Research on Medical Information Processing Based on Data Mining Technology

Zhiying Cao(✉)

The Affiliated Changshu Hospital of Soochow University (Changshu No.1 People's Hospital), Suzhou 215500, Jiangsu, China
502921758@qq.com

Abstract. Big data construction has become a national strategic policy, and medical big data related to human health is an important part of it. Data mining uses computers to extract useful information from massive, incomplete, noisy, fuzzy and random data. This paper introduces that data mining technology is divided into association classification rule technology, cluster analysis and rough set theory. As well as the feature pattern polymorphism, data fuzziness, data timing and data redundancy of medical data mining. Data mining architecture is composed of preprocessing module, mining process module, result evaluation module, knowledge guidance module and mining object management module. It is a system integrating information management, retrieval analysis and evaluation, and data warehouse. The application prospect of data mining in medical field is very broad. With the deepening of research, the value of data mining is constantly reflected, which can better serve the public health.

Keywords: Data mining · Data warehouse · Medical information processing

1 Introduction

Big data construction has become a national strategic policy, and medical big data related to human health is an important part of it [1]. In June 2016, the Guiding Opinions of the General Office of the State Council on Promoting and Regulating the Application and Development of Big Data in Health Care clearly stated that "Big Data in Health Care is an important basic strategic resource of the country" [2, 3]. The application of medical big data is of great significance to clinical medical research, scientific management and the transformation and development of medical service mode, and its research is bound to become an important development direction in the next two decades. The value of big data center can not be realized without data mining technology [4–6].

Data Mining (DM) is an information processing technology developed in 1990s. It is a process of extracting potentially useful information and knowledge hidden in a large number of incomplete, noisy, fuzzy and random data by computer, involving knowledge in many fields such as database, artificial intelligence and statistics [7]. By applying data mining technology to medicine, we can find the rules and patterns of medical

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2022
Published by Springer Nature Switzerland AG 2022. All Rights Reserved
S.-H. Wang and Y.-D. Zhang (Eds.): ICMTEL 2022, LNICST 446, pp. 510–516, 2022.
https://doi.org/10.1007/978-3-031-18123-8_39

diagnosis, thus assisting doctors in disease diagnosis and providing reliable basis for scientific management and medical research in hospitals [8–11]. The data mining process can be divided into nine stages: data preparation, data selection, data preprocessing, data reduction, data mining target determination, mining algorithm determination, DM, pattern interpretation and knowledge evaluation [12].

2 Research Status

(1) Association classification rules in 1998, association classification method cBA31 was put forward. CBA integrated the process of classification rule mining and association rule mining, and achieved better classification effect than decision tree classification algorithm C4.5 based on rule mining. Since then, some people have put forward various improvement methods aiming at the shortcomings of CBA, typically CARL2 and ARC. The basic idea of these methods is to use the existing association rule mining algorithm 3 to generate feature words or feature word itemsets that frequently appear in various categories, and use these frequent feature word itemsets to construct classification rules to classify test samples. The more frequent feature words the test sample contains and the higher the confidence level, the more likely it is that the test sample belongs to this category; otherwise, the less likely it is to belong to this category [13, 14]. Compared with other classification methods, association classification generates classification patterns in the training stage, and only needs to compare and match the documents to be classified with the classification patterns in the classification stage, so it has the advantages of short training time and classification time.

(2) Chaster analysis is also a commonly used technology in data mining, which refers to the process of dividing a data into several categories according to the principle that the distance between data objects in the same group is small, and the distance between data objects in different groups is large, among which there are many methods to define the distance. A group of abstract or physical objects are divided into several groups according to their similarity: the more similar objects are divided into one group, and this process is the clustering process. A collection of similar objects is called a cluster: objects in different clusters are not similar. Clustering is to search for valuable associations between data items from a given data set [15, 16].

Clustering is also called unsupervised induction in machine learning. The biggest difference between clustering and classification is that the classification problem is to classify data objects into different known classes when the classification attributes of training samples are known. In clustering problem, for unknown data objects, the final classification results need to be found in training samples. Cluster analysis has a wide range of applications, including business, insurance, biology, geography, medicine and so on. In business, cluster analysis can help market workers find different groups of customers and describe their different characteristics by purchasing patterns. In the field of biological research, cluster analysis can be used to obtain the hierarchical structure of animals and plants, and classify them according to their gene functions. Clustering analysis can also be used alone

as a tool to understand the characteristics of data and analyze the distribution of data, or as a preprocessing step of other algorithms, such as qualitative induction algorithm. In the medical field, cluster analysis has been widely used in DNA analysis, automatic analysis of medical image data, analysis of disease risk factors and so on [17]. Many researchers in China have been doing research in related fields, such as classification of Neisseria gonorrhoeae drug-resistant epidemic strains and clustering analysis of coronary atherosclerotic heart disease.

(3) In the actual system, there are uncertain factors of varying degrees in many cases. The collected data often contain noise, inaccuracy or even incompleteness, and roughness and theory are the mathematical tools to deal with such uncertain factors. In 1970s, scientists of Polish Academy of Sciences first put forward rough set theory. Rough set theory defines the concepts of fuzziness and uncertainty in the sense of classification, which can effectively analyze various uncertain information such as inconsistency, inaccuracy and incompleteness, and can also analyze and reason data to find hidden patterns and knowledge. The main goal of rough set theory is to get decisions and rules on the basis of keeping the classification ability unchanged by reducing knowledge. To understand the basic concepts of rough set theory, we must first understand a series of concepts such as knowledge attributes closely related to it. In rough set theory, "knowledge" can be regarded as an ability to classify realistic and abstract objects according to the attributes of features. People's behavior is based on the ability to distinguish real or abstract objects. For example, in ancient times, in order to survive, human beings had the ability to distinguish whether food was edible or not. Doctors must be able to distinguish which disease the patient was suffering from when diagnosing the patient. This ability to classify things according to their characteristics can be regarded as some kind of knowledge.

3 Characteristics of Medical Data Mining

Particularity of medical data Clinical medical data is the main object of medical data mining research, which contains all data resources of patients and diseases in the whole medical process. Compared with ordinary data, these data have the following particularity:

(1) pattern polymorphism: medical information includes pure document data manually entered by doctors, image data such as MRI and CT obtained by medical imaging equipment, signal data and voice data of otolaryngology, etc., which are polymorphic.

(2) Data fuzziness: the objective incompleteness of disease and case information and subjective inaccuracy in describing disease make it impossible for medical data to fully reflect the situation of patients or diseases, and at the same time, data design, data collection, data entry and other links may all lead to the missing of the final medical database, which leads to greater fuzziness of medical data.

(3) Timing of data: the patient's visit to a doctor or the onset of a disease has a progress in time, and the images obtained by medical equipment such as electrocardiograph are also a function of time. These medical data have higher timing than ordinary data.

(4) Data redundancy: Data redundancy means that the same information is stored in multiple places. There is a great deal of identical information in medical database, which may lead to wrong or meaningless patterns in medical data mining activities.

It is precisely because of these characteristics of medical data, and because it involves many ethical and legal issues, that medical data mining has its particularity, which requires some unique technologies in data mining.

4 Architecture of Data Mining

Data mining is based on artificial intelligence, machine learning and statistics. However, the data mining system does not simply combine these technologies, but adds a lot of auxiliary technical support to form a complete system. A typical data mining system is shown in Fig. 1, which has the following main components: It can be seen from Fig. 1 that the data mining system is composed of preprocessing module, mining process module, result evaluation module, knowledge guidance module and mining object management module, and is an application software system integrating information management, retrieval analysis and evaluation, data warehouse and so on. It can complete a series of tasks such as data collection, preprocessing, data analysis and result expression, and finally output the results to users. See Fig. 1.

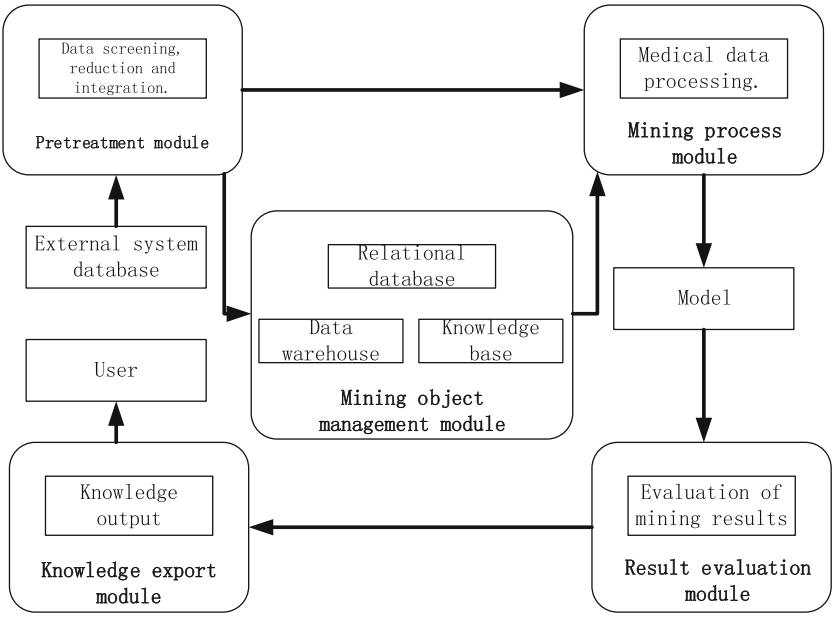

Fig. 1. Architecture of data mining

(1) Pre-processing module: The pre-mining processing module performs various processing on the collected data, including removing noise, integrating various data sources, selecting data related to the problem, changing the selected data into a mineable form, and then generating a data warehouse or data mining library. In the process of mining, if the pattern evaluation finds that the mining pattern is affected by data problems, it will return to the module for data processing before data mining.
(2) Mining process module: Mining operation module is the core part of the whole data mining system. It uses various data mining algorithms and technologies, such as decision tree induction, regression analysis, Bayesian classification, association analysis, online analytical processing, text and multimedia data mining technology, to mine and discover knowledge for databases and data warehouses, and by means of mining rules, experiences, methods and factual data in knowledge base.
(3) Result evaluation module: The main purpose of the mode evaluation module is to evaluate the knowledge and results obtained from data mining. Analyze and compare the user's interest degree with many patterns excavated, evaluate the value of the patterns, and analyze the defects. The difference between the user's interest degree and the mined module is too large, which needs to be returned to the corresponding module for re-execution. And some patterns that meet the user's interest will be directly transmitted to the knowledge output module.
(4) Knowledge export module: Knowledge export module is the interface and bridge between users and data mining system, which translates and explains the patterns obtained from data mining and provides them to decision makers in a way that users can easily understand. Users can directly interact with the system to provide information, formulate mining tasks, and carry out progressive and exploratory data mining according to the results of each step of data mining.
(5) Mining object management module: The database management module is responsible for maintaining and managing all kinds of databases in the system, including databases, data warehouses and mining knowledge bases. These internal databases are obtained by exchanging, cleaning and purifying with external databases, which is the basis of data mining. Mining knowledge base, also known as domain knowledge base, contains experience, methods, techniques, theories, rules, facts and knowledge used or obtained in the mining process. The main purpose of this module is to guide the mining process and evaluate the candidate patterns obtained from mining. From the above introduction, it can be seen that it is difficult to fully realize all functions of a complete data mining system. At present, many data mining systems are incomplete in a strict sense. If it can't handle a large amount of data, it should be called a machine learning system, or a statistical analysis tool, or an experimental system prototype. Similarly, if a system can only perform some data and information retrieval tasks, including summation operation or deductive query and answer, it can only be called an information retrieval system.

5 Summary

With the maturity of data mining technology, its application in medical field will be more and more extensive. Further research work has the following prospects.

(1) In the horizontal aspect, other technologies such as decision tree classification and artificial neural network in data mining also have their unique advantages. It is an important research topic to give full play to these advantages and apply them to other medical fields including disease diagnosis and analysis, and prediction after medication guidance.
(2) In the vertical aspect, the association rule mining and rough set mining in this paper are both single-layer mining, which has certain limitations and requires Boolean processing of features in advance, which to a certain extent affects the richness of extracted rules and the accuracy of classification by rules. If we can conduct in-depth research and multi-dimensional and multi-level image mining, the effect will be better.

In a word, the application prospect of data mining in the medical field is very broad. With the deepening of research, the value of data mining is constantly reflected, which can better serve the public health.

References

1. Sirichanya, C., Kraisak, K.: Semantic data mining in the information age: a systematic review. Int. J. Intell. Syst. **36**(8), 3880–3916 (2021)
2. Li, Z.: Research on the new path of internet of things data mining under the background of cloud computing. J. Phys. Conf. Ser. **1915**, 042089 (6 pp) (2021)
3. Istratova, E., Sin, D., Strokin, K.: A comparative analysis of data mining analysis tools. In: Pattnaik, P.K., Sain, M., Al-Absi, A.A., Kumar, P. (eds.) Proceedings of International Conference on Smart Computing and Cyber Security. Strategic Foresight, Security Challenges and Innovation (SMARTCYBER 2020). LNNS, vol. 149, pp. 165–72. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-7990-5_16
4. Chen, X., Zhao, D., Zhong, W., Ye, J.: Research on brain image segmentation based on FCM algorithm optimization. In: Fu, W., Xu, Y., Wang, SH., Zhang, Y. (eds.) Multimedia Technology and Enhanced Learning. ICMTEL 2021, LNICST, vol. 388, pp. 278–289. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-82565-2_23
5. Gupta, P., Hoi, C.S.H., Leung, C.K., Ye, Y., Xiaoke, Z., Zhida Z.: Vertical data mining from relational data and its application to COVID-19 data. Big Data Analyses, Services, and Smart Data. Advances in Intelligent Systems and Computing (AISC 899), pp. 106–116 (2021)
6. Rao, A.S., D'Mello, D.A., Anand, R., Nayak, S.: Clinical significance of measles and its prediction using data mining techniques: a systematic review. In: Chiplunkar, N., Fukao, T. (eds.) Advances in Artificial Intelligence and Data Engineering. Select Proceedings of AIDE 2019. AISC, vol. 1133, pp 737–59, Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-3514-7_56
7. Marimuthu, V.K., Lakshmi, C.: Performance analysis of privacy preserving distributed data mining based on cryptographic techniques. In: Proceedings of the 7th International Conference on Electrical Energy Systems (ICEES 2021), pp. 635–40 (2021)

8. Mandan, N., Agrawal, K., Kumar, S.: Analyzing different domains using data mining techniques. In: 2020 International Conference on Computer Communication and Informatics (ICCCI), p. 6 (2020)
9. Chen, X., Zhao, D., Zhong, W.: Auxiliary recognition of alzheimer's disease based on Gaussian probability brain image segmentation model. In: Ning, H. (eds.) Cyberspace Data and Intelligence, and Cyber-Living, Syndrome, and Health. CyberDI CyberLife 2019. CCIS, vol. 1138, pp. 513–520. Springer, Singapore (2019). https://doi.org/10.1007/978-981-15-1925-3_37
10. Mahmud, H., et al.: Technologies in medical information processing. Advances in Telemedicine for Health Monitoring: Technologies, design and applications, pp. 31–54 (2020)
11. Kishor, A., Chakraborty, C., Jeberson, W.: Reinforcement learning for medical information processing over heterogeneous networks. Multimedia Tools Appl. **80**(16), 23983–24004 (2021). https://doi.org/10.1007/s11042-021-10840-0
12. Melnykova, N., Mukalov, P., Koziy, D.: The special ways of application of neural networks for medical information processing. In: 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). Proceedings, pp. 428–431 (2018)
13. Karali, E.: Novel approaches to medical information processing and analysis. In: Lambropoulou, S., Theodorou, D., Stefaneas, P., Kauffman, L. (eds.) Algebraic Modeling of Topological and Computational Structures and Applications. Springer Proceedings in Mathematics and Statistics. PROMS, vol. 219, pp. 453–482. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68103-0_23
14. Andrikov, D.A., Kuchin, A.S.: Development of a prototype of a medical information system for a clinical diagnostic center. Procedia Comput. Sci. **186**, 287–292 (2021). (14th International Symposium "Intelligent Systems", INTELS 2020)
15. Chen, X., Zhao, D., Zhong, W., Ye, J., Gao, F.: Research on early warning monitoring model of serious mental disorder based on multi-source heterogeneous data sources. In: Zhang, YD., Wang, SH., Liu, S. (eds.) Multimedia Technology and Enhanced Learning. ICMTEL 2020. LNICST, vol. 327, pp. 403–410. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-51103-6_36
16. Xinlei, C., Xiaogang, R., Yue, W., Jiufeng, Y.: Design and realization of a comprehensive management system for severe mental disorders based on FLUX mode. J. Med. Imaging Health Inf. ASP **10**(2), 522–527(6) (2020)
17. Zhang, Y., Wu, L., Wang, S.: Magnetic resonance brain image classification by an improved artificial bee colony algorithm. Progress Electromagn. Res. **116**(2011), 65–79 (2011)