Journal of Medical Systems (2019) 43: 125
https://doi.org/10.1007/s10916-019-1214-8

IMAGE & SIGNAL PROCESSING

Schizophrenia Auxiliary Diagnosis System Based on Data Mining Technology

Xiaohong Wang[1] · Na Zhao[1] · Peng Ouyang[2] · Jiayi Lin[3] · Jian Hu[1]

Received: 26 December 2018 / Accepted: 13 February 2019 / Published online: 27 March 2019

Abstract

In order to use digital medical technology to develop and design an auxiliary diagnosis system for schizophrenia to assist doctors at all levels to diagnose and predict the cure of patients, improve the accuracy of diagnosis of symptoms, find complications in advance, and reduce the risk of disease, the application of Bayesian network in auxiliary diagnosis system of schizophrenia is studied, and an auxiliary diagnosis system of schizophrenia is designed. Based on data mining technology, knowledge information can be found from patient data and used to diagnose the nature of patients. The demand analysis of auxiliary diagnosis system is briefly introduced, and an auxiliary diagnosis system for schizophrenia based on Bayesian network is designed.

Keywords Bayesian network · Auxiliary diagnosis system · Demand analysis · Functional design · Data mining

Introduction

Schizophrenia is a serious mental disease, and the incidence is unexplained. Emotion, perception, thinking and behavior and other aspects of the disorder and mental activities and other symptoms are often related to syndrome in clinical [1]. The incidence rate is 0.007% - 0.014%, often having onset in young adults, and the clinical cure rate is low, which brings a lot of burden to the patients and their families [2]. In life, the general awareness of patients with schizophrenia are relatively clear and intelligence is basically normal, but in the disease process, the course generally has a long period of time, and the disease will deteriorate fast. As a result, the cognitive function of patients is damaged, and it will cause mental decline or mental disability [3]. But some patients, after a scientific and reasonable treatment, can achieve rehabilitation or basic rehabilitation.

If digital medical technology can be used to develop a schizophrenia auxiliary diagnosis system to assist doctors at all levels to diagnose and predict schizophrenia patients, the accuracy of diagnosis can be improved, symptoms can be detected in advance and the risk of onset can be reduced [4]. Digital intelligent medical technology is the trend of the development of medical information technology. It solves the “information island” phenomenon among different medical institutions and realizes the sharing of medical information among different institutions [5]. With the continuous development of the scale of hospital clinics, the number of patients in hospital clinics is also increasing, the number of patrols per unit time is increasing sharply, the patient information management is also very complex and diversified, and the traditional way is gradually difficult to meet the requirements of patients on the level of service, seriously affecting the operational efficiency of medical institutions and hindering the development of medical institutions [6]. Therefore, it is necessary to build a mobile patrol information management and query platform for schizophrenics. Wireless technology is an important method and means to update the information and data of schizophrenia patients to the central system quickly and effectively, to solve various problems in the mobile patrol of schizophrenia patients for medical institutions, and to

This article is part of the Topical Collection on *Image & Signal Processing*

✉ Jian Hu
drhujiangbest1@163.com

[1] Department of Psychiatry, The First Affiliated Hospital of Harbin Medical University, 23 Youzheng Street, Nangang District, Harbin 150001, China

[2] School of Management, Harbin Institute of Technology, Harbin 150001, China

[3] Beijing Electro-Mechanical Engineering Institute, Beijing 100074, China

improve the service level and management efficiency of departments [7].

The research content of schizophrenia auxiliary diagnosis system is to integrate and analyse individual genetic background data, health data, disease-related molecular biology data and drug clinical trial data, and to form a network electronic health system technology and data analysis system to study disease prediction, diagnosis, treatment, and prevention digital medical knowledge analysis method and its integrated software [8]. From a biological and medical point of view, it is difficult for biologists to discover the effects of a single or several genes on organisms and the relationship between them as a whole by manipulating them. However, with the development of technology, it is now possible to analyse personal health indicators, medical records, drug reactions and other data [9]. At the same time, genetic information, protein family tree information, genome-wide expression and methylation information, as well as epigenetic information can also be analysed. If biologically multi-dimensional and multi-directional data can be fused organically, a patient can be described completely, thus achieving precise medical purposes for schizophrenics [10].

Method

Demand analysis of auxiliary diagnosis system for schizophrenia

Data requirement analysis based on Bayesian algorithms: According to the data characteristic information required, case report forms are designed and subjects are selected in the research hospital centers. Finally, 316 schizophrenic patients are selected as data to verify the model of auxiliary diagnosis system. Among them, the selected patient data contains 237 dimension attributes and 8 of 237 dimension attributes are category attributes. 17 attributes have more vacancy values, vacancy rate is about 13%, 79 attributes have discrete data values, and other attributes are continuous data values [11]. In the latter study, the selected schizophrenic patients' sample data will be used for learning and training to obtain decision rules, and to explore whether the data demand characteristics can better meet the needs of the system, and then be used for clinical auxiliary diagnosis.

316 patients with high-dimensional small sample data are selected for two main purposes: to study the impact of the attributes of clinical samples on their categories, and to explore whether the patient data meet the needs of the auxiliary diagnosis model to guide the diagnosis process; to find a method of mining high-dimensional small sample data. The main reason for the analysis of high-dimensional small sample data is that in some cases or in a short time only some data can be obtained, and knowledge can be obtained from these data, so it is necessary to study high-dimensional small sample data [12]. The most important thing here is how to ensure and improve the accuracy of the results of this auxiliary diagnosis model.

By mining and analyzing the small sample data in clinical diagnosis, we can no longer be limited by the number of samples; the important condition attribute set obtained can help doctors check only a few important items when examining patients, which can not only reduce the diagnosis cost of patients, but also optimize the allocation of medical resources. The results obtained after mining and analysis can be applied to the auxiliary diagnosis system, and then help or assist doctors to diagnose schizophrenia of patients.

System function business process requirement analysis: The main purpose of data pre-processing is to process the data with redundancy, incompleteness, noise and high dimensionality that cannot directly use Bayesian network, to provide simple, clean, accurate and normal data for the auxiliary diagnosis system of schizophrenia, and to improve the efficiency and accuracy of the auxiliary diagnosis system information processing of schizophrenia.

Therefore, the flow chart of data pre-processing that the auxiliary diagnosis system should adopt is shown in Fig. 1.

The auxiliary diagnosis system of schizophrenia based on Bayesian network is composed of Bayesian network structure and parameters. The Bayesian network can be used to obtain the Bayesian network structure from the patient sample data set through structural learning, then to learn the parameters, and finally to obtain the parameters of the Bayesian network. The construction process of Bayesian network is basically the same, and its workflow is shown in Fig. 2.

In fact, the process of auxiliary diagnosis of schizophrenia by Bayesian network is to use the Bayesian network has been built to calculate and analyse the newly added patients' data, and to judge the type of the input patients' data. As a result, reasoning diagnosis is actually a problem of classification. The process of diagnosis should first preprocess the original information of patients, standardize the patient records, and then calculate the patient records using Bayesian network to get the diagnosis results. The diagnostic workflow is shown in Fig. 3.

The update of the sample database mainly refers to adding new patient data information to the sample database. The process of updating is to manage the pre-processing of the patient information that has been diagnosed and input it into the

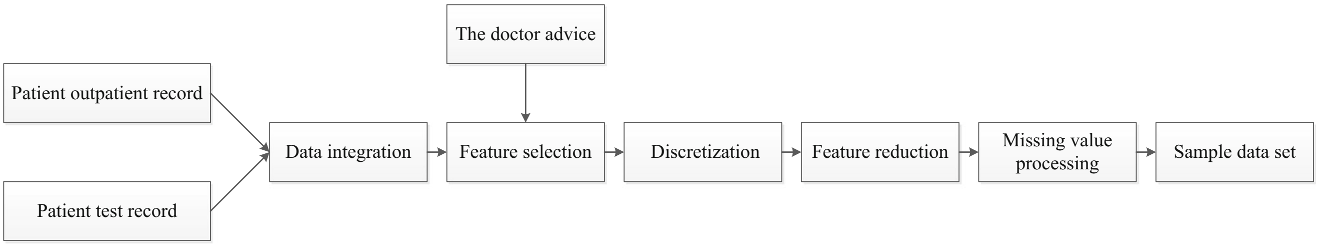


Fig. 1 Data pre-processing workflow diagram

sample database to obtain a new sample database. The workflow is shown in Fig. 4.

Demand analysis of auxiliary diagnosis function model based on Bayesian network. Naive Bayesian is an important branch of Bayesian decision theory. Naive Bayesian hypothesis requires that the value of an attribute affects a given class independently of other attribute values and it is a supervised learning method. Although this harsh restriction is often not met in reality, naive Bayesian reasoning usually implements attribute selection process first in data sets, which improves the independence of attributes. Moreover, naive Bayesian reasoning can generate more complex non-linear decision-making surfaces, and can fit fairly complex surfaces and achieve great success.

Based on the improved naive Bayesian method of attribute weighting, the predictive formula of the patient's untreated probability can be obtained according to the formula as follows:

$$P\left(C_2/X_j\right)=\frac{P\left(X_j/C_2\right)P(C_2)}{P\left(X_j/C_1\right)+P\left(X_j/C_2\right)P(C_2)} \tag{1}$$

This involves the estimation of the class conditional probability density. $P(X_k|C_j)$ can be obtained from the training set by fitting the class conditional probability density (that is, the probability density function of X_k) of the characteristic attribute component X_k in each grouping C_j. According to the value type of attribute variable X_k, the estimation methods of class conditional probability density are different.

When X_K is a discrete numerical value, then:

$$P\left(K_k/C_j\right)=\frac{N_{jk}}{N_i} \tag{2}$$

When it is a continuous numerical value, according to the improved naive Bayesian model method mentioned above, that is, the conditional probability density function fitting Xk by the kernel density estimation method according to formula X_k, $P(X_k|C_j)$ is calculated as follows:

$$P\left(X_k/C_j\right)=\frac{1}{nh}\sum_{t=1}^{n}K\left(\frac{X_k-X_t}{h}\right) \tag{3}$$

K(x) is called the kernel function, h is called the window width of the kernel function, that is, if the larger h is chosen, the deviation may be larger, and the estimated probability density function will be smoother; if smaller, the estimated probability density function will not be so smoother, but the probability density curve and sample fitting will be relatively better.

The Logistic regression method and the naive Bayesian method before and after improvement are used to establish a model to predict the cure probability (PHM) of schizophrenia patients in the course of treatment. The model is applied to the auxiliary diagnosis system. The resolution performance of the three models on validating the data set of schizophrenia patients is shown in Fig. 5.

Among them, the area under ROC (Receiver Operating Characteristic) curve of Logistic regression model is AUC (Area under concentration-time curve) = 0.5142 ± 0.1095, standard naive Bayesian model is AUC = 0.5899 ± 0.1063, and improved naive Bayesian model is AUC = 0.7721 ± 0.0865. The difference has statistical significance ($P < 0.0001$), which shows that the improved naive Bayesian model can better distinguish between the cured schizophrenics in the treatment process. Namely, the performance of the improved Naive Bayesian model applied in the auxiliary diagnosis system is better than that of the other two models.

The aim of this study is to design an auxiliary diagnosis system for schizophrenia based on Bayesian network. Firstly, the data requirement of the auxiliary diagnosis

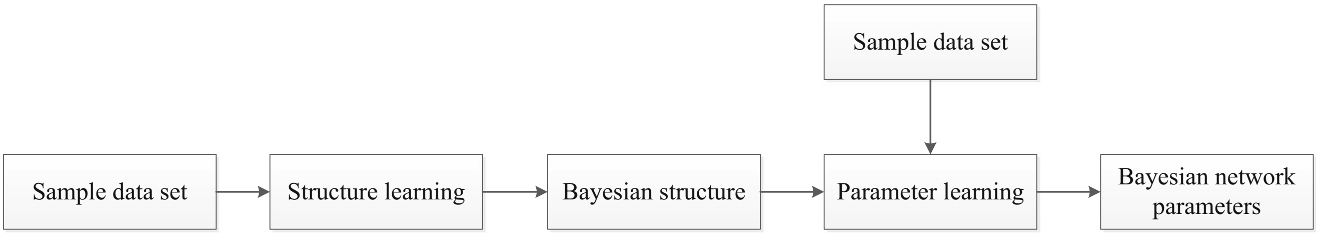


Fig. 2 Work flow chart of Bayesian network construction

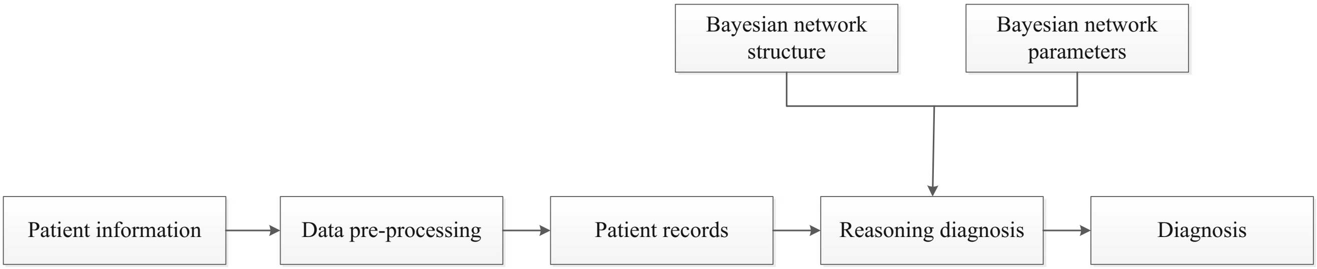


Fig. 3 Auxiliary diagnosis workflow

system is elaborated in detail, and the business process requirement of the auxiliary diagnosis system is discussed based on naive Bayesian model and functional method. At the same time, the functional requirements of the auxiliary diagnosis system in the application of the model and the relationship between the processes in each stage are discussed and analysed. Moreover, case data are selected to verify and analyse the model system, so as to further deepen the design and research of the reasoning and diagnosis function of the subsequent diagnosis system. Finally, some other requirements constraints of the auxiliary diagnostic system are simply supplemented and explained.

Overall design of auxiliary diagnosis system for schizophrenia

Based on the consideration of doctors at all levels, the software architecture design of schizophrenia auxiliary diagnosis system can truly reflect and meet users' needs for software, thus improving the software requirements and quality of software design. It is a bridge between software requirements and software design.

Considering that the overall framework design of the auxiliary diagnosis system for schizophrenia is an important process, synthesizing the design principles of the auxiliary diagnosis system, as shown in Fig. 6, the overall functional structure design of the system is that the client of the doctor receives the pathological parameter detection data, and the patient self-test data, doctor measurement data, etc. collected from the detection instrument through the network. After the initial treatment by the client, the information is displayed on the display screen, and the data is transmitted to the hospital group server through the network. Combined with schizophrenia knowledge in the knowledge base and the comparison of expert cases, the doctor client finally gives diagnosis and treatment suggestions.

All sample data are transmitted by network in this system, and real-time remote diagnosis can be carried out among network users at the same time.

Results and discussion

Design of data pre-processing

According to the previous research and the characteristics of the original data of schizophrenia patients, as well as the goals of data pre-processing in various formats of users or patients, and the process requirements of data pre-processing, the data pre-processing system needs to meet the needs of data integration, feature screening, data discretization, feature reduction and missing value processing and other data management functions. That is to say, the data pre-processing function can be further divided into different sub-functional modules as shown in Fig. 7.

Data integration function is mainly to integrate the feature information of the same patient from different data sources into a record; the feature screening function is to remove the redundant feature information according to the opinions of doctors and experts, and preliminarily screen the valuable features; discretization is to discretize the continuous feature values. Feature reduction is to use the algorithm to reduce the dimension of the data, to further

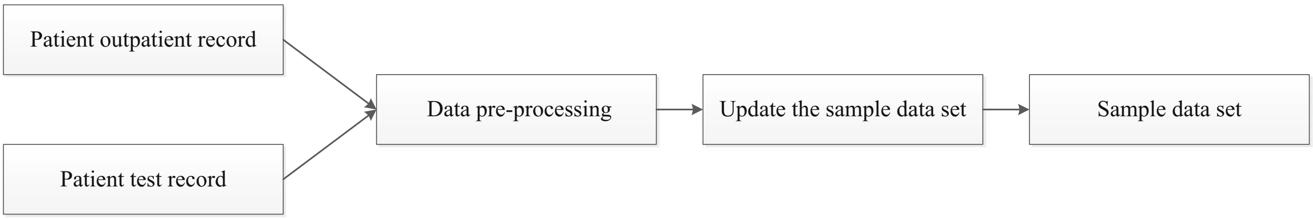


Fig. 4 Workflow of updating sample base

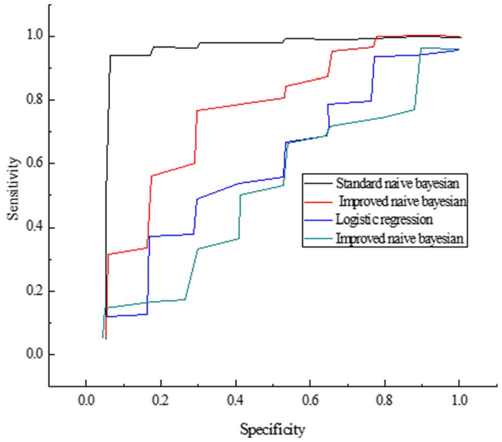


Fig. 5 Resolution performance

process and screen out valuable feature information, and to deal with missing feature value information.

Data integration processing mainly integrates the same patient data in different source data formats into the same record, and eliminates duplicate data records. According to the characteristics of patient data, the processing functions of conversion and merging should be included, as shown in Fig. 8.

The integrated data may also have the same or similar feature values such as patient's name and name, patient's ID and ID, or the features such as education level, origin, and occupation, which are totally helpless for disease prediction. It is necessary to remove these feature values so that they can contain all of efficient information and concise data record.

According to the need of experts' opinions, the similar features predicting schizophrenia will be integrated into a new feature, and the original features will be removed, so that it can be more efficiently operated. For instance, family history prediction of schizophrenia patients will be very helpful.

Design of Bayesian network construction module

In order to ensure the accuracy of the auxiliary diagnosis system, the system adopts three different structures of Bayesian networks, namely NB (Naive Bayesian Network), INB (Improved Naive Bayesian network) and AINB (the new Bayesian network model proposed). Among them, AINB model is a new Bayesian network model proposed here. Based on the practice of INB model in this system, it is found that INB model cannot make better use of medical expert's experience to assist the diagnosis of schizophrenia. Therefore, AINB Bayesian network is proposed based on INB. Therefore, the construction of Bayesian network here includes three types of modules: NB, INB and AINB. The module construction structure is shown in Fig. 9.

Because each type of Bayesian network consists of two parts: structure and parameters, the construction of each Bayesian network includes two processes: the determination of the structure of Bayesian network and the calculation of parameters of Bayesian network. The structure and parameters of NB can be obtained directly from sample data sets, and NB network is also the simplest type of Bayesian network, so it is usually constructed in reasoning.

After completing the above operations, case operation can be carried out and current cases can be operated. When data input is completed, the system automatically matches similar cases according to intelligent algorithm and calls similar diagnosis scheme in database for doctors' reference. At the same time, according to the input patient information, intelligent matching estimation method gives the corresponding diagnosis results, and refers to similar cases in the past to provide some diagnosis and treatment programs; the visiting record module that day shows the basic information of patients and visiting time, case creation time and date; for online information module, expert users can not only communicate directly with patients or their families, but also respond to the message questions. In the experience query module, users can gradually diagnose patients' diseases according to certain reasoning strategies based on the knowledge stored in the knowledge base.

In the design of the knowledge base of the system, the expert diagnosis results of the relevant case and the suggestions for the case can be given. The case management module is to select the cases according to the standard and put them

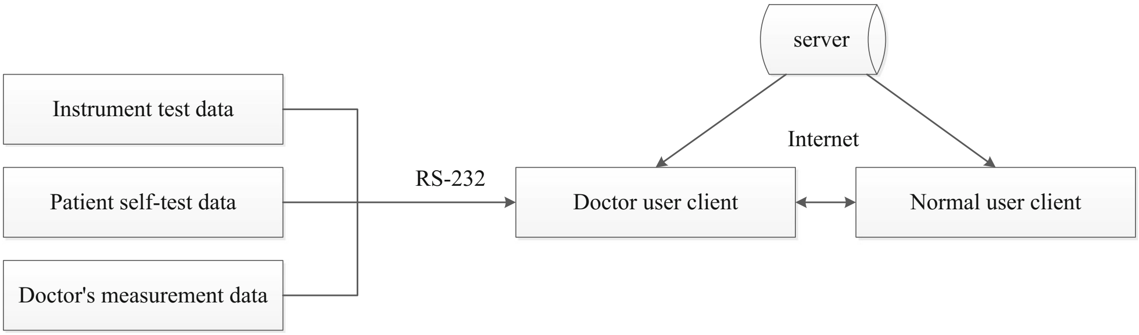


Fig. 6 System overall framework diagram

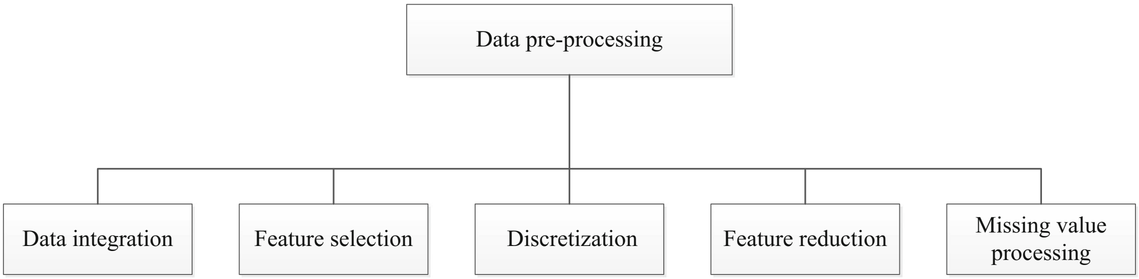


Fig. 7 Functional structure diagram of data pre-processing module

into the knowledge base to form the expert case base. The cases must be submitted through three or more doctors and users to examine and submit together before they can be successfully put into the expert case database. Ordinary cases can only be queried, modified and deleted, not used as cases for expert experience query. Exit the system after completing the operation.

Design of feature screening module

The data pre-processing module mainly processes a series of raw data to obtain the available sample data. Firstly, the overall structure of the data processing module is introduced, and then the design of each module is further elaborated. Different sub-functional modules can be further divided into different data pre-processing functions.

In the collected data of schizophrenic patients, laboratory data and outpatient data are stored in different ways. In order to facilitate the storage of the two types of data, it is necessary to convert the two types of data into a unified data format.

In the process of collecting raw data, the outpatient data of a schizophrenic patient may be stored in different rows. Some of the information in these rows may be duplicated or different variable information may be stored in different rows. The information in a data record is integrated into a data record, and redundant data information is eliminated in the merging process, which is what outpatient data row merging should do.

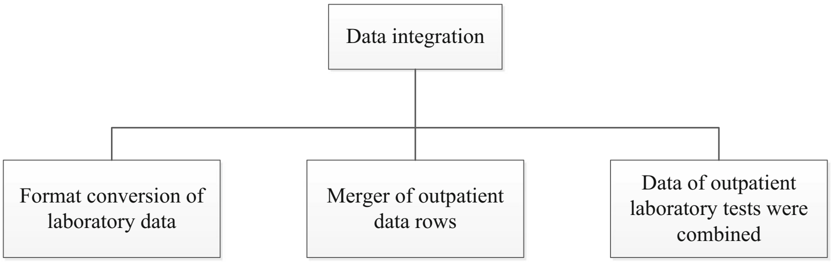
Because the original outpatient records and laboratory data of schizophrenia patients are stored in different databases, in order to get the complete data information of the same patient, it is necessary to integrate the outpatient data records and laboratory data records of the same patient in the same data information record, that is, to increase the value of laboratory data in the number of outpatient clinics and obtain the complete information of the patients.

In order to improve the efficiency and accuracy of the auxiliary diagnosis system for schizophrenia, it is necessary to remove duplicate, similar and useless features (columns) from the integrated data information, and construct a feature screening module under the guidance of medical experts.

Because Bayesian networks generally require using discrete values, in the research of this system, according to the experience of experts, the standard of laboratory indicators and the support of medical experts, the first step is to discretize the continuous eigenvalues in the original data, which is realized by program. The task of feature reduction is still to further reduce the number of features. In the pre-processing of feature screening, some features (attributes) that are obviously unrelated to the prediction of schizophrenia have been preliminarily eliminated. Nevertheless, the number of remaining features still needs to be further screened, and the attributes that are unrelated to the prediction or have little correlation need to be eliminated.

Feature reduction refers to the selection of a feature subset from all features, which makes the constructed model more excellent, also known as feature selection or attribute selection. In practical project applications, on the one hand, the more the number of features are, the higher the existence of irrelevant features or interdependent features will be, and with the increase of the number of features, the performance of the classifier will decline when it reaches a certain limit. On the other hand, due to the limited training samples obtained, with the increase of feature dimension, the demand of learning algorithm for time and space will gradually increase. In some cases, it will lead to dimension disaster, which makes the model more complex, and thus greatly reduces the reasoning

Fig. 8 Function diagram of data integration module


Data integration
Format conversion of laboratory data
Merger of outpatient data rows
Data of outpatient laboratory tests were combined

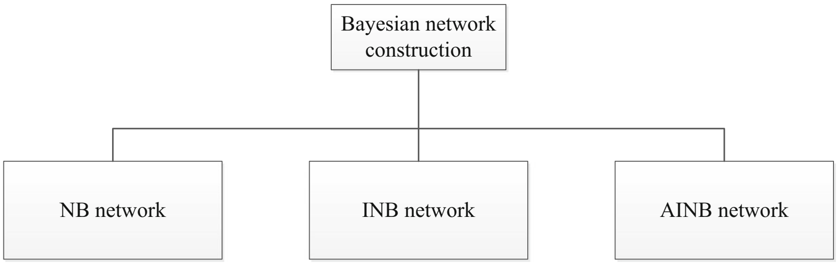


Fig. 9 Structural diagram of function modules for constructing Bayesian networks

ability and application efficiency of the model. In the general process of feature selection, an evaluation function is used to evaluate a feature subset generated from the feature set, and the evaluation result is compared with the stop criterion. If the evaluation result is better than the stop criterion, it stops; otherwise, the next feature subset is generated and the selection of feature subset is continued. Generally, the validity of the selected feature subset is verified.

Conclusion

The existing network technology, database technology and data mining technology are used to construct and design an auxiliary diagnosis system for schizophrenic patients and expert doctors engaged in diagnosis and treatment of schizophrenia. The introduction to the function and architecture of schizophrenia auxiliary diagnosis system is focused on, and the design of the function and database of schizophrenia auxiliary diagnosis system based on Bayesian network is completed. The system fully and reasonably utilizes the greatest advantages and functions of experts to help more schizophrenic patients. The system can effectively reduce the misdiagnosis rate of schizophrenia and early detect the disease and predict the development of the disease, providing more professional services for patients.

Funding There was no dedicated funding regarding this study.

Compliance with Ethical Standards

Conflict of Interest Author Xiaohong Wang declares that he has no conflict of interest. Author Na Zhao declares that he has no conflict of interest. Author Peng Ouyang declares that he has no conflict of interest. Author Jiayi Lin declares that he has no conflict of interest. Author Jian Hu declares that he has no conflict of interest.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

This article does not contain any studies with animals performed by any of the authors.

Informed Consent Informed consent was obtained from all individual participants included in the study.

References

1. Sacchi, S., Pieragnoli, P., and Ricciardi, G., Impact of haemodynamic SonR sensor on monitoring of left ventricular function in patients undergoing cardiac resynchronization therapy. EP Europace 19(10):1695–1699, 2016.
2. Quintero, D. G., Taylor, R. B., and Miller, M. B., Air-abrasive disinfection of implant surfaces in a simulated model of periimplantitis. Implant Dentistry 26(3):423, 2017.
3. Wang, X., Lian, Y., and Wang, X., Study of regional left ventricular longitudinal function in fetuses with gestational diabetes mellitus by velocity vector imaging. Echocardiography 33(8):1228–1233, 2016.
4. Versiani, M. A., Ordinola-Zapata, R., and Keleş, A., Middle mesial canals in mandibular first molars: A micro-CT study in different populations. Archives of Oral Biology 61:130–137, 2016.
5. Tardif, R., Catto, C., and Haddad, S., Assessment of air and water contamination by disinfection by-products at 41 indoor swimming pools. Environmental Research 148:411–420, 2016.
6. Ojaghi-Haghighi, Z., Mohebbi, B., and Moladoust, H., Left ventricular torsional parameters before and after atrial fibrillation ablation: A velocity vector imaging study. Electronic Physician 9(9): 5395–5401, 2017.
7. Chauveau, S., Anyukhovsky, E. P., and Benari, M., Induced pluripotent stem cell-derived cardiomyocytes provide in vivo biological pacemaker function. Circulation Arrhythmia & Electrophysiology 10(5):e004508, 2017.
8. Lo, Q., Haluska, B., and Chia, E. M., Alterations in regional myocardial deformation assessed by strain imaging in cardiac amyloidosis. Echocardiography 33(12):1844, 2016.
9. Shen, Y., Xu, J., Li, Z., Analysis of gut microbiota diversity and auxiliary diagnosis as a biomarker in patients with schizophrenia: A cross-sectional study. *Schizophr. Res.* 197, 2018.
10. Ildiz, G. O., Arslan, M., and Unsalan, O., FT-IR spectroscopy and multivariate analysis as an auxiliary tool for diagnosis of mental disorders: Bipolar and schizophrenia cases. Spectrochimica Acta Part A Molecular & Biomolecular Spectroscopy 152:551–556, 2016.
11. Zhang, L. X., Gu, W. J., Li, Y. J., Wang, Y., Wang, W. B., Wang, A. P. et al., PTH is a promising auxiliary index for the clinical diagnosis of aldosterone-producing adenoma. American Journal of Hypertension 29(5):575–581, 2015.
12. Qi, X. K., Liu, J. G., Li, C. Q., and Ning, B. O., 27 cases of atypical viral encephalitis with obvious psychiatric symptom but negative auxiliary diagnosis. Journal of the Neurological Sciences 357: e115–e116, 2015.