

## Research on Data Mining in Computer Aided Medical Diagnosing System

Keke He<sup>1</sup>, Junfeng Jia<sup>2</sup>,

<sup>1</sup> School of Information Engineering, Xi'an University, Xi'an, 710065, China

<sup>2</sup> Xijing Hospital, the Fourth Military Medical University, Xi'an, 710032, China

**Keywords:** Computer aided medical diagnosing, Data mining, Medical images

**Abstract:** In the computer aided diagnosis system, the use of data mining in medical database is of great significance. This paper analyzes the characteristics of medical diagnosis system and expounds the key technology of data mining in medical diagnosis system. Based on the above analysis, the paper explores the application process of data mining in medical diagnosing system, including data collection, standardization, integration and disease diagnosis to provide some references for the relevant researchers.

### 1. Introduction

Data mining is a new kind of new information technology developed in recent years with the development of database and artificial intelligence technology [1]. It is from the practical application of a large amount of data, incomplete, noisy, fuzzy, in which the extraction of implicit, people do not know in advance, but is potentially useful information and knowledge, it can be highly automated analysis of the original data warehouse data, make inductive reasoning, dig out the potential of the model, the maximum efficiency of the use of existing information and data. Data mining can generally be an atmosphere of three stages. Data integration, data selection and data transformation can be divided into one part, which is a data preprocessing stage. The effect of data preprocessing will directly affect the quality of data mining. The discovery of patterns is the intermediate stage, is extracted from the database by using the phase of pattern data mining algorithm, the last phase is the subsequent processing stage, it includes the pattern evaluation and knowledge representation of two parts, the subsequent processing stage through the standards have been developed to identify good useful knowledge. The medical information system has greatly promoted the digitalization of medical data, and brought great place to the analysis of medical data in the future. But at the same time, it also brings trouble. We must know that medical data is very huge. In order to improve efficiency, we must get valuable information related to these data in a short time. We put the application of data mining in medical data analysis, so that doctors can get the most valuable information to analyze the medical data in the medical information system in the large in a short time, it will not only improve the efficiency of doctors, but also allow doctors to determine more accurately [2].

### 2. Data Characteristics of Medical Diagnosing System

Medical data mining technology helps to improve the efficiency of data analysis and increase the speed of generating new knowledge. Medical database is a huge information base. There are huge amounts of data and data in various forms. We need to find valuable information in such data. Because of the limitations of medicine, the medical data produced by clinical patients may involve patients' privacy. If these privacy is not allowed by the patients, the public will cause invasion to some extent. This requires data diggers to protect the security and confidentiality of medical data when they are working. Medical data contain many things, but not only words and images, there will be many voices, signals and so on. This is not the other information. The multi-attribute mode of medical information has brought a lot of trouble to data mining. The variety of medical diseases, and the limitations of the case, the medical information system cannot reflect all the information of any

kind of disease. At the same time, with the limitations of the recording itself, many medical data cannot be well recorded, which leads to the incomplete medical information. There are many time functions in medical data, some are delay, some are ahead of time, and some are the time functions of the waveform. At the same time, there are still some static data. Although these static records are not timed, this part of medical records must be detailed to a certain time when recording. As we mentioned earlier, the number of medical data is huge, it is similar to that of different diseases sometimes by medical means the same, it will have the same lot of medical data is recorded on the medical information system, which will give future data to find the inconvenience [3].

### **3. Key Technologies of Medical Diagnosing System**

#### **3.1 Neural network**

In data mining, the improvement of neural network is focused on solving two problems: knowledge expression and knowledge acquisition. Knowledge expression is to make the abstract weights in the neural network represent certain knowledge. Knowledge acquisition is a given neural network that has been trained to extract explicit knowledge from it. Neural network is an adaptive function estimator does not depend on the model, its outstanding advantage is capable of parallel processing, and has the learning ability, adaptability and strong fault tolerant ability, can construct the nonlinear model, the model has high accuracy, better generalization of the unknown input, you can also get the right output, can accept different kinds of variables as input, can be applied to the field of a wide range of model construction ability, has the ability of fuzzy inference, allow the output to input variables is fuzzy. Based on the research of theoretical models, a specific neural network model is constructed to achieve computer simulation or preparation of hardware, including the research of network learning algorithm. This work is also called technical model research. The algorithm used by neural networks is vector multiplication, and the symbolic function and its various approximations are widely used. Parallel, fault-tolerant, hardware implementations, and self-learning features are a few of the basic advantages of neural networks [4].

#### **3.2 Association rules**

The mutual dependence of association rules are used to reveal the unknown data, his task is: given a transaction database, based on the support / confidence framework, a number of interesting relationship was found between data and project, generate all association rules support and confidence were higher than the user specified minimum support and minimum reliability. The first phase of association rules mining must be found from the original set of data to find all the high frequency project groups. High frequency means that the frequency of a project group appears to be at a certain level compared to all records. The second stage of association rule mining is to produce association rules. The association rules generated from the high-frequency project group are generated by using the previous high-frequency project group. Under the threshold of the minimum confidence level, if the trust degree obtained by a rule satisfies the minimum confidence level, it is called the association rule. The data in the association rules can be divided into single and multidimensional. In a single dimensional association rule, we only relate to one dimension of the data; in multidimensional association rules, the data to be processed will involve multiple dimensions. A single dimension association rule is a relationship in dealing with a single attribute; multidimensional association rules are some relationships between the processing of various attributes.

#### **3.3 Cluster analysis**

Clustering analysis is a grouping of samples to find differences in multidimensional data points. It is different from the discriminant analysis. The classification way of clustering analysis does not need to set a pointer variable in advance. It belongs to a nonparametric analysis method, so there is no very rigorous mathematical basis and no need to assume that the population is normal distribution. The

general process of clustering analysis is to collect data and convert it into a similar matrix. From a machine learning point of view, a cluster is equivalent to a hidden pattern. Clustering is an unsupervised learning process of a search cluster. Different from classification, unsupervised learning does not rely on pre-defined classes or training instances with class tags. Automatic marking is needed by clustering learning algorithm, while classified learning instances or data objects have category labels. Clustering is an observational learning, rather than an example type of learning. Cluster analysis can be used as an independent tool for obtaining data distribution, observing the characteristics of each class and further analysis of a particular class. By clustering, it can identify the dense and sparse regions, find the global distribution pattern, and the relationship between the data attributes.

#### **4. Application Process of Data Mining in Medical Diagnosing System**

##### **4.1 Data collection and standardization**

Biomedical database is a complex database, including electronic medical records, medical images, pathological parameters, laboratory results, and so on. At present, data mining technology is mainly applied to relational databases, transaction databases and data warehouses based on structured data. Mining of complex type data is still in its infancy. While collecting and recording the data of the system, the dictionary of key words for the clinical symptoms of chronic diseases is designed to standardize it. The clinical manifestations and symptoms of chronic diseases are different from each other. Large amounts of data cause massive information. The basic conditions of patients were collected, such as gender, age, weight, family history, disease symptom classification, including simple diseases and complex diseases. The basic principle is grading and grading. That is to set up a set of general standard criteria to quantify data collected by users. For example, when describing the symptoms of hypertensive patients, they should be collected according to the medical hypertension classification standard instead of light, medium, heavy, high and high fuzzy words. In the data collection and input, there are two interactive ways. One is an active type, has the advantages of flexibility, directly by the user input the symptoms; the other is a guide, has the advantages of high efficiency of gathering input, but the flexibility and expansibility, the system provides is an ordered set of symptoms of the auxiliary information list for the user to choose. In the design of the system, two kinds of input forms are adopted, combined with each other, complemented each other. Guided by typical symptoms, active information is fuzzy, which is conducive to the flexible expansion and operation efficiency of the system. The purpose of data selection is to identify the data set that needs to be analyzed, to reduce the range of processing and to improve the quality of data mining. Preprocessing is to overcome the limitations of current data mining tools.

##### **4.2 Data integration and data warehouse**

The correct diagnosis is obviously important for guiding the medication and rehabilitation of patients. In clinic, some diseases are complicated. The classification analysis of data mining can be applied to the diagnosis of diseases. The integration of diagnosis and treatment information of chronic diseases is mainly from the hospital to collect the data of chronic disease diagnosis and treatment of thousands of medical records, and to form a database for diagnosis and treatment of chronic diseases. The main contents are from the data of clinical diagnosis and treatment. In the past, the establishment of database is subjective, and we advocate the use of objective information such as clinical medicine and evidence-based medicine to build database. The list of disease data is constructed by a two-dimensional table, which is beneficial to the storage of different kinds of disease information. A disease should correspond to a set of clinical information, such as symptomatic sweating, pain, fever, etc., such as white blood cells, bilirubin, carcinoembryonic protein and so on. The current medical multimedia data mainly comes from some imaging instruments in the hospital. It promotes the development and application of medical image archiving and communication system, making it possible for hospitals to manage and use medical images from different devices centrally

and uniformly. Data mining is a systematic theory of the latest achievements of data processing technology, especially for the multidimensional data such as medical image data analysis. A large number of patients and patients in the medical record information in the database of personal information, including age, gender, residence, occupation, living conditions, analysis of association rules can be found between meaningful patterns and the information in the database, the related risk factors of a disease analysis can instruct patients how to prevent the disease.

#### **4.3 Data mining and disease diagnosis**

Medical diagnosis is a sequential diagnosis problem based on knowledge. Doctors acquire knowledge through certain ways, form inference network, and case data are stored in database. Therefore, how to extract diagnosis rules from case database has become the theme of research. For example, using association rules to find indications of head trauma patients for examination and data mining for automatic detection of liver cancer genetic syndrome, etc., all show that data mining technology has broad application prospects in disease diagnosis. Data mining can extract valuable diagnostic rules by processing a large number of historical data in patient database, so that we can make diagnostic conclusions based on patients' age, gender, physiological and biochemical indicators, and exclude the interference of human factors. In addition, because of the large amount of data processed, the diagnosis rules have good application universality to diagnose by the results of the data mining library. It is specific to four modules. The diagnostic module mainly entered the patient's name, gender, age, family history, personal history and current medical history. According to the main complaint symptoms, the diagnosis ranked the first three diagnoses which were closest to the disease. First of all, exclude other diseases, ask patients for past medication and clinical symptoms. According to the diagnostic content of the evaluation, we systematically retrieve and identify the drug decision information that accords with the disease. The premise is the result of diagnosis after the diagnosis and prediction. Check the inspection module. For example, diagnosis is made by various auxiliary examinations of coronary heart disease, including electrocardiogram, two steps, flat plate test, cardiac pacing test and so on. We collect the examination results of patients, systematically evaluate the top three diagnostic contents, and give medical records at the same time. The effect evaluation module system combined with symptoms, signs, examination results and dosage to track chronic disease patients after a period of time, regularly give treatment quality evaluation and prediction.

#### **5. Conclusion**

Data mining technology has a broad application prospect in medical diagnosis system. With the popularization and application of large databases and network technology, there will be a large number of electronic format data appearing in all walks of life, especially in the medical sector. The research and development personnel of the information profession should seize the opportunity to prepare for the technology and meet the challenges.

#### **Acknowledgement**

The paper is the results of Science and Technology Planning Project of Xi'an City (Grant No. 2017CGWL36).

#### **References**

- [1] Shi Xiaojing. Data Mining and Its Application in Medical information [J]. Journal of Medical Informatics, 2013, 34(5): 2-6.
- [2] Yao Yuan, Liu Lihua. Construction of Chronic Computer-aided Diagnosis System [J]. Chinese Journal of Health Informatics and Management, 2014, 11(2): 130-133.

- [3] Qin Wenzhe, Chen Jin, Dong Li. Progress and Application of Medical Data Mining under the Background of Big Data [J]. Chinese Journal of Clinical Thoracic and Cardiovascular Surgery, 2016, 23(1): 55-60.
- [4] Mu Dongei, Feng Chao, Wang Ping. Applications of Data Mining in Medical Field as Well as SWOT Analysis [J]. Journal of Medical Informatics, 2015, 36(1): 53-57.