects and challenges. Data mining, as a powerful tool for acquiring knowledge in massive data, is increasingly applied to all aspects of Chinese medicine. This will increase the academic level of Chinese medicine, the progress of modernization research, and the expansion of living space.


## Acknowldgements

This research was supported by the foundation of Jiangxi University Of Traditional Chinese Medicine(No.2013YHS013).


## References


[1] Wang Z , She K K . The Application of Data Mining Technology in the Data Analysis of Traditional Chinese Medicine[J]. Applied Mechanics and Materials, 2014, 687-691:1266-1269.

[2] Yu C , Ying X . International Forum on Computer Science-Technology and Applications - Application of Data Mining Technology in E-Commerce[J]. 2009:291-293.

[3] Wilson A M,ThabaneL,Holbrook A.Application of data mining techniques in pharmacovigilance. [J]. Br J Clin Pharmacol, 2015, 57(2):127-134.

[4] Jianjun H U . Design of Preprocessing in Data Mining System for Traditional Chinese Medicine Prescriptions Information[J]. Computer Engineering, 2008, 34(21):1223-1227.

[5] He, Y. , Zheng, X. , Sit, C. , Loo, W. T. , Wang, Z. Y. , & Xie, T. , et al. (2012). Using association rules mining to explore pattern of chinese medicinal formulae (prescription) in treating and preventing breast cancer recurrence and metastasis. Journal of Translational Medicine, 10(1), 1-8.

[6] Zhou, X. , Chen, S. , Liu, B. , Zhang, R. , Wang, Y. , & Li, P. , et al. (2010). Development of traditional chinese medicine clinical data warehouse for medical knowledge discovery and decision support. Artificial Intelligence in Medicine, 48(2-3), 139-152.