Application and Exploration of Big Data Mining in Clinical Medicine

Yue Zhang[1], Shu-Li Guo[2], Li-Na Han[1], Tie-Ling Li[3]

[1]Department of Cardiovascular Internal Medicine, Nanlou Branch of Chinese People's Liberation Army General Hospital, Beijing 100853, China
[2]State Key Laboratory of Intelligent Control and Decision, School of Automation, Beijing Institute of Technology, Beijing 100081, China
[3]Department of Cadre Physiotherapy, Chinese People's Liberation Army General Hospital, Beijing 100853, China

Yue Zhang and Shu-Li Guo contributed equally to this work.

Abstract

Objective: To review theories and technologies of big data mining and their application in clinical medicine.
Data Sources: Literatures published in English or Chinese regarding theories and technologies of big data mining and the concrete applications of data mining technology in clinical medicine were obtained from PubMed and Chinese Hospital Knowledge Database from 1975 to 2015.
Study Selection: Original articles regarding big data mining theory/technology and big data mining's application in the medical field were selected.
Results: This review characterized the basic theories and technologies of big data mining including fuzzy theory, rough set theory, cloud theory, Dempster–Shafer theory, artificial neural network, genetic algorithm, inductive learning theory, Bayesian network, decision tree, pattern recognition, high-performance computing, and statistical analysis. The application of big data mining in clinical medicine was analyzed in the fields of disease risk assessment, clinical decision support, prediction of disease development, guidance of rational use of drugs, medical management, and evidence-based medicine.
Conclusion: Big data mining has the potential to play an important role in clinical medicine.

Key words: Big Data; Clinical Medicine; Data Mining

Introduction

With continuous developments in science and technology, especially in the technology revolution of cloud computing, internet, and mobile internet, the mass of data grows at an incredible rate. Humanity has quietly entered the era of "Big Data," which refers to large-scale data sets that cannot be captured, managed, or processed by common software tools. Big data are characterized by the four Vs − volume, variety, velocity, and value.[1] Medical big data are the application of big data in the medical field after the data related to human health and medicine have been stored, searched, shared, analyzed, and presented in innovative ways. With advancements in medical technology and hospital information, the scale of medical big data is continuously growing. For example, the content of an ordinary computed tomography (CT) film is about 150 MB while the image information of a pathological slide can reach 5 GB. By the end of November 2013, the total number of medical institutions in the country was 962,000, which included 24,000 hospitals and 923,000 other medical institutions. In 2012, a total of 6.9 billion people sought medical help at hospitals across the country. At present, medical data of a medium-sized hospital can reach dozens of trillion bytes in a year.[2] Data mining technology is also in constant development so that data can be extracted from the database for analysis. Compared to the original decision support system, data mining technology is more convenient for integrating data. It can be said that big data

Access this article online
Quick Response Code:
Website:
www.cmj.org
DOI:
10.4103/0366-6999.178019

Address for correspondence: Dr. Li-Na Han,
Department of Cardiovascular Internal Medicine, Nanlou Branch of
Chinese People's Liberation Army General Hospital,
Beijing 100853, China
E-Mail: hanlina3399111@sina.com

Received: 18-09-2015 **Edited by:** Yuan-Yuan Ji
How to cite this article: Zhang Y, Guo SL, Han LN, Li TL. Application and Exploration of Big Data Mining in Clinical Medicine. Chin Med J 2016;129:731-8.

and data mining technologies open the door to success. In this paper, relevant concepts of big data and data mining, the classification and characteristics of data mining technologies, and the application of data mining in medical and health fields are discussed.

Basic Theories of Data Mining

Fuzzy theory

The mathematic foundations of fuzzy theory include fuzzy sets and fuzzy reasoning, which determine the relationship of relevant things by discovering the membership function and reasoning based on certain rules. Fuzzy theory is especially suitable for ideas with certain rules that are widely used in clinical medicine. Kimiafar *et al*.[3] used a fuzzy hierarchy process to analyze factors affecting the satisfaction of nurses in hospital information systems. They considered three main factors (service quality, system, and information) and 22 secondary factors. Ultimately, they learned that information quality was the most important index (58% weight). As another kind of example, Kuo *et al*.[4] used two-stage fuzzy neural networks to predict the prognosis of prostate cancer. This method studied the relationship between clinical features and prognoses of prostate cancer. After clinical data are acquired, the prognosis of patients with prostate cancer can be predicted and a more accurate health expectation can be obtained.

Rough set theory

Rough set theory is a mathematical tool to describe and solve inexact problems. This theory has become one of the most powerful tools in the field of data mining. Its core feature is that no prior information is required to solve problems outside the data set; as such, it is simple and easy to operate. However, the disadvantage is that it is difficult to deal with continuous attributes directly. It is widely used in data reduction, data discovery, data evaluation, and data analysis. Gil-Herrera *et al*.[5] used a meta-analysis method based on rough set theory to help individual patients identify homogeneous subgroups. This technique involves distinguished and nondistinguished rough set theories to create patient's homogeneous groups. A total of 1111 patients were divided into nine randomized controlled trials to study the effects of two-step management of malignant tumor transplantation. Three subgroups of patients created using this method had a significantly lower statistical heterogeneity (16.8%, 0, and 0, respectively). This method has great potential in the automation and standardization of the detection process as well as in meta-analysis of the management of intermittent peritoneal dialysis and will ultimately help personalize medical decision-making. Li *et al*.[6] proposed a classification algorithm of AdaBoost-support vector machine (SVM) and combined it with cluster boundary sampling preprocessing techniques (CBS-AdaBoost-SVM) for the early diagnosis of breast cancer. The algorithm used a machine learning method to diagnose unknown image data. At the same time, a genetic algorithm based on rough set reduction algorithm was used to remove redundant features. The effectiveness of the proposed methods was examined on digital data switching matrix by calculating the accuracy, confusion matrix, and receiver operating characteristic curves, which gave the physicians important insights for the early diagnosis of breast cancer.

Cloud theory

Cloud theory is an extension of fuzzy theory. It is considered an uncertain transformative model for a qualitative concept expressed in terms of language values and its quantitative representation. It avoids the limitations of the traditional fuzzy set theory by integrating the fuzziness and randomness of qualitative concepts, which provides a powerful method for the information processing of qualitative and quantitative integration. Liu and Xiao[7] designed a new health management model based on the cloud platform. They brought together elements of the health system by "vertical stratification" and "horizontal crossing" and then used the interconnections to share and collaborate with all kinds of clouds. A new health management system based on survey, identification, and regulation was established. The system manages health networks and intervenes preventively before diseases happen. In doing so, it could potentially block, delay, and even reverse the development of disease processes.

Dempster–Shafer theory

Dempster–Shafer (D-S) theory is the extension of the classic probability theory, which is an inaccuracy theory.[8] It satisfies conditions weaker than the Bayesian probability theory and has the ability to express directly. At present, the development aim is to combine the D-S theory with rough set theory, fuzzy theory, and artificial neural network (ANN) to improve data processing. This theory is widely used in clinical medicine. For example, Lian and Denoeux[9] used ELT-FS (evidential low-dimesional transformation with feature selection) to assess the results of cancer treatment and achieved high prediction accuracy in both lung cancer and esophageal cancer groups. Worsley *et al*.[10] used the Cardiff D-S theory to assess changes in subjective and objective functions from pre- to post-knee arthroplasty. They collected pain levels, joint stability, activity levels, and function in subjective aspects and objective functional assessments including goniometry, ultrasound imaging, three-dimensional motion analysis/conversion modeling of gait, and sit-stand. An optimal set of variables was used to classify the functions with the D-S theory. By calculation, the classification accuracy of healthy individuals and preknee arthroplasty patients was between 90% and 94%, and postknee arthroplasty subjective function improved 74% compared to healthy individuals. This method can be used to distinguish true patients who needed knee replacement surgery.

Artificial neural network

ANN is an algorithm developed from simulations of the human brain. It learns statistical law of data in a way similar to the mind's memory and induces a data model that can describe the features of a sample. Then, the data model that has been learned is used to classify new data. It has features such as massively parallel process, high degree of fault tolerance, self-organization, adaptive ability, and association functions. Compared to the fuzzy logic system, an ANN system is more intelligent. In the medical area, this method makes a leap of progress. Recently, Sunkaria *et al*.[11] developed a heart rate classification algorithm based on an ANN to evaluate the prognoses of patients' heart health. They chose the electrocardiograph (ECG) data of 46 patients in a hospital outpatient information system, input the information of each patient's ECG into the classification algorithm, and compared the outcomes with the results obtained from the heart disease experts. The results of both the algorithm and the experts were highly consistent with the patients' actual heart health.

Genetic algorithm

The first genetic algorithm was proposed by Holland in 1975.[12] This method encodes and computes data by simulating biology evolution in nature and induces selection, crossover, and mutation operations to filter data and eliminate invalid information through the rule of survival of the fittest; in the end, the required data are kept. Genetic algorithms are easy to parallel and have been widely used in classification and optimization problems. Johnson *et al*.[13] used this algorithm to forecast the progression of Alzheimer's disease. Compared to the stepwise selection method in their research, the genetic algorithm was better at predicting the development from a healthy state to mild cognitive impairment, and then to the final evolution into Alzheimer's disease. The results suggested that, in models of disease progression prediction, the contribution of combined variables is more important than that of a single variable.

Inductive learning theory

Inductive learning belongs to the category of machine learning, which extracts general rules and patterns by summarizing many pieces of experienced data. It is an important method of data mining. Birnbaum *et al*.[14] were devoted to studying the relationship between interleaving and inductive learning. They tested the discriminative contrast hypothesis in depth by examining the influence of interleaving and spacing as well as their combined effects. Interleaving enhances inductive learning with the roles of difference and retrieval.

Commonly used Data Mining Technologies

Bayesian network

A Bayesian network (BN) is a model of uncertain knowledge expression and reasoning based on probability analysis and graph theory. It is represented as a graph of assigned complex causal relationship networks. Each node in the network represents a variable, and the arc between variables expresses the direct causal relationship between the events. Because of the conditional independence of BN, the difficulty of solving a problem can be greatly simplified by considering the finite variables associated with each variable. In some areas, its performance is comparable to the neural network and decision tree. Lee *et al*.[15] used this technique as a means to predict the risk of radiation pneumonia. They collected 54 patients with nonsmall-cell lung cancer who underwent three-dimensional conformal radiation therapies and studied 19 patients. Serum levels of related biomarkers (α-2 microspheres, angiotensin-converting enzyme, transforming growth factor, interleukin-6) in the initial and intermediate treatment stages were detected, and the results showed that the optimal size for predicting radiation pneumonia using the BN method was 200 cases. The optimized performance record area under the receiver operating characteristic curve (AUC) in the BN model was 0.83, which was significantly higher than multivariate logistic regression (0.77). It was easy to see that BN methodology provided the flexibility to model hierarchical interactions between radiation pneumonia covariates, which was applied to probabilistic inference on radiation pneumonia. In addition, Dain *et al*.[16] used this method to predict low glycemic index and hypoglycemia risk in patients with type 1 diabetes. The results showed that if the basal rate of daily insulin dose was above 50%, the metabolic control was poor and the risk of hypoglycemia increased in type 1 diabetes patients wearing insulin pumps. However, different resources, such as a bolus calculator, temporary basal rates adjustments, and systematic educative training, probably reduced the hypoglycemia risk. These results provided experience for the set-up of insulin pumps.

Decision tree

A decision tree is a prediction model in machine learning. It classifies data by a series of rules and finds the field of the largest information in the database on the basis of information gain in information theory. Then, a node of the decision tree is established, and the branch of the tree is set-up according to the different values of the field. Finally, the tree's lower nodes and branches in each of the branches are established. A decision tree shows a clear and distinct feature that can generate rules which are easily understood. Furthermore, the calculation volume is low, so it is a popular data mining technique. At present, commonly used decision tree methods include the classification and regression tree, C4.5, iterative dichotomizer 3 (ID3), and supervised learning in quest. Among them, the ID3 method is currently the most popular. Yang *et al*.[17] used this method to explore risk assessments of public health incidents and achieved good results. In addition, Yaël *et al*.[18] used a decision tree method to analyze blood samples to improve the detection accuracy of mean corpuscular hemoglobin concentration (MCHC). A new fluorescent flow cytometry was used to obtain parameters (rbc-o and hgb-o), and correlation analysis was conducted on the red blood cell and hemoglobin levels. The results showed that rbc-o and hgb-o accurately detected abnormal MCHC levels. Through data learning, especially with the research of new parameters, it could improve guidance for biological interpretation and enable rapid follow-up.

Pattern recognition

Pattern recognition refers to the various forms of information processing and analysis of the characteristics of different items (numerical, literal, and logical relations) to describe, identify, classify, and interpret the items.[19] Patterns are data that have time and space distributions, and different data have different patterns. A computer can be used to classify massive data and filter it according to the different patterns of information to extract useful data. This technique is called pattern recognition. The pattern recognition method can be subdivided into the decision theoretic method, syntax method, and ANN method. It is widely used in the fields of product defect monitoring, voice and fingerprint identification, image signal analysis, and others. In the fields of medicine and health, pattern recognition has achieved great success in clinical automatic testing and analysis, extraction and analysis of ECG and electroencephalograph (EEG), medical image processing and analysis, automatic treatment planning, and auxiliary diagnosis. For example, Shi and Luo[20] used this technique in the noise reduction of magnetic resonance images (MRIs) and achieved remarkable results. In addition, in the process of medical equipment procurement risk assessment, Zhang *et al*.[21] proposed a method of the fuzzy comprehensive evaluation that facilitated the quantitative comparison of qualitative descriptions and drew the quantitative comparison results of procurement risks.

High-performance computing

High-performance computing (HPC) generally refers to the use of a number of processors or a cluster of several computer systems to carry out calculation analyses. This not only relates to the development of hardware devices, such as computer processors, servers, and networks, but also includes the research and development of HPC technology from the aspects of system structure, parallel algorithm, and software development. It is currently one of the hottest technologies of data mining. De Coninck *et al*.[22] proposed a new method of HPC called random regression - best linear unbiased prediction (DAIRRy-BLUP) to predict genomes. It is a kind of parallel and distributed memory algorithm based on the single character observation value (y). It is also an average information algorithm that uses the maximum likelihood estimation of variance components. The purpose of DAIRRy-BLUP is to provide more accurate marker effects and estimate breeding values by analyzing large data sets. Preliminary results showed that the DAIRRy-BLUP method was able to analyze large-scale data sets (1 million people, 360 thousand single nucleotide polymorphism [SNP]). The results also indicated that the increase in the record number of phenotypes and genotypes was more significant in prediction accuracy than that of the SNP array. Fleck *et al*.[23] developed parallel software for the entropy calculation of biological molecule conformation. It was a hybrid of multipoint interface and open multipoint C^{++} called PARENT, which was a particularly optimized HPC that provided an efficient estimation of entropic effects in different biological processes (e.g., protein-protein interactions) as well as allowed for a detailed mapping of intermolecular allosteric networks. The code was extensively benchmarked in a molecular dynamics trajectory set of protein complexes (in a range of 300–6000 atoms), ribonucleic acid, and lipid bilayers, exceeding 40 µs in total length; it demonstrated robustness and good scalability.

Statistical analysis

Statistical analysis is the oldest data mining technology. It uses the principle of statistics and probability theory to analyze relationship attributes to find out their correlation. The most commonly used models include linear and nonlinear analysis, continuous regression analysis, logistic regression analysis, university and multivariate analysis, and time series analysis. These models are simple and easy to operate and widely used in the fields of marketing, scientific research, management, and other fields for solving practical problems on an intuitive basis. Hortigüela-Saeta *et al*.[24] used statistical analysis to investigate epilepsy in children. For 6 years, they studied 68 patients (55.9% males with an average age of 3.7 years) with epilepsy in the Pediatric Intensive Care Unit of a referral hospital. According to statistics results, the most common symptom (50.0%) was generalized tonic-clonic seizures. The average duration of an epileptic episode was 51.44 min, the average number of antiepileptic drugs taken to prevent seizures was 3.21, and the average number of drugs taken before entering the intensive care unit was 2.37. The most commonly used drug was diazepam (83.8%) by rectal administration (75.0%). The second most used therapy was an intravenous injection of diazepam (52.9%), and the third was Phenytoin. The most common reason for epilepsy status was previous epileptic events (33.9%). Dravet syndrome was also discovered as a reason for epilepsy diagnosis. Through statistical analysis, the data becomes simple and clear, which could help doctors fine-tune clinical decisions.

Data Mining Applications in Clinical Medicine

In the medical field, very large volumes of data are generated, which have a broad prospect in clinical applications. The feasibility of some applications of clinical data mining was explored as described below.

Assessing disease risks

Disease risk assessment has a great influence on prognosis and clinical intervention strategies. Rapid and accurate assessment can help clinicians determine a patient's condition with which to decide the optimal treatment strategy. Relationships between risk factors and prognosis of diseases are complex. The same risk factors can occur in many different diseases, and one disease can be composed of many risk factors. As such, the relationship between risk factors and diseases has great correlativity rather than a simple causality. Therefore, data mining can be used to assess the disease risks to find the key factors associated with the disease prognosis in order to provide effective treatment for a disease's tertiary prevention. In this regard, this country's scholars are constantly exploring. Dong *et al*.[25] developed an algorithm based on the genetic and fuzzy theory (genetic fuzzy system) to evaluate the risk of unstable angina. They evaluated the related indexes of 54 cases including age, sex, blood pressure, creatinine, aspartate transaminase, lactate dehydrogenase, creatine kinase, MB isoenzyme of creatine kinase, and Troponin T as well as the previous history of cardiovascular diseases. The calculated results were compared to a senior clinician's judgment, and the two were found to be highly consistent. However, the data mining algorithm was more convenient and obtained results in a more timely fashion.

Supporting clinical decisions

In clinical diagnoses, doctors are often concerned with various symptoms of diseases. One disease can present different symptoms in different patients, while many diseases have the same or similar clinical manifestations. This makes it difficult to diagnose a disease, which can lead to misdiagnoses or missed diagnoses. According to published research, the average misdiagnose worldwide is 30.0%, with 27.8% occurring in China. Misdiagnosis not only leads missing the best window of opportunity for treatment, but also increases the financial burden on patients and their families, further inducing conflicts and disputes between doctors and patients. Data mining technology has brought new hope to disease diagnosis. By use of big data mining technology, a patient's previous diagnosis can be extracted, and the laboratorial results and clinical symptoms can be quickly obtained to help analyze similar manifestations. Doctors' experiences and the advantages of data mining can be combined to provide the best chance for accurate disease diagnosis. Armstrong *et al*.[26] used neural network technology to carry out a detailed analysis for characteristic indexes of 240 microcalcifications in mammographs across 220 cases. It was eventually concluded that data mining accurately predicted microcalcifications to be benign or malignant in early images of patients suspected of having breast cancer. In addition, Rastgarpour and Shanbehzadeh[27] developed a new kernel-based fuzzy level set to process medical image classification and recognition (management information system). They chose image samples including CT images of the blood vessels and heart, MRIs of the brain and breast, and silver microscopic images of nuclei. The image information was processed and analyzed by Matlab R2008 software (proivded by The MathWorks Company, USA) software, and the processed image was then compared to the original. The results showed that the image boundaries based on the new fuzzy algorithm were clearer than those in the original images. This method provided more reliable support for clinicians' diagnoses.

Predicting disease developments

At present, disease prognosis is mainly based on the classification of the disease diagnosis or key indicators of clinical laboratory testing. The majority of diagnostic grading standards are issued by foreign authoritative medical institutions. Due to ethnic differences, these criteria can lead to errors when applied to domestic patients and can affect doctors' judgments to some extent. In this regard, big data mining technology can provide more accurate results. ANN technology is widely used in various contexts; in the medical field, it can simulate the thinking mode of the human brain and has the ability to learn, think, and memorize, allowing it to deal with complex data. Sato *et al*.[28] applied ANN to predict survival rates of esophageal cancer. A total of 199 pieces of clinical and pathological data from 418 patients were collected. It was found that the neural network model had a significantly higher accuracy rate in predicting 1-year and 5-year survival rates ($P < 0.0001$, AUC = 0.884) than the prognosis method based on tumor, node, and metastasis staging criteria. Heckerling *et al*.[29] combined a neural network and genetic algorithm to predict the prognosis of patients with urinary tract infections. In this study, nine indexes (e.g., frequent micturition, dysuria, etc.) from 212 women with urinary tract infections were used as predictor variables for training. A satisfactory prognosis system was also similarly obtained.

Practical drug use guidance

The US has been at the forefront of developing the application of big data mining in health and medicine. Its government applied the results of big data research in the clinical field, the representative product of which is Pillbox. A drug guidance device based on big data, Pillbox contains all the information of a variety of drugs including color, shape, dosage, and interactions with other drugs. This device is capable of providing accurate information to patients for a variety of drugs, especially for rare clinical drugs, can help patients understand drug characteristics, and offer sensible drug use guidance. The operation of Pillbox is simple, making it compatible with elderly patients. Pillbox is believed to reduce the confusion of identifying different drugs. More importantly, it can help people understand a drug's performance and reduce side effects. Japanese scholars Suzuki *et al*.[30] used the data in the World Health Organization VigiBase to discover that combined drugs can change the occurrence frequency of drug-induced liver damage. They used acetaminophen, isoniazid, valproic acid, and amoxicillin-clavulanic acid to test drug hepatotoxicity, and evaluated the frequency of liver events with these four drugs in the presence of comedications. The conclusion was made that comedications can modify drug hepatic safety. Han *et al*.[31] summarized the types and mechanisms of experimental drugs used in autoimmune myocarditis, and carefully analyzed the advantages and characteristics of various treatments. Big data mining technology can also provide an important technical support for predicting the effects of experimental drugs in clinical medicine.

Strengthening medical management

Big data mining has a brilliant future in medical management. It can help to analyze medical operation indicators of hospitals for a period of time, including medical information management, medical quality management, medical supplies management, financial management, medical dispute management, doctors work performance, hospital management, decision-making management, hospital resource management, etc., to help hospital administrators provide data support for medical decision-making. In the study of Chen,[32] the factors influencing the number of patients received in a hospital were analyzed by using a gray correlation analysis method in data mining. It was found that the number of bed turnover times and the number of patients treated with surgery were related most significantly to the number of admitted patients, followed by the average number of open beds, and the average number of doctors. In addition, Peking Union Medical College Hospital carried out data mining research in clinical medical management to provide support for hospital business management and to assist the leadership in decision-making.[32] Guangzhou Nanfang Hospital used a data mining query system to verify whether doctors prescribed excessive medications and whether patients caused overspending on medical treatments.[33] In the future, big data mining will cover all aspects of medical management to allow hospitals to work more efficiently.

Benefiting evidence-based medicine

Evidence-based medicine has become the main direction of clinical work. For each specific clinical issue, medical workers can search the latest research progress to solve the problem, which will avoid rashness caused by experience. Data mining technology can take large amounts of evidence-based on medicine research into the data warehouse with the innovation of the data mining technology; then, a more efficient search method will be conducted and more accurate search results will be provided. The project of the wisdom health care system in Ningbo was based on the construction project of evidence-based medicine. The data sharing and exchange of this project have been completed. After eight municipal medical institutions were docked with platforms of the city, urban, and district (Yinzhou district, Haishu district, Jiangdong district, Jiangbei district, and Fenghua county), it realized the collection and exchange of diagnosis information, and treatment and health care information including most patients' electronic medical records and medical information for the entire city. The mining of the rich health information provided doctors with a direct practical basis for the clinical diagnosis and treatment of diseases to more accurately and quickly treat patients.[34]

Future Prospects

At present, the theories and technologies of big data mining are being explored for improvement. Each mining theory and technology has its own characteristics and limitations. Fuzzy theory is more frequently studied because it can deal with incomplete data and is suitable for data with a few rules, but the disadvantage is the incompleteness. Rough set theory is simple and does not need transcendent information except data that need to be processed. Its disadvantage is that it cannot deal with continuous attributes. Cloud theory is characterized by virtualization, high extension, and service diversity although its security is questionable. The D-S theory can satisfy conditions that are weaker than Bayesian probability and express uncertainties directly. ANN works for classification and continuous variables and has good robustness and self-organizing adaptability, but the disadvantage is opacity and "black boxes." Genetic theory can concurrently deal with all kinds of data and is easy to combine with other models, but encoding and calculation are difficult, resulting in only the guarantee of global optimization. Inductive learning theory can make fast classification and is suitable for large database without need of priori knowledge. However, it will be misled by implicit inductive bias when training data are insufficient. At the same time, along with data mining theories, each data mining technology also has its own characteristic and disadvantage. The BN has conditional independence. It can simplify difficulties in solving problems and is easier to obtain and reason the knowledge, but it is only suitable for static analysis. The decision tree is intuitive, clear, concise, and of high classification speed. It is suitable for large-scale data processing, but it has difficulty in expressing complex concepts and insufficiently emphasizes mutual relationships between the same characters; additionally, it has a poor ability to denoise. The HPC method is powerful but has the disadvantages of high cost and difficulty in programming design. The statistical analysis method is simple, and its workload is small, but it requires higher data integrity. The pattern recognition method can reflect structural characteristics of a pattern and has a strong anti-interference ability for an image. The disadvantages are the rejection and error rates. Technical features and application of various mining theories are shown in Table 1.

In the future, big data mining technology will continue to improve and expand. In the technical field, a programming language specially used for data mining will be developed to make data mining more formal and standardized. At the same time, a visualized data mining method will be established, which will be convenient for the user to understand and manipulate. Weiss *et al*.[35] explored a combination of visual three-dimensional printing technology and biological data mining in order to improve the efficiency of data mining. In addition, data mining based on the Wed network will be the focus of research direction. This will establish a data mining server on the internet that cooperates with the database server to realize data mining in order to establish a powerful data mining engine and data mining services market. Research on various mining methods of unstructured data, such as text data, graphics data, video data, audio data, and multimedia data mining, will greatly expand the data source and realize tight data integration. In application fields, big data mining technologies will permeate all aspects of medicine to provide more rapid and accurate results for clinicians to assess disease conditions, and make predictions for patients, eventually even real-time predictions. Zhang *et al*. proposed a new system that analyzed the medical data stream and made predictions in real-time. This system was based on a stream mining algorithm called very fast decision tree, which avoided the need for an offline clustering process and reduced resource consumption. The prototype of this

Table 1: Technical features and application of various data mining theories

Data mining theories	Advantage	Weakness	Application examples	Author	Reference
Fuzzy theory	Deals with incomplete data; does not need a complex mathematical model; it is easy to understand and use	Not thorough	Predicts the prognosis of prostate cancer; can classify and recognize medical images	Kuo *et al.*, Rastgarpour and Shanbehzadeh	[4,27]
Rough set theory	No prior information is required to process data; is able to handle data that cannot be distinguished from available properties; is simple and easy to operate	Difficult to deal with continuous discrete attributes directly; unable to obtain sufficient support for objective facts	Helps individual patients identify homogeneous subgroups	Gil-Herrera *et al.*	[5]
Cloud theory	Provides the model for quantitative and qualitative analysis of the uncertain; is characterized by virtualization, high extension, and service diversity	Data security and privacy protection properties are questionable	Provides new health management model based on cloud platform	Liu and Xiao	[7]
Dempster-Shafer theory	Satisfies the condition weaker than the Bayesian probability theory; has the ability to express uncertainty directly	Conflicting evidence fusion may obtain counterintuitive conclusion	Assesses cancer treatment outcomes	Lian and Denoeux	[9]
Artificial neural network	Has a strong ability to deal with the uncertain information; can deal with both categorical and continuous variables; good robustness, self–adaptability, parallel processing, distributed storage, and high fault tolerance	Not suitable for high-dimensional variables; difficult to understand learning and decision-making process of network; it has opacity	Assesses prognoses of heart health	Sunkaria *et al.*	[11]
Genetic algorithm	Can handle all types of data in parallel; is easy to combine with other models; solves the optimization of overall situation problems with solution that is independent of the initial conditions	Too many parameters are needed; encoding is difficult and the amount of calculation is large; can only ensure optimization trend of overall situation; cannot ensure the optimal results reached by the probability	Predicts progress in Alzheimer's disease	Johnson *et al.*	[13]
Inductive learning theory	Has high speed of classification; is suitable for large database without need of priori knowledge	Will be misled by implicit inductive bias when training data is insufficient	Test difference control hypothesis	Birnbaum *et al.*	[14]
Bayesian network	Has conditional independence; simplifies problem solving; complexity of knowledge acquisition and reasoning is low	Uses acyclic assumption and applies to static analysis only	Predicts the risk of radiation pneumonia	Lee *et al.*	[15]
Decision tree	Uses the decision tree diagram; is intuitive, simple, and clear; Has high speed of classification; the decision-making process is visible and suitable for large–scale data processing	Difficult to express complex concepts; insufficient emphasis on the relationship between the same characters; its noise immunity is poor	Assesses risk of public health events	Yang *et al.*	[17]
Pattern recognition	Easily identified; reflects pattern structures; has a strong anti-interference ability	Has rejection rate and error rate; difficult to select base unit when interference is encountered	Denoises magnetic resonance image	Shi and Luo	[20]
High-performance computing	High computing power	High cost and difficulty in designing programming	A new method for high-performance computing is proposed: DAIRRy-BLUP	De Coninck	[22]
Statistical analysis	The most basic data mining technology; operation is simple with less workload	Poor accuracy and reliability; high requirement for data integrity	Provides treatment plans for children with epilepsy	Hortigüela-Saeta *et al.*	[24]

BLUP: Best linear unbiased prediction.

system is currently being developed in order to combine appropriate medical data to verify the effectiveness of the design model. The advantage of the system is that it will run online and advance with the changing environment. If the system is successfully developed, it will be very useful for emergency rescue and on-site care.[36]

Conclusions

Big data mining technology opens up a new era in which guidelines and characteristics of many things are readily available from the mass of basic data. Although there are difficulties and problems that remain to be solved in its current infancy, the medical data mining field is promising and will become, with the relentless progress of science and technology, a powerful assistant in the care and prevention of human disease.

Financial support and sponsorship

This study was supported by grants from Hainan Province Introduced and Integrated Demonstration Projects (No. YJJC20130009) and Hainan Province Science and Technology Cooperation Fund Project (No. KJHZ2015-36).

Conflicts of interest

There are no conflicts of interest.

References

1. Wu ZJ, Guo Q. Research on the development strategy of China's health management industry in the era of big data. Health Econ Res 2014;6:14-6.
2. Yu GP, Bao XY, Huang XT, Liu W, Xu BB, Yu N, *et al*. Medical and health big data: Types, characteristics, and relevant issues. J Med Inform 2014;35:9-12. doi: 10.3969/j.issn.1673-6036.2014.06.002.
3. Kimiafar K, Sadoughi F, Sheikhtaheri A, Sarbaz M. Prioritizing factors influencing nurses' satisfaction with hospital information systems: A fuzzy analytic hierarchy process approach. Comput Inform Nurs 2014;32:174-81. doi: 10.1097/CIN.0000000000000031.
4. Kuo RJ, Huang MH, Cheng WC, Lin CC, Wu YH. Application of a two-stage fuzzy neural network to a prostate cancer prognosis system. Artif Intell Med 2015;63:119-33. doi: 10.1016/j.artmed.2014.12.008.
5. Gil-Herrera E, Tsalatsanis A, Kumar A, Mhaskar R, Miladinovic B, Yalcin A, *et al*. Identifying homogenous subgroups for individual patient meta-analysis based on rough set theory. Conf Proc IEEE Eng Med Biol Soc 2014;2014:3434-7. doi: 10.1109/EMBC.2014.6944361.
6. Li P, Bi T, Huang J, Li S. Breast cancer early diagnosis based on hybrid strategy. Biomed Mater Eng 2014;24:3397-404. doi: 10.3233/BME-141163.
7. Liu KH, Xiao LF. New health management model based on cloud platform and big data. J Public Health Prev Med 2014;25:89-91.
8. Meng XM, Ling PL, Gong XH. Research on techniques and tactics intelligent decision support system of net antagonistic event competitions project. Comput Eng 2012;21:148-52. doi: 10.3969/j.issn.1000-3428.2012.21.040.
9. Lian C, Denoeux T. Cancer therapy outcome prediction based on Dempster-Shafer Theory and PET imaging. Med Phys 2015;42:3549. doi: 10.1118/1.4925280.
10. Worsley PR, Whatling G, Barrett D, Holt C, Stokes M, Taylor M. Assessing changes in subjective and objective function from pre-to post-knee arthroplasty using the Cardiff Dempster-Shafer Theory classifier. Comput Methods Biomech Biomed Engin 2015;22:1-10. doi: 10.1080/10255842.
11. Sunkaria RK, Kumar V, Saxena SC, Singhal AM. An ANN-based HRV classifier for cardiac health prognosis. Electron Healthc 2014;7:315-30. doi: 10.1504/IJEH.2014.064332.
12. Holland JH. Adaptation in Natural and Artificial Systems. Ann Arbor: University of Michigan Press; 1975.
13. Johnson P, Vandewater L, Wilson W, Maruff P, Savage G, Graham P, *et al*. Genetic algorithm with logistic regression for prediction of progression to Alzheimer's disease. BMC Bioinformatics 2014;15 Suppl 16:S11. doi: 10.1186/1471-2105-15-S16-S11.
14. Birnbaum MS, Kornell N, Bjork EL, Bjork RA. Why interleaving enhances inductive learning: The roles of discrimination and retrieval. Mem Cognit 2013;41:392-402. doi: 10.3758/s13421-012-0272-7.
15. Lee S, Ybarra N, Jeyaseelan K, Faria S, Kopek N, Brisebois P. Bayesian network ensemble as a multivariate strategy to predict radiation pneumonitis risk. Med Phys 2015;42:2421-30. doi: 10.1118/1.4915284.
16. Dain A, Ruiz M, Rista L, Flores A, Muratore C, Ladino C. Basal and bolus rates in insulin accu-check combo pump type 1 diabetes patients and its correlation with age, frequency of self monitoring and metabolic control. Diabetes Technol Ther 2015;17:A55.
17. Yang Y, Sun H, Kang Z, Wu QH. Application of decision tree model ID3 algorithm in risk assessment of public health emergencies. Chin Prev Med 2015;16:60-4.
18. Yaël B, Pérol J, Dignat-George F. Managing samples with implausibly high mean cell hemoglobin concentration (MCHC) values by using red blood cell (RBC) and hemoglobin (HGB) values obtained by fluorescence flow cytometry. Clin Chem Lab Med 2015;53 Suppl 1:S156. doi: 10.1515/cclm-2015-5003.
19. Theodoridis S. Pattern Recognition. 4th ed. Beijing. Publishing House of Electronic Industry: Publishing House of Electronic Industry; 2010.
20. Shi HL, Luo SQ. MR image denoising based on nearly shift-insensitive and nonredundancy discrete wavelet transform. J Clin Rehabil Tissue Eng Res 2010;14:9739-43. doi: 10.3969/j.issn.1673-8225.2010. 52.013
21. Zhang HP, Zhou D, He MH. Evaluation of medical equipment purchasing risk. Inf Med Equip 2007;22:58-62.
22. De Coninck A, Fostier J, Maenhout S, De Baets B. DAIRRy-BLUP: A high-performance computing approach to genomic prediction. Genetics 2014;197:813-22. doi: 10.1534/genetics.114.163683.
23. Fleck M, Polyansky AA, Zagrovic B. PARENT: A parallel software for the calculation of conformational entropy in biomolecular systems. Eur Biophys J 2015;44:243-8.
24. Hortigüela-Saeta MM, Conejo-Moreno D, Gutiérrez-Moreno M, Gómez-Saiz L. Descriptive statistical analysis of the treatment of status epilepticus in a referral hospital. Rev Neurol 2015;60:433-8.
25. Dong W, Huang Z, Ji L, Duan H. A genetic fuzzy system for unstable angina risk assessment. BMC Med Inform Decis Mak 2014;14:12. doi: 10.1186/1472-6947-14-12.
26. Armstrong AJ, Marengo MS, Oltean S, Kemeny G, Bitting RL, Turnbull JD, *et al.* Circulating tumor cells from patients with advanced prostate and breast cancer display both epithelial and mesenchymal markers. Mol Cancer Res 2011;9:997-1007. doi: 10.1158/1541-7786.MCR-10-0490.
27. Rastgarpour M, Shanbehzadeh J. A new kernel-based fuzzy level set method for automated segmentation of medical images in the presence of intensity inhomogeneity. Comput Math Methods Med 2014;2014:978373. doi: 10.1155/2014/978373.
28. Sato F, Shimada Y, Selaru FM, Shibata D, Maeda M, Watanabe G, *et al*. Prediction of survival in patients with esophageal carcinoma using artificial neural networks. Cancer 2005;103:1596-605. doi: 10.1002/cncr.20938.
29. Heckerling PS, Canaris GJ, Flach SD, Tape TG, Wigton RS, Gerber BS. Predictors of urinary tract infection based on artificial neural networks and genetic algorithms. Int J Med Inform 2007;76:289-96. doi: 10.1016/j.ijmedinf.2006.01.005.
30. Suzuki A, Yuen NA, Ilic K, Miller RT, Reese MJ, Brown HR, *et al.* Comedications alter drug-induced liver injury reporting frequency: Data mining in the WHO VigiBase™. Regul Toxicol Pharmacol 2015;72:481-90. doi: 10.1016/j.yrtph.2015.05.004.
31. Han L, Guo S, Wang Y, Yang L, Liu S. Experimental drugs for treatment of autoimmune myocarditis. Chin Med J 2014;127:2850-9. doi: 10.3760/cma.j.issn.0366-6999.20140748.
32. Chen YH. Application research of data mining technology in hospital management (In Chinese). China Med Equip 2014;11:62-5. doi: 10.3969/j.issn.1672-8270.2014.01.022.
33. Gong WN. Application research of data mining technology in hospital management. Guide China Med 2012;10:722-5. doi: 10.3969/j.issn.1673-6036.2014.03.003.
34. Sun XD, Huang XQ, Zhu CL, Zhang K, Zhang H, Lu CT. Research on massive medical data mining analysis method based on evidence-based medicine. J Med Inform 2015;36:11-6.
35. Weiss TL, Zieselman A, Hill DP, Diamond SG, Shen L, Saykin AJ, *et al.* The role of visualization and 3-D printing in biological data mining. BioData Min 2015;8:22. doi: 10.1186/s13040-015-0056-2.
36. Zhang Y, Fong S, Fiaidhi J, Mohammed S. Real-time clinical decision support system with data stream mining. J Biomed Biotechnol 2012;2012:580186. doi: 10.1155/2012/580186.