Research Progress and Application of Medical Data Mining under the Background of Big Data

Keke He[1], Junfeng Jia[2]

[1] School of Information Engineering, Xi'an University, Xi'an, China

[2] Xijing Hospital, Air Force Medical University, Xi'an, China

Keywords: Big Data Area, Health Progress, Data Mining

Abstract: the Advent of the Era of Big Data Has Brought about Major Changes in Life, Work and Thinking. in the Medical Field, Since the 21st Century, with the Development of Social Economy and Medical and Health Progress, the Spectrum of Human Diseases is Changing, the Types of Diseases Are Increasing, and the Complexity of Etiology, Diagnosis and Treatment is Gradually Increasing. in Order to Improve the Health of Human Beings and Explore the Law of the Occurrence and Development of Diseases, We Need to Constantly Explore and Discover the Hidden Laws from the Vast Amount of Knowledge through the Full Exploitation and Utilization of Medical Big Data. This Will Play an Important Role in Improving the Level of Medical Information Management, Providing Theoretical and Methodological Support for the Diagnosis and Treatment of Diseases, and Promoting Clinical Practice and Decision-Making.

1. Introduction

In the context of the era of big data, medical information contains a wealth of information, mainly related to individual health or disease state, not only the amount of data is large, but also reflects the complexity and diversity of data. Medical big data involves multidisciplinary information such as pathology, diagnosis, treatment methods, and imaging. The complexity is self-evident. In the process of acquisition, structured and semi-structured data such as text, video, image, and web page will be generated. There are often some complex connections, which undoubtedly add to the difficulty of medical big data analysis. The medical information includes personal information about the patient's personal information, medical information, and other medical data related to privacy, as well as confidential medical research information. When analyzing and researching these data, it is necessary to consider the security of the information to ensure that the data is not leaked. At the same time, the era of big data has also brought new ethical controversies, such as the debate on the specificity of genetic information and the dispute over the right to know. It is urgent for the state to establish a rules system to regulate the research of medical information big data.

2. Characteristics of the Big Data Era

Big data is an abstract concept. Although its importance has been recognized, its definitions vary from field to field. Currently, there are several types of recognition: Wikipedia believes that big data means that it is difficult to manage with existing databases. Data integration that combines massive features and complexity with tool processing. In general, big data is a collection of data that cannot be perceived, acquired, managed, processed, and serviced by traditional IT technologies and hardware and software tools in a tolerable time. Apache Hadoop points out that data sets that cannot be captured, managed, and processed within the acceptable range of a general-purpose computer. Based on this definition In May 2011, the McKinsey Group announced that big data is at the forefront of innovation, competitiveness and productivity. The IDC Report states that “Big Data Technology describes a new generation of technologies and architectures that extract valuable information from a wide variety of data by enabling high-speed capture, discovery and analysis

  DOI: 10.25236/iiicec.2019.033

techniques. NIST (United States) According to the National Institute of Standards and Technology, “Big data refers to the amount of data, the speed of acquisition, or data that cannot be analyzed effectively using traditional methods, or can be effectively processed with important horizontal amplification techniques. Data, it focuses on the technical aspects of big data. In addition, there are quite a few other discussions on the definition of big data in industry and academia. In short, no matter which definition, big data is not a new product or New technology, it is only a phenomenon or feature that emerged in the digital age. What we should pay most attention to is not its definition, but the value it brings. Compared with other existing technologies, big data has “cheap, rapid, optimized”. The advantage of increasing the amount of data for human analysis and use by analyzing and storing massive amounts of data; The exchange, integration and analysis, can help people discover new knowledge, create new values, to bring “great knowledge” and “big development.”

According to the structural state classification of digital collections, big data can be divided into: (1) structured data, such as enterprise financial systems, personnel systems; (2) semi-structured data, such as e-mail, HTML pages; (3) unstructured Data, such as video, mobile terminals, sensors, etc. According to the application type of data, big data can also be divided into three types: massive transaction data, massive interaction data, and massive processing data that combine transaction and interaction data for processing. According to the source of the data, big data can be divided into three categories: administrative record data, business record data, Internet and search engine data.

The characteristics of big data are generally summarized as four “V”s: (1) volume is “large volume”, jumping from terabytes to petabytes (1PB is equivalent to 50% of all US library collections), along with data The massive generation and collection of data has become larger and larger, and has gone beyond traditional storage and analysis techniques. (2) Velocity is the “speed”, which is the timeliness of big data, which means that the collection and analysis of data must be carried out quickly and in a timely manner. It is generally required to give an analysis result in the second time range to maximize its value. (3) Variability is “a wide variety of data types”, including semi-structured and unstructured data such as audio, video, web and text, and traditional structured data. (4) value is “value”, which is concentrated in the low value density and high commercial value. Big data is mostly unstructured and semi-structured data. It takes too much time and money to analyze. Take the monitoring video as an example. In the continuous monitoring process, the useful data may only be one or two seconds. In recent years, based on the four “V” features, the three methods of data acquisition and transmission, such as vender, veracity and complexity, have been added. The key question when using big data is how to find value from a large, rapidly generating, and diverse data set. If you don't take advantage of the collected data, you can only have “a bunch of data” instead of “big data.” It can be seen from the definitions of the definition, type, value and characteristics of integrated big data that big data has the characteristics of early warning, predictive, differentiating, sharing and dynamic.

3. Medical Big Data Application Field

Precision medicine. Precision medicine is an important application area for big data in the field of clinical medicine. Precision medicine is based on personalized medicine and is part of clinical translational medicine. It will integrate multiple post-study techniques, second-generation sequencing technology, genomics, computer biology analysis, medical informatics, and clinical informatics. Multidisciplinary domain big data resources. The existing medical big data can provide the necessary data guarantee for precision medicine to support the large sample population, disease characteristics, gene sequencing, etc. required by the precision medical research institute, and then comprehensively consider various pathogenic factors to accurately lock the cause and Therapeutic targets will ultimately achieve personalized and precise treatment of diseases and patients.

Disease warning and prevention. As early as 2009, engineers used medical big data technology to successfully predict the spread of H1N1 throughout the United States. In 2014, when the Ebola outbreak broke out, the medical aid agencies quickly analyzed the geographical location of the epidemic area and predicted the spread of the virus by analyzing the communication data of the

local residents, which provided the first hand for the later medical resource allocation, material allocation and rescue route formulation. data. The well-known medical technology company CardioDX discovered a genome-wide association analysis of 100 million gene samples in 2010, and found that 15 nucleotide polymorphisms were associated with lipid metabolism, and found that 6 of the 15 gene fragments were found. In addition to affecting lipid metabolism, it also has a significant impact on glucose metabolism. Another example is the analysis of the data of a large number of sample populations, and found that there is a certain proportional relationship between the incidence of diabetes and the waist circumference value. When the waist circumference increases, the risk of the disease increases, and the main risk factor for diabetes is abdominal obesity. In this way, the prediction of the risk of diabetes can be achieved by the three abdominal obesity indicators of waist circumference, waist-to-hip ratio and waist height ratio, and enhance the risk resistance ability of high-risk population.

The diagnosis of current diseases relies mainly on the doctor's professional ability and diagnostic experience. However, the doctor's experience is limited, especially for young doctors. The experience needs to be accumulated. Therefore, the diagnosis results of the disease have certain subjectivity. There are certain uncertainties in the treatment plan. On the one hand, the doctors work hard, the doctor-patient relationship is tense, and on the other hand, the medical expenses are high. Using data mining methods such as decision tree, support vector machine, neural network, machine learning, etc., it is possible to fully exploit medical information, obtain the correlation between patients and diseases, and obtain a complete mapping between appearance and cause, so as to accurately construct an auxiliary diagnosis system. To assist the doctor's daily diagnosis and treatment. Not only that, but also further correlate the etiological diagnosis information of various departments of the hospital, and effectively eliminate the island effect of hospital diagnostic information, realize the informationization and scientificization of medical diagnosis, and improve the accuracy of disease diagnosis and the level of medical services. Further reduce the rate of misdiagnosis and diagnosis.

4. Basic Steps of Data Mining

They create a mining model A mining model is a container for storing patterns mined by mining algorithms. There are many input columns, predictable columns, and related algorithms in the definition of the mining model. 2 Model Training (Mode Processing) The collected data is provided to the data engine, and the mining algorithm analyzes the input data, finds rules between attribute values to extract patterns, and stores the patterns in the mining model. 3 Model Prediction The data mining engine applies the rules found during the training process to the new data set, predicts the predictable columns of each new case, and assigns the prediction results to each input case. In the data mining process, each model is created in association with a data mining algorithm, and the patterns in the data set are discovered by using specified data mining algorithms and appropriate algorithm parameter values. Explore the sources of current popular data mining algorithms, mainly from the fields of statistics, machine learning, and databases. The decision tree uses the attributes of the collected data as nodes and uses the values of the attributes as branches to analyze and summarize large amounts of data. The decision tree is composed of decision nodes, branches and leaves. The root node is the most informative attribute of all data. The intermediate node is the attribute with the largest amount of information in the data subset contained in the subtree rooted at the node. The leaf node of the decision tree is the category value of the data. Each leaf node represents a possible classification result, and each path from the root junction to the leaf node is a rule. Typical decision tree algorithms include the ID3 algorithm and the C4.5 algorithm based on the ID3 algorithm.

In the 18th century, British scholar Bayes proposed a formula for calculating the conditional probability to solve the following problem: Hypothesis H[,1], H[,2]... mutually exclusive and constitute a complete event, the probability of which is known P(H[,i],i=1,2,..., it is observed that an event A occurs with H[,1], H[,2]..., and the conditional probability P(A/H) is known. [,i]), find P(H[,i]/A), and use Bayes' theorem.

Bayes' theorem: P(H[,i]/A)=P(H[,i])P(A ｜ H[,i])/[P(H[,1])P(A ｜ H[,1]) +P(H[,2])P(A ｜ H[,2])+...]

Bayes' theorem is used for predictive modeling. Assume that the relevant project B information is known, and the direct data of the demonstration project A is lacking. The state and probability of occurrence of the A project are derived by analyzing the relevant state and probability of occurrence of the B project. That is, when the probability P(Bi) of the event Bi and the probability P(A ｜ Bi) of the event A under the condition that the event Bi has occurred are known, the Bayes' theorem can be used to calculate the probability P of the event Bi under the event A. (Bi ｜ A), expressed as:

$$P(B_i \mid A) = \frac{P(B_i)P(A \mid B_i)}{\sum_{i=1}^{n} P(B_i) \cdot P(A \mid B_i)}$$

Cluster Analysis Algorithm The class of data objects processed by cluster analysis is unknown. The clustering algorithm relies on guesses and assumptions, guesses the grouping of data, and groups the collection of objects into multiple classes of similar objects. Based on the initial hypothesis, the algorithm will calculate the extent to which the classification fits the real world, and then attempt to re-assume the grouping to create a classification that better represents the data, and the algorithm will loop through the process until it can no longer redefine the classification. To improve the results.

The neural network consists of a set of nodes (neurons) and edges. There are three types of nodes: input, implicit, and output. Each edge connects two nodes through an associated weight. The direction of the edge represents the prediction process. The data stream. Input neurons define all input attribute values and their probabilities of the data mining model; implicit neurons receive input from input neurons or previous implicit neurons, assign weights to various input probabilities, process some calculations, and output to the neural network The element provides the output; the output neuron represents the predictable attribute value of the data mining model. The core process of neural network model training is: 1 algorithm randomly assigns values to the ownership values in the network at the initial stage (range usually from -1.0 to 1.0). 2 For each training case (or each set of training cases), it calculates the output based on the current weights in the network. 3 Calculate the output error, then the backpropagation process calculates the error for each output and implicit neuron in the network, and the weights in the network are updated. 4 Repeat the steps until the conditions are met.

5. Data Mining Technology

Data mining is so powerful because it stands on the shoulders of giants from birth. One shoulder is the prosperity and rapid development of modern computers and related information processing technologies. Macroscopically, the proposed and universal acceptance of data mining technology is due to the feasibility of research and application of the development of computers and related technologies. Microscopically, the technical background generated by data mining mainly includes the development of information technology such as database, data warehouse and Internet; the improvement of modern computer performance; the research and application of statistical science and artificial intelligence in data analysis.

The development of computers has led to an increase in the processing and storage capabilities of computers. After decades of development, computer architecture, especially parallel processing technology, has matured and been widely used, and has become the basis for supporting large-scale data processing applications. The improvement of computing performance and the development of advanced architecture make the research and application of data mining technology possible. Relevant theoretical and technical achievements such as statistics and artificial intelligence have been successfully applied to commercial processing and analysis. These applications have greatly promoted the development and development of data mining technology to some extent. The core technologies and algorithms of data mining systems are inseparable from the support of these theories and technologies. The development and application of these theories themselves provide

valuable theoretical and application accumulation for data mining. Mathematical statistics is one of the most important and active disciplines in applied mathematics. Today's powerful and effective mathematical statistics methods and tools have become the basis of the information consulting industry. Artificial intelligence is one of the important research fields in computer science. The expert system used to be the brilliance of artificial intelligence. The fundamental of this method is based on the experience world created by experts in a certain field. Its information comes from the thinking activities of the human brain. But the expert system is subjective knowledge, which is inevitably biased and wrong, thus limiting the application of expert systems. Data mining inherits the highly practical features of the expert system, and uses the data as the basis to objectively mine knowledge. Machine learning has been fully researched and developed, but people have not satisfied the small sample learning model, and turned to a large number of incomplete, noisy, fuzzy, random big data samples in life, which is data mining. The specialty is where it is.

Data mining involves using a variety of algorithms to accomplish different tasks. All of these algorithms attempt to build a suitable model for the data. An algorithm is used to analyze the data and determine the model that best matches the characteristics of the analyzed data. Data mining models can be divided into two types: predictive models and descriptive models. The predictive model predicts the value of the data. Data mining tasks that predictive models can accomplish include classification, regression, time series analysis, and prediction. A descriptive model identifies patterns or relationships in the data. Unlike predictive models, descriptive models provide a way to explore the nature of the data being analyzed, rather than predicting new properties. The mining tasks of descriptor models are mainly clustering, summarization, association rules and sequence discovery. The main tasks of data mining are as follows:

Classification refers to mapping data to pre-defined groups or classes. Since the category has been determined before analyzing the test data, the classification is also referred to as guided learning. Classification algorithms require that categories be defined based on data attribute values. It usually describes the category by observing the characteristics of the data of the class it knows. Pattern recognition can be considered as a classification problem. The input mode is divided into classes based on its similarity to pre-defined categories.

Regression refers to the mapping of data items to an initial value predictor. Regression is learning a function that can do this mapping. Regression first assumes that some known functions can fit the target data and then use error analysis to determine a function that best fits the target data.

The data attribute values of time series analysis are constantly changing over time. The data is obtained at equal time intervals. Time series data can be visualized by time series diagrams. In Figure 1, it is easy to see that the Y sequence is similar to the Z sequence, they all have some variability, and X is relatively stable. Time series analysis has three basic functions. First, the distance metric is used to determine the similarity of different time series; second, the structure of the line in the time series graph is tested to determine the behavior of the time series; and third, the historical time series map is used to predict the future value of the data.

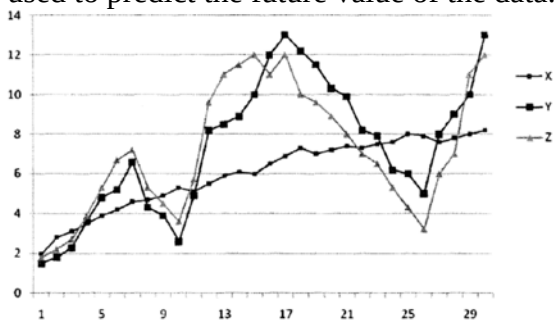

Fig.1 Time Sequence

Forecasting requires forecasting future data states based on past and current data. Forecasting can actually be seen as a classification. The difference is that the prediction is mainly to predict the state of future data rather than the current state. Clustering refers to unsupervised learning or segmentation. Clustering can be thought of as the process of dividing or splitting data into

intersecting or disjoint groups. The clustering task is accomplished by determining the similarity between the data on pre-specified attributes.

Summary is the mapping of data to a subset with a simple description. Summary is also called characterization or generalization. The summary extracts or obtains representative information from the database, which can be done by retrieving part of the data, or getting some summary information from the data. Summary can succinctly characterize content in the database. Association refers to the existence of a relationship between the values of two or more variables, and this relationship is not directly represented in the data. The purpose of association analysis is to find out the associated networks in the database.

6. Medical Data Mining Applications in the Context of Big Data

The early warning of big data lies in the fact that when the data is abnormal, a warning can be issued through a certain mechanism, so that the corresponding measures can be quickly taken to solve the problem in time. Teng Qi et al. designed and developed a healthy cloud platform by using emerging cutting-edge cloud computing technologies. Distributed cloud storage technology is used to store large-scale heterogeneous multi-modal physiological signal data, and the data mining model (L1-Logistic) is integrated into the MapReduce framework to quickly mine user health information and high-risk factors of major diseases, so that users can Real-time understanding of your physical condition, while giving early warning information to the user's abnormal conditions, and informing them to go to the hospital for medical treatment, to achieve early warning of major sudden illness. MIT, the University of Michigan, and a women's hospital created a computer model to analyze ECG data from heart patients and predict the risk of heart disease in the next year. Through machine learning and data mining, the model can analyze the accumulated data and find high-risk indicators, which changes the past doctors' lack of pre-judgment on 70% of patients with heart disease due to lack of comparative analysis of previous data. Phenomenon. As Viktor Mayer-Schönberger puts it: “Forecast, the core of big data”. Data mining is also used more in predictive modeling in clinical practice, using patient-specific information to predict disease outcomes, aiding disease diagnosis and recommending treatments to support clinical decision making. Predictive modeling mainly uses the method of independent variables to model the target variables, including two modes: classification and regression. Classification is the prediction of discrete data. In clinical medicine, the diagnosis of disease is a typical classification process. Liu Juan studied and discussed three classification prediction algorithms (C5.0, BP- artificial neural network and TAN Bayesian network) for data mining, and constructed a suitable model for early warning and diagnosis of gastric cancer. And automation of classification. Regression is mainly to predict continuous and ordered data, and can be widely used in disease diagnosis, prognosis and drug dose prediction. For example, Consortium et al. They used a least squares regression model to establish a warfarin dose prediction algorithm to predict the stable maintenance dose of warfarin. Using online search records and search terms that are closely related to the flu, Google has built specific systems and 450 million mathematical models to predict the spread of the flu and even predict where the flu will occur. After comparing the predicted results with the actual flu cases recorded by the US CDC, their predictions are 97% relevant to official data. This kind of prediction is based on big data. This is a new type of capability unique to today's society. In an unprecedented way, through the analysis of massive data, we can get great value products, services and profound Insights.

The diversity of big data highlights the personalization of medical services. Gene sequencing is the representative of medical service personalization. Bina Technology uses big data to analyze human gene sequences and discover rare lesion information in genes. As more and more genetic information is obtained from genetic sequencing, this technology will have a significant impact on our health. The continuous development of gene sequencing technology has promoted the emergence of new types of disease treatment measures such as personalized drug development. Apple's president, Jobs, sorts all his DNA and tumor DNA in the fight against cancer, allowing doctors to use the specific effects of his specific genetic makeup. If the cancer lesion causes the drug to fail, the doctor can change another drug in time. His life has been extended for several years

by personalizing medication. The UK Health Care Authority announced that it will establish the world's largest database of cancer patients to provide a foundational support for personalized cancer treatment. The purpose of this database is to promote “personalized medicine” that is symptomatic for each patient's cancer category and specific circumstances. Data from cases of medical institutions across the UK and 11 million historical records, and sharing information with health care databases in Wales, Scotland and Northern Ireland.

Data sharing is the cornerstone of big data applications and complements the “four V features” of big data. Through information sharing, each information island is connected to maximize the amount of data and provide data support for more and newer applications. Users can access more types and more time-series data content, provide a more reliable basis for analysis and decision-making, greatly speed up information circulation, increase its timeliness and availability, and generate greater value. The medical field has accumulated a huge amount of data, but most of the data resources are scattered among different countries, research units and researchers. China officially launched the “National Medical and Health Science Data Sharing Project” in April 2004. It includes 1 network, 6 data centers, 40 or so principal databases, and 300 or so databases (data set series). This framework contains many different levels of data integration and resource organization. Providing data resources and information services for government health decision-making, medical technology innovation, health care, medical talent development, and universal health (sharing international biomedical data.. The sharing of medical data at the international level has continued to develop. In 1997, the Human Brain Program was officially launched in the United States, with more than 20 well-known research institutes and universities participating. The goal is to establish a global management system and network collaborative research environment for all knowledge of the nervous system, so that the experimental data and research results of the brain can be managed flexibly and effectively, so as to maximize the use of these experimental data and results, sharing international nerves. Informatics resources reduce unnecessary duplication of research and waste of human and material resources.

Clinical databases, electronic medical records, and semi-structured data such as medical images are the concrete manifestations of big data in clinical medicine. Different from the latter two, the process of collecting information in the database is purposeful and proactive, and professionals have the information to enter, organize and unify the structure. Therefore, research based on databases is also more convenient and feasible. The application of the database to clinical work is an innovation. Clinical research is no longer limited to prospective RCTs, but rather to reflect the real world situation, gradually transitioning from RCTs to BCT (Big-data Clinical Trial), can be expected to be large Clinical studies in the data age BCT will replace RCTs as the dominant type of research. Founded in 1989, the American Thoracic Surgery Association (STS) database has covered 95% of heart surgery in the United States and collected 5 million surgical records. The Congenital Heart Surgery (CHSD) database is an important part of the STS database and is the largest database of congenital heart malformations in children in North America and is considered the gold standard for medical professional clinical outcome databases. In recent years, data mining based on the CHSD database has been increasing, and the positive effect of large databases on improving medical quality is becoming increasingly prominent. For example, Welke et al based on the CHSD database to explore the complex relationship between the number of cases and mortality in pediatric cardiac surgery; Pasquali et al based on the CHSD database to explore mortality after neonatal Blalock-taussig shunt; Jacobs et al Based on the CHSD database, the multivariate analysis method was used to study the importance of preoperative factors in patients; Dibardino et al used a multivariate analysis method to explore the effects of gender and ethnicity on the outcome of congenital cardiac surgery based on the CHSD database. In recent years, a series of high-quality clinical databases have emerged in the field of cardiovascular surgery at home and abroad, such as the British Thoracic Heart Surgery Association (SCTS) database, the Australian and New Zealand Cardiothoracic Surgery Database, and the Chinese Adult Cardiac Surgery Database of the Department of Cardiovascular Diseases. Both greatly improved the success rate of cardiovascular surgery. The anticoagulant therapy database for Chinese patients after heart valve replacement

established by West China Hospital of Sichuan University, collected tens of thousands of hospitalized and follow-up data on anticoagulant therapy after heart valve replacement, and anticoagulation for heart valve replacement in China. Treatment research provides solid data support. The establishment of a large database can better support the clinical data mining work, thus forming a clinical data collection - mining - closed loop of clinical decision support, to achieve continuous improvement and improvement of medical quality.

7. Conclusion

Data mining has been favored in academia and industry. At the same time, the data mining algorithms that are gradually improved are more efficient and accurate, and the field of data mining applications is more broad. Medical data mining is a new interdisciplinary subject with wide-ranging and technical difficulties. It needs to work closely with computer, intelligent information processing, statistics and medical experts to break through various technical difficulties. With the further in-depth study of data mining theory and the application in the medical field, the huge and unique medical field will provide a broad space for data mining, and data mining will certainly bring new vitality to medical development.

Acknowledgement

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

References

[1] Jiang Guangkun. Research on Medical Data Mining Algorithm Based on Hadoop Platform in Big Data Environment [J]. Machine Tool & Hydraulics, 2018, 46(18):168-173.

[2] Qian Fengcui, Yin Jiaqi, Lu Xingyu, et al. Medical Big Data Composition and Application: A Case Study Based on Life Science Data [J]. Medical Information, 2016, 29(28).

[3] Li Yutong, Yao Dengju, Li Zhe, et al. Research on medical big data mining system [J]. Journal of Harbin University of Science and Technology, 2016, v.21(02): 42-47.

[4] Jing Yanmei, Zhu Guangju. Application Research of Library Data Mining under the Background of Big Data [J]. Studies Theory:-.

[5] Wang Shaobo, Jing Jianwen, Fang Xuanqi. Discussion on Data Mining Method and Its Application in Big Data Background [J]. Management Review, 2017(14): 102-103.

[6] Ding Xiangwu, Yang Ying. Application of Data Mining in Medicine [J]. Journal of Hubei Medical College, 1999, 18(3): 149-151.

[7] Qin Wenzhe, Chen Jin, Dong Li. Research progress and application of medical data mining under the background of big data [J]. Chinese Journal of Thoracic and Cardiovascular Surgery, 2016(1): 55-60, total 6 pages.