Problems and Countermeasures for the Data Mining of Medical Records of Traditional Chinese Medicine in the Era of Big Data

Xiaofang Wang, Xiaojie Hu and Cheng Liu*

School of Humanities, Jiangxi University of Traditional Chinese Medicine, Nanchang 330004, Jiangxi Province, China

*Corresponding author e-mail: jxjzlc@jxutcm.edu.cn

Abstract. The medical records of Traditional Chinese Medicine (TCM) such as Huangdi’s Canon of Medicine and Treatise on Typhoid and Miscellaneous Diseases have always been the hot research objects in the field of TCM. Many papers on the summary and statistical analysis of TCM medical records have appeared one after another, which has played a huge role in promoting the development of TCM. However, with the progress of the times, the research of Traditional Chinese Medicine should keep pace with the new times. The authors in this paper introduced the research status and role of the data mining technology and mining methods, the problems and countermeasures of data mining technology in the research of Traditional Chinese Medicine medical records at this stage.

Keywords: Traditional Chinese Medicine, Medical Records, Data Mining, Problems and Countermeasures

1. Research Status of the Informatization of TCM Medical Records

The establishment of database can improve the efficiency of consulting medical records, and it is also the basis of intelligent medical record content statistics. At present, the existing medical record database construction includes: the Traditional Chinese Medicine medical record database established by Beijing Zhongda Anxin Technology Development Co., Ltd. is one of the 10 subsystems of the "Traditional Chinese Medicine basic database system" developed by the company [1]; On the basis of the query system of "Wu Jin" medical records, the query system of "Wu Jin" medical records is developed, which is convenient for the analysis of the topic words in the medical records of Shanghai University[2]; Zhang Qiming of Shandong University of Traditional Chinese Medicine, etc. standardized the terms of Traditional Chinese Medicine by using the terms of Traditional Chinese Medicine to form a "database structure key". With the help of self-made medical record recording program, they recorded the medical record data, and established the database of Traditional Chinese Medicine medical records of past dynasties. They also made statistics on 332 common clinical symptoms and 110 common clinical syndromes, and wrote the statistical diagnosis of Traditional Chinese Medicine. It opens up a new way of TCM diagnosis [3].
With the continuous emergence of TCM medical records, how to sort out and grasp a large number of medical records and find the hidden knowledge and rules is an important issue in medical records research. Mining and utilizing the internal knowledge of medical records, combining with mathematics, statistics and data mining methods, to explore some TCM clinical diagnosis and treatment rules from medical records and provide clinical diagnosis and treatment reference [4]. Some experts put forward that data mining technology (such as descriptive analysis, cluster analysis, association analysis, etc.) can be applied to the collation of the rules of syndrome differentiation and treatment, the characteristics of doctors, medication experience, etc. contained in medical records, and believe that the implementation of medical records collation and mining can promote the development of Traditional Chinese Medicine theory and clinical practice [5]. Wang Guoyin and others believe that the core of the existing database in Traditional Chinese Medicine is not query, but analysis. Only through the analysis and mining of a large number of data in the database, can we find the scientific components. This is the result of statistical analysis, not experience or intuition. Its epoch-making significance lies in the analysis of some clinical data. The database of bed makes the empirical components of Traditional Chinese Medicine more scientific [6].

2. The Value of Medical Records and Data Mining Technology
Data mining is a process of extracting information and knowledge from a large number of, incomplete, noisy, fuzzy and random data that people don't know in advance but are potentially useful. It is a key step of knowledge discovery [7]. At present, the application of data mining technology in TCM literature related research involves syndrome, diagnosis, prescription, compatibility of Traditional Chinese Medicine, academic thought and so on. The medical record is the objective record of diagnosis and treatment of diseases. It is the essence of the experience of famous doctors in past dynasties of China. It is a concentrated embodiment of its profound and extensive knowledge theory and profound clinical skills. Recently, many researchers have applied data mining technology to the study of TCM medical records.

Data mining technology is an intelligent database technology, which integrates artificial intelligence, pattern recognition, fuzzy mathematics, database, mathematical statistics and other technical methods, and is specially used for massive data processing. It is not only the query and traversal of the past data, but also can find out the potential relationship between the data, and find the unknown relationship and law in advance. Its purpose is to let the data owner get very clear and useful results. Data mining technology is a massive and complex computing and statistics system, and TCM medical record database is a large system containing complex information. Database is one of the preferred methods to scientifically evaluate and classify such a system. After Database sorting out the medical records, it can be found that not only the medical records themselves have the elucidation of the theory of Traditional Chinese Medicine, but also the knowledge mining can find some hidden contents, which can also sublimate the theory of Traditional Chinese Medicine.

3. Application of Data Mining Technology in the Research of TCM Medical Records
3.1. Steps of Medical Record Information Mining
In the process of medical record information mining, we should try to do the following aspects: ①. Making clear the purpose and put forward the questions. Question is the key to retrieve the desired results, and it is also the first step to improve work efficiency. This can refer to the patient intervention control consequence (PICO) questioning principle in Evidence Based Medicine. In view of the special diagnosis and treatment process of TCM syndrome differentiation and the characteristics of the database itself, a more feasible questioning principle can be designed, but the principles and norms should be matched. ② Selecting information. Select the information suitable for data mining applications, such as doctor, disease name, prescription name, drug name, drug nature, taste, meridian and processing, symptoms, syndromes, disease location, etiology, pathogenesis, etc. ③ Processing information. a. Delete useless or duplicate information; B. unify the expression and type of
information. ④ Digging. Find out the unknown, useful and relevant information and integrate it into new data. ⑤ Explaining and evaluating the new results. ⑥ Application guides decision making. In addition, the results of data mining are obtained through the circular mining process, which can not be completed at one time. Because the result of data mining technology is the result of statistical analysis of a large number of medical records of Traditional Chinese Medicine, it has great universality of application.

3.2. Common Data Mining Methods

3.2.1. Association rule. Association rule mining is to discover interesting association or correlation between item-sets in data, and find interesting association from a large number of transaction records, which helps us better analyze transactions. To explore the regularity between the values of two or more variables, such as "simultaneous occurrence" or "from one object to another" is called Association. Association rule mining is to find out the hidden knowledge in the database through association analysis, which can be used to speculate the unknown problems according to the known situation. By using the method of association rules analysis, we can find the rules between symptoms and syndromes, between syndromes and prescriptions, and between symptoms and prescriptions in medical records.

3.2.2. Cluster analysis. It is also known as cluster analysis, which refers to the use of the principle of birds of a feather flock together, through the analysis of the degree of similarity between different variables, so that the variables with high degree of similarity are aggregated into one class, and the variables with low degree of similarity are aggregated into another class. So repeatedly, the variables with high degree of similarity (closeness) are aggregated into a smaller class, and the variables with low degree of similarity (alienation) are aggregated into a smaller class. In this way, we can form a classification system from closeness to alienation, from small to large, so as to express the relationship between variables. By dividing a large amount of disordered data into several categories, it is helpful to understand the rules in a large amount of data. In the research of TCM medical records, cluster analysis can avoid the subjective factors in the process of classification, reflect the research object objectively and accurately, and find its internal objective laws. Through the cluster analysis method, we can objectively analyze the core drug group in a large number of medical records to treat a certain disease and syndrome, so as to explore the ancient doctors' prescription and medication rules.

3.2.3. Factor analysis. Factor analysis is a method to analyze the main potential factors of influencing variables by analyzing the correlation of different variables, comprehensively classifying and simplifying the number of variables, and replacing the information contained in the original variables with fewer variables. Factor analysis groups the data according to the correlation, and the data in the same group has a high correlation, that is to say, the data in the same group is related to a comprehensive factor. Corresponding to the concept of Traditional Chinese Medicine, the data of the same group is the patient's related symptoms, and this comprehensive factor is a type of syndrome related to this group of symptoms. Through the method of factor analysis, the syndrome differentiation of Traditional Chinese Medicine medical records can be summarized.

3.2.4. Rough set. Rough set is a mathematical tool to study the imprecision, uncertainty and incompleteness of existing information or knowledge. The starting point of rough set is to divide the object into domains, and then determine the support of each part to a certain concept. The rough set method can be used to identify the symptom information in medical records, divide the syndrome types of diseases, and explore the thinking process of syndrome differentiation and treatment of doctors in each medical record. Sun Jijia et al. [8] used the data mining method of rough set and support vector machine to study the syndrome differentiation of 293 patients with cirrhosis of Traditional Chinese Medicine. The correct rate of syndrome differentiation was 71.3% - 84.4%.
4. The Existing Problems of Data Mining Technology in Traditional Chinese Medicine

A large number of researchers have gradually focused on data mining technology to explore the medical records of Traditional Chinese Medicine. At present, they have accumulated effective methods of data mining of medical records. At the same time, they found that data mining can fundamentally find the underlying basic laws of medical records and solve the individual problems existing in the theory and practice of Traditional Chinese Medicine. However, there is no systematic mining of medical records at this stage. The application of data mining in medical record research has the following problems: (1) the main content of the research is the analysis of diagnosis and treatment of diseases, while the basic research is rare, but it has gradually increased in recent years; (2) although some research results are consistent with clinical data, there are still a large number of conclusions inconsistent with the results of clinical practice; (3) the method of medical record mining is based on statistics; (4) there are few researchers engaged in medical case mining, and they have strong concentration.

In the future, we need to focus on the following two aspects: (1) comprehensively studying the characteristics of medical records, selecting targeted mining methods, so that the mining results can more directly reflect the hidden laws in medical records, and lay a good foundation for improving the level of clinical practice of Traditional Chinese Medicine; (2) focusing on the development of text mining in computer technology, so as to make more comprehensive use of the mining methods of clustering analysis, artificial neural network and so on.

5. Countermeasures

In order to solve the existing problems, interdisciplinary is needed. In the face of the dynamic characteristics of TCM data, this paper attempts to use the latest artificial intelligence technology to explore the mining of TCM medical record data, such as artificial neural network and case-based reasoning technology, to establish a more accurate, practical and efficient medical database, in order to achieve quantitative and qualitative analysis.

Artificial neural network is an abstract simulation of the information processing of human brain nervous system. It has been used in the early research of TCM data mining. With the progress of algorithm, it is more skillful in medicine. Because of the complexity and unpredictability of human physiology and pathology, the acquired data have very complex nonlinear relationships. Artificial neural network is very suitable for dealing with these complex nonlinear relationships. Artificial neural network has the ability of parallel processing, which can simulate the distributed storage of human brain, fault tolerance and adaptability. Its disadvantage is that it requires sufficient data sets, otherwise it will affect the accuracy. Zhao Liang et al. [9] applied BP neural network based on LM algorithm to TCM syndrome differentiation of epigastric pain, and the accuracy and diagnostic accuracy of predicting "disharmony between liver and stomach" and "deficiency of stomach Yang" syndrome types with enough cases were more than 95%.

Case based reasoning is a research hotspot in the field of artificial intelligence, which originates from cognitive psychology. The reasoning process simulates the way of human solving problems. That is to say, when meeting a new problem, choose an old case which is closest to the current problem from the previous case, and improve the solution of the old case according to the situation of the new problem It will be recorded in the database. Its characteristic is that it can combine quantitative analysis with qualitative analysis, and has dynamic database, which is exactly what is lacking in modern Chinese medicine case mining. Yang Li et al. [10] established the system of TCM syndrome differentiation and treatment based on case-based reasoning.

Mr. Zhang Taiyan, a modern sage, pointed out: "the achievements of Traditional Chinese Medicine are most remarkable in medical records. If we want to seek the experience of our predecessors, we can find the most clues in medical records. We can get twice the result with half the effort by studying them.” With the development of modern information technology, in the face of massive data in medical records, we should make rational use of these data, excavate the knowledge connotation contained in them, and explore the hidden rules. Through the study of medical records, we constantly
sum up the previous experience, explore their academic ideas, and dig out the rules hidden in a large number of medical records, so as to better serve the clinical and scientific research, and reflect the value of literature research.

6. Conclusion

To sum up, many medical workers have done a lot of work in the research of TCM medical records. Through the research, it is helpful to understand the academic origin of the predecessors, inherit and master the academic thoughts of the predecessors, summarize or draw the successful experience or failure lessons of the predecessors, so as to develop ideas and guide clinical practice. Although data mining technology has just started in the research of medical records, it is certain that the introduction of this new research method will greatly improve the research level of TCM literature (including TCM medical records), improve the utilization efficiency of TCM literature, and then improve the efficiency of TCM scientific research.

Acknowledgments

This work was financially supported by the fund of Research Project of Humanities and Social Sciences in Colleges and Universities of Jiangxi Province (No. YY19108).

References

[1] Ren Tingge, Liu Xiaofeng, Gao Jianbo, et al. Introduction of basic database system of traditional Chinese Medicine. Chinese Journal of traditional Chinese medicine information, 2001, 8(11):90-91.

[2] Wu Zhongping, He Xinhui, Ke Xuefan, et al. Introduction to medical records query and statistical analysis system. Shanghai Journal of Traditional Chinese Medicine, 2003, 37 (3): 54-56.

[3] Zhang Qiming, Wang Yongyan, Zhang Zhibin, et al. Establishment and statistical methods of medical records database of Traditional Chinese Medicine. Journal of Shandong University of Traditional Chinese Medicine, 2005, (29): 298-299.

[4] Hong Yanzhu, Zhou Changle, Zhang Zhifeng, et al. Research progress of medical records of Traditional Chinese Medicine. Bulletin of Traditional Chinese Medicine, 2008, 7 (8): 62-65.

[5] Sun Yan, Zang Jiaxin, Ren Tingge, discussion on medical records sorting method based on data mining technology. Chinese Journal of Traditional Chinese Medicine information, 2006, 13 (11): 106-107.

[6] Wang Guoyin, Gu Renyue. Establishment and mining of TCM medical records database. China Medical Journal, 2007, 2 (4): 55-56.

[7] Margaret H. Dunham. Data Mining Tutorial. Beijing: Tsinghua University Press, 2008. 8.

[8] Sun Jijia, Su Shibing, Lu Yiyu, et al. Research on data mining method of TCM syndrome differentiation based on rough set and support vector machine. Journal of Mathematical Medicine, 2010,23(3): 261-265.

[9] Zhao Liang, Zhang Ye, Cao Yue, et al. Study on BP neural network syndrome differentiation model of epigastric pain based on Levenberg Marquardt algorithm. Journal of Chengdu University of traditional Chinese medicine, 2018, 41 (2): 97-101.

[10] Yang Li, Zhou Xuezhong, Bi Lanxin, et al. Decision support system for clinical diagnosis and treatment of traditional Chinese medicine based on case-based reasoning. World science and technology modernization of traditional Chinese medicine, 2014, 16 (3): 474-480.