Research on Intelligent Decision of Pulmonary Tuberculosis Disease Based on Data Mining

Guifen Chen, Wang Ke, Ma Li

Research on Intelligent decision of pulmonary tuberculosis disease based on Data Mining

Guifen Chen,Wang Ke,Ma Li(✉)

College of Information and Technology Science, Jilin Agricultural University,Chang Chun, Jilin, China

1045093555@qq.com

Abstract. Aiming at the problem that the low diagnostic efficiency and low accuracy of the single data mining method for Diagnosis of pulmonary tuberculosis, In this study, the electronic records of 1203 cases of tuberculosis patients in Changping District City, Beijing City of Beijng and Beijing Institute of tuberculosis control and tuberculosis control were build, Tuberculosis disease diagnosis model is built by application of rough set and decision tree method, On the basis of this, the diagnosis system of pulmonary tuberculosis was constructed. In this study, The combining method of rough set and decision tree was approached to attribute reduction, the model reduced redundant 57attributes and remained 22 attributes, and articled 7 the decision rules. The model accuracy is 89.46%. Compared with the non reduction method, the decision rule was reduced by 128%, and the accuracy of the model remained unchanged. The research results showed that the algorithm can reduce the time and space complexity of the algorithm while ensuring the accuracy of the model, so as to improve the efficiency of the mining, and provide some references for clinical diagnosis.

Key words: pulmonary tuberculosis disease, rough set,decision tree, intelligent diagnosis

1 Introduction

The digitization of medical equipment and instruments makes the information capacity of the hospital database expand, including a large number of clinical information about the patient's medical history, diagnosis, examination and treatment[1]. How through the efficient, intelligent algorithm of massive pulmonary tuberculosis disease diagnosis and treatment data data mining. According to the hidden relationship between treatment outcome and medical records data, looking for a reliable and feasible method for diagnosis and treatment of that early and effective for medical personnel to provide for auxiliary diagnosis and treatment, has important clinical significance[2-4].

Pulmonary tuberculosis is a chronic respiratory disease which is seriously harmful to human health. At present ,the diagnosis of typical tuberculosis is mainly through the observation of the clinical manifestations, sputum sputum, chest imaging, bronchoscopy judgments[5].How to treatment of TB suspicious symptoms and suspected pulmonary tuberculosis patients with reasonable examination and early diagnosis and treatment, Reduce the further spread of Mycobacterium tuberculosis[6]; How to get rid of the single index, establish the patient's multi modal clinical information, and dig out the clinical index which is closely related to the pathology, to realize the clinical identification of pulmonary tuberculosis is one of the important clinical needs in the Department of respiration[7].

This paper attempts to improve the existing mining algorithms according to the data features of pulmonary tuberculosis, by the combination of rough set and decision tree method. Attribute reduction is carried out, and the decision tree rule set is extracted, Mining the data of medical records with implicit diagnosis rules, to obtain new knowledge discovery, to provide reference for clinical treatment of tuberculosis patients. The method of mining the medical record data with implicit diagnosis rules, obtain new knowledge discovery, to provide reference for clinical treatment of tuberculosis patients.

Guifen Chen (1956-),College of Information Technology, Jilin Agricultural University,ChangChun,China

2 Data Mining Foundation

2.1 Rough Set Theory

Rough set theory is the core of knowledge and approximate set,and deals with imprecise, uncertain and incomplete data. In the problem solving process, it is not required to provide any prior information of the problem data set, so it can deal with the uncertainty problem more objectively[8].

Rough set theory can describe knowledge with four yuan ordered groups, namely: k = (U, a, V, d). Where u is the universe; A is all attributes; V= ∪a∈ A Va, Va is the attribute range; d: U× A→ V is a function of information, dx: A→V，x∈ U， says the object x in K complete information, in which dx (a)= d(x , a). For such information systems, each attribute subset defines an equivalence relation on the domain, namely B A, defined RB: xRB y dx (b) = dy (b) , b∈ B[9].

Note by the equivalence relation attribute set B⊆A T, is derived for the RB. a∈ A T， if RAT = RA T\ {a}, said the a attribute is redundant; If there is no redundant attribute in the system, then the AT is independent; IfRB = RAT and B is not redundant subset of attributes, then B AT is called A T reduction, denoted as red (AT), The intersection of all reduction of A T is called the A T kernel, denoted as core (AT) [10].

2.2 Decision Tree Method

The decision tree method represents as a decision tree on the basis of the function that will be learned from a set of training data by a large amount of data to be classified according to a certain target, from which to find useful, potential information, commonly used in the classification of the prediction algorithm[11]. Decision tree method has the characteristics of high speed, high precision, simple model and so on. It has a wide application in data mining[12, 13].

The decision tree is constructed including two steps: generating decision trees and decision tree pruning.

The decision tree is generated from a root node, which is to construct a tree from top to bottom by constantly dividing the sample into subsets[14]. The test value for each attribute is represented as a non leaf node on the tree. Each result is represented as a branch of the tree, and the final classification class is the leaf node of the tree. In the decision tree structure, the information gain is used as the criterion for dividing the

nodes[15].

Because of noise and outliers, so the decision tree can cause abnormal branching, so need pruning strategy[16]. In the decision tree pruning, usually choose leaf nodes to replace one or more sub tree, then select the class with the highest probability for the node of the category, can also replace the subtree with the branches.

3 Model Construction

3.1 Construction of Electronic Medical Records

Original sample data are form the medical records of Changping District Beijing tuberculosis prevention and Beijing Tuberculosis Control Research Institute,and data acquisition time for November 2015 ~ 2016 years 5 months, we uses Microsoft SQL2010 to integrate data from different data sources, which involves 1203 copies of disease history archives. And the data of this study are mainly three categories:

(1)Patient general information：①medical record number②sex③birthday④Group (initial treatment group and re treatment group), ⑤The types of household registration (the city and other provinces), ⑥Ethnic (Han, Hui, Manchu and other),、⑦Contact history (yes,no), Previous associated with other diseases (diabetes, silicosis, hepatitis, epilepsy, lung cancer, lung infection, pulmonary heart disease, chronic bronchitis, other)⑨ Previous lung combined with the history of tuberculosis pleurisy, tuberculosis of lymph node, bone tuberculosis, cutaneous tuberculosis, renal tuberculosis, peritoneal tuberculosis, pelvic tuberculosis, tuberculosis of the intestines, fallopian tube tuberculosis)

(2)Main symptoms before treatment ： ① Cough, expectoration less than 2 weeks ② Cough and sputum >2 weeks ③ Hemoptysis / sputum with blood ④ pectoralgia ⑤ afternoon fever ⑥ night sweat ⑦ feeble ⑧ anorexia ⑨ lose weight⑩irregular menses⑪Physical examination found no symptoms⑫other

(3)Check the project before treatment： ①erythrocyte sedimentation rate (ESR)② C reactive protein（no，yes）③Sputum smear before treatment (Not checked, the results)④ Before treatment sputum culture (Not checked, the results) ⑤ Rapid culture of Mycobacterium tuberculosis Not checked, the results) ⑥ Identification of culture positive patients (traditional / rapid) (not checked, the results) ⑦ Tuberculin test(not checked, the results), ⑧Tuberculosis antibody (not checked, the results) T-SPOT（not checked, the results）⑩ Sputum Mycobacterium tuberculosis PCR (not checked, the results), ⑪ Sputum Mycobacterium tuberculosis Hain test (not checked, the results)

12 Sputum detection of Mycobacterium tuberculosis X-pert (not checked, the results), 13 The blood tumor marker examination (not checked, the results) 14 bronchoscopy （not checked, the results）15 biopsy（Lung tissue / pleural / pleural effusion）（not checked, the results）16（not checked, the results）17 CT（not checked, the results） (4)The final diagnosis (pulmonary tuberculosis / pulmonary tuberculosis, pleurisy, not NTM)。The raw data is shown in figure1。

sex	date	group	Household type	race	Contact history	Hemoptysis	Chest pair
male	05/11/1999	Initial treatment group	This city	Han	no	no	no
female	01/12/1986	Initial treatment group	other provinces	Han	no	no	no
male	10/03/1987	Initial treatment group	This city	Han	no	no	yes
male	12/05/1990	Initial treatment group	other provinces	Han	no	no	yes
male	01/02/1994	Initial treatment group	other provinces	Han	no	no	no
female	04/11/1942	Initial treatment group	This city	Han	no	no	no
female	01/07/1949	Initial treatment group	other provinces	Han	no	yes	no
female	05/06/1939	Initial treatment group	This city	Han	no	no	no
female	06/01/1986	Initial treatment group	other provinces	Han	no	no	no
female	20/11/1991	Initial treatment group	other provinces	Han	yes	no	no
male	12/06/1934	Initial treatment group	This city	Han	no	no	no

Fig. 1 part of the original data

3.2 Tuberculosis Diagnosis Based on Data Mining

3.2.1 Data Preprocessing

Because medical records are manually entered into the text or database by doctors or non medical professionals, there will be a large number of records, records are not uniform, record errors and noise, and therefore need to pre process the data of the original medical records[17]. Data preprocessing mainly includes case data acquisition, attribute selection, discretization of continuous attributes, noise and missing data in data processing, case selection and so on[18]. In order to further data mining, it is necessary to encode the values and fields in the information data table. The coding of the tuberculosis disease is shown in Table 1, and the pre processed data is shown in table 1.

Table 1 part of Pulmonary tuberculosis disease attribute code table

Attribute	Data discretization and coding	Attribute	Data discretization and coding
sex	(1) male (2) female	Cough and expectoration	(1) no (2)yes
Grouping	(1) the treatment group (2) in the first group	Cough, expectoration less than 2 weeks	(1) no (2)yes

Household type	(1) the city (2) other provinces	night sweat	(1) no (2)yes
Nation	(1) Han (2) Hui (3) Man (4) others	afternoon fever	(1) no (2)yes
Contact history	(1) no (2)yes	Hemoptysis / sputum with blood	(1) no (2)yes
Prior to the merger of other diseases	(1) no (2)yes	Sputum smear before treatment	(1) not checked, (2)the results
Diabetes	(1) no (2)yes	Before treatment, sputum culture	(1) not checked, (2)the results

3.2.2 Attribute Reduction Based on Rough Set

There are many attributes in the medical record data, and there is often a certain degree of dependence between each attribute, which can not be simply deleted. Reduction under the premise without loss of information, can simply indicate the decision system，and the decision attributes set dependencies of condition attribute set to remove unnecessary attributes from the condition attributes， and simplify the condition attribute, and improves the efficiency of data mining[19，20]. In this paper, the remaining attributes are 22, and the redundant attributes are reduced by 54, and the attributes of the reduced attributes are shown in Table 2.

Table 2 attributes after reduction

General information of patients (3 items)	Contact history, diabetes, tuberculosis pleurisy
Main symptoms before treatment (5 items)	Cough, sputum or hemoptysis / 2 weeks, bloody sputum, chest pain, physical examination found, without any other symptoms.
Check project development before treatment (14 items)	Erythrocyte sedimentation rate, C-reactive protein value, before treatment of acid fast bacilli in sputum culture have to do as a result of, sputum Mycobacterium tuberculosis rapid culture results has been done, tuberculin test, tuberculosis antibody has something to do with the results, interferon gamma release test /T-SPOT, sputum

	Mycobacterium tuberculosis Hain test, sputum Mycobacterium tuberculosis X-pert detection, blood tumor mark examination, bronchoscopy, biopsy (lung tissue / pleura in water), treatment of chest DR is to do as a result of, before treatment chest CT have to do as a result of

3.2.3 Decision Tree Model

The C5.0 algorithm is used to construct decision tree, importance degree sequence is as follows according to the information gain to establish attribute: before treatment of acid fast bacilli in sputum culture have to do as a result of, rapid culture of Mycobacterium tuberculosis in sputum results and treatment of chest DR is the, tuberculosis antibody, sputum Mycobacterium tuberculosis X-pert detection, tuberculin test, gamma - IFN release test /T-SPOT. (lung tissue biopsy / pleura in water). This and clinical diagnosis of pulmonary tuberculosis patients diagnosis and treatment standard is basically the same. The results are shown in Figure 2

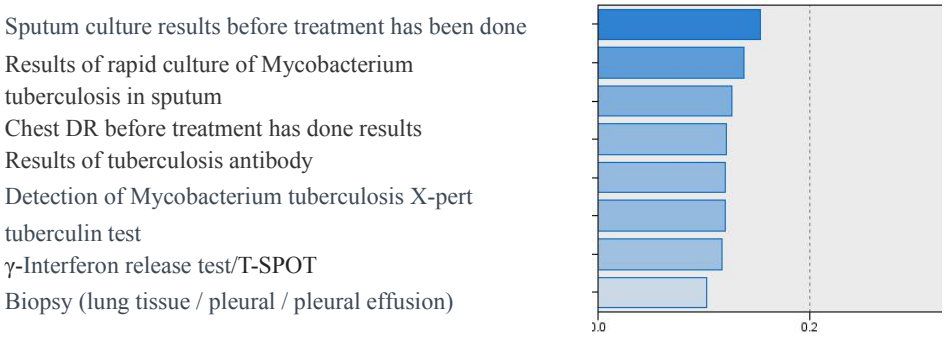

Fig.2 attribute importance ranking

According to C5.0 decision tree construction, in order to avoid the effect of sampling errors of a single partition of the results, improve the accuracy of the model, the decision tree model uses of ten fold cross validation, model accuracy is 83.46 percent, support data number is 1004. and the 7 decision rules are dig out. In accordance with the support, confidence in descending order as shown in Table 3.

Table 3 in accordance with the support and confidence of the rules results in descending order

Rule	instance numbers	Confidence(%)
If before treatment sputum culture results > 1 then Pulmonary tuberculosis / pleurisy	269	98.2
If DR results <= 2 then Not a pulmonary tuberculosis	234	96.7
If tuberculosis antibody results > 1 and DR Results > 2 then Pulmonary tuberculosis / pleurisy	32	94.1
If tuberculin test < = 1 and tuberculosis antibody result < = 1 and γ-interferon release test /T-SPOT < = 1 and sputum Mycobacterium tuberculosis X-pert detection > 1 and DR results > 2 then Pulmonary tuberculosis / pleurisy	1050	85.3
If DR results >2 then pulmonary tuberculosis /pleurisy	121	74.1
If before the treatment of acid fast bacilli in sputum culture result < = 1 and Mycobacterium tuberculosis in sputum rapid culture result < = 1 and tuberculin test < = 1 and γ-interferon release test /T-SPOT < = 1 and biopsy tissues (lung / pleural / pleural)< = 1 then Not a pulmonary tuberculosis	338	71.5
If before treatment of acid fast bacilli in sputum culture result < = 1 and tuberculin test < = 1 and γ-interferon release test /T-SPOT < = 1 then not a pulmonary tuberculosis	593	60.7

3.3 Results Analysis

Excavated from the rules, most of the samples are concentrated in device reliability 60% - 99% of the value, include the interval from the actual clinical diagnostic results, the number of instances and confidence degree is high, i.e., a strong association rules data and of pulmonary tuberculosis disease diagnosis has higher clinical value in value. This research model has been applied to the Changping District Beijing tuberculosis prevention and Control Institute and Application of Tuberculosis Control

Research Institute in Beijing, and the effect is good.

In this study, the combination of rough set and decision tree method and a single decision tree method are compared in terms of rule number, accuracy, and confidence interval and modeling time. The comparison results are shown in Table 4. From the table, we can see that a single decision tree method is extensive in classification, there are still some redundant attributes ,that resulting in the construction of a large decision tree size, more rules are extracted, and mining efficiency is not high. In this study, we use existing pulmonary tuberculosis disease archives data, by combining rough set and decision tree optimization algorithm for screening of 22 attribute variables to establish TB prediction and classification model and delete redundant attributes, simplifying the decision model, the mining efficiency is improved. Redundant attributes are removed, the decision model is simplified, and the mining efficiency is improved.

Table 4 Comparison of model results

Model	rules number	accuracy rate	confidence interval	modeling time
Combinatorial optimization method	7	83.46%	60%–99%	0.02
Single decision tree method	16	82.18%	63% -98%	0.25

4 Establishment of the Decision System

According to the function and application requirement of the diagnosis of the disease, this paper constructs the system architecture of the tuberculosis disease diagnosis system, which includes three levels: data layer, service layer, application layer, as shown in figure 3, In System construction, Java is used as the development language, B/S mode is used as development model, the application of Spring3 +Hibernate4 +Struts2 model is used to develop. Development platform is : the operating system is Windows 7 Professional Edition, server is Tomcat_7.0.14, the data base tool is MySQL Server 5.0, using JDBC is use to link database, LINQ to SQL is used to realize data access layer.

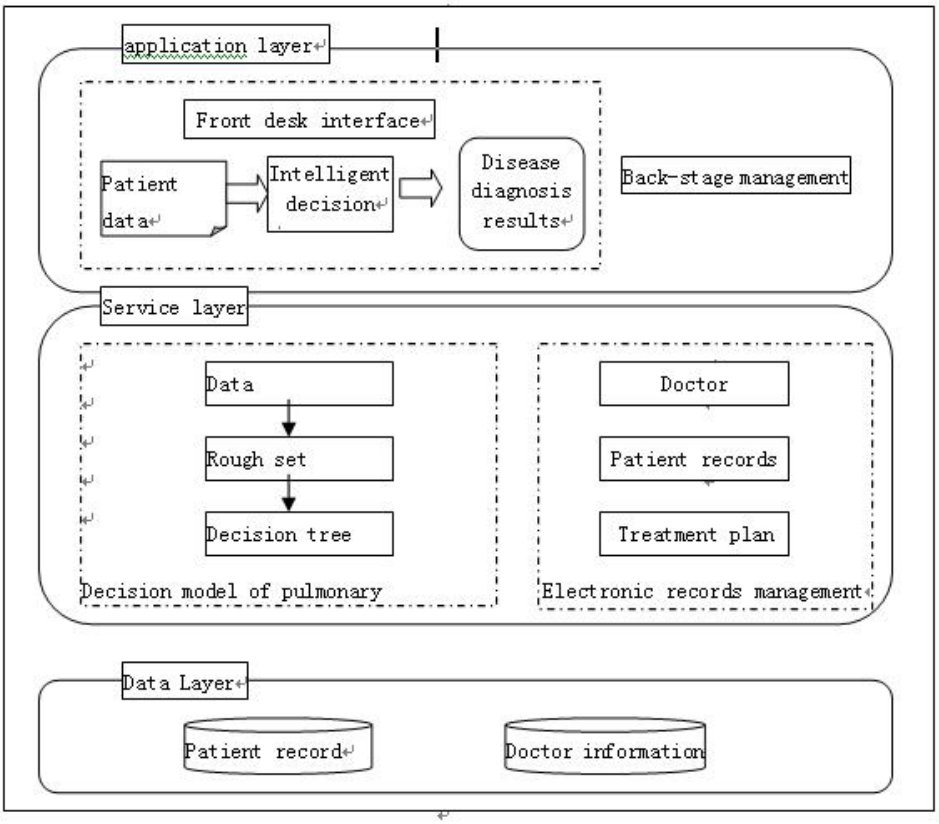

Fig 3 the system architecture of the diagnosis of pulmonary tuberculosis

(1) Application layer: this layer is at the top of the platform, mainly to the user provide software application service and user interaction interface. This layer provides services platform portal website, the web form is to show the various service functions in a unified interface and operation to the user.

(2) Service layer: This layer is the core part of the service platform of medical information resource sharing, main contents include pulmonary tuberculosis disease decision model and management of electronic file. The pulmonary tuberculosis disease decision model includes model building and visualization services, electronic archives management includes query, audit, data collection and so on medical information service. The release of medical information service is mainly based on Web services.

(3) Data layer: this layer mainly use database virtualization technology, middleware technology, a unified scheduling and allocation of medical information resources in each node of the database resources. Data layer, including all kinds of electronic information resources, the layer is the core task of data synchronization, data management, data reduplication, data backup.

5 results and discussion

Medicine development has been the experience of medical and medical experiments to evidence as the basis of evidence-based medicine; this produces the large amount of medical data and objectivity. Clinicians should be combined with history, clinical symptoms, and gradually learn to use large samples to establish the model of data

mining, and analyze the condition, and make reasonable treatment and predicting disease development[21].Computer aided medical data mining implements the redundancy elimination of medical data, standardized storage and data seamless integration and sharing, knowledge extraction automation and visual expression and other functions[22].

Decision tree method of data mining in the diagnosis of pulmonary tuberculosis has been applied, such as Zhangqi "decision tree model for classification of tuberculosis treatment regimens and pre sentenced", which illustrate that the data mining method is suitable for pulmonary tuberculosis disease diagnosis and classification problems, but single application of decision tree constructing diagnosis model, there are a large number of research variables included, which may cause problems in the test of efficacy decreased.

This study uses existing pulmonary tuberculosis disease archives data, by combining rough set and decision tree optimization algorithm for screening of 22 attribute variables establish TB prediction and classification model, this model deletes redundant attributes, improves the efficiency of mining, and provide a reference for clinical diagnosis. In addition, the research ideas and methods of this study can also apply to the choice of regimen in the treatment of other chronic diseases, such as hypertension, diabetes.Practice has proved that the optimization algorithm based on the combination of rough set and decision tree can effectively deal with the data of uncertainty reasoning, is a powerful tool for data mining, and in the future of medical data mining and utilization has broad application prospects.

Acknowledgments.

Funding for this research was provided by The national Spark plan "based on the IOT of maize precision technology integration and demonstration" (2015GA66004)

References

1.Chen Chuntao. Construction of digital hospital information system and empirical study [D]. Huazhong University of Science and Technology, 2008

2.Wang Xin, Weng Weng, Zu Aihua, Guo Yanfei, Zhou Ting, Chen Weihong. A case control study on the risk factors of pulmonary tuberculosis [J]. industrial health and occupational disease, 2011,04:208-213.

3.Zhang Qi, Zhou Lin, Chen Liang, Zhang Jinxin, Wen Xingxuan, He Xianying. Decision tree model for the classification and prediction of tuberculosis treatment program [J]. Journal of the Chinese Journal of disease control, 2015,05:510-513.

4.Ren Zheng Hong. The epidemiological characteristics and trends of the time of onset of pulmonary tuberculosis in China from 2005 to 2011 [J]. Chinese health statistics, 2013,02:158-161.

5.Chen Dachuan, Wang Zaiyi. Research progress in the diagnosis of pulmonary tuberculosis [J]. Journal of clinical pulmonary, 2016,01:145-148.

6.Wu Tengyan. Study on the application effect of tuberculosis control mode in Guangxi [D]. Guangxi Medical University, 2014

7. Lu. Vega electronic medical records system and the data of diagnosis and treatment of chronic respiratory diseases [D]. mining research of Shanghai Univer, 2015

8.Chen Guifen, Ma Li, Dongwei, Xin Mingang. Clustering, rough set and combination algorithm of decision tree in soil fertility evaluation [J]. China Agricultural Science and 2011,23:4833-4840.

9. Wang Guoyin, Yao Yiyu, Yuhong. Journal of rough set theory and Application Research on [J]. computer, 2009,07:1229-1246.

10. Zhang Ming. Research on the method of knowledge acquisition and reduction in rough set theory [D]. Nanjing University of Science and Technology, 2012

11.Feng Xinghua. Research on fuzzy decision tree algorithm based on axiomatic fuzzy set [D]. Dalian University of Technology, 2013

12.Shi Shan. Research on Network Intrusion Detection Based on decision tree [D]. algorithm C4.5 Soochow University, 2012

13.Wang Jiali. Application of data mining in financial diagnosis [D]. Guangxi University, 2012

14.Zhang Rui. ID3 decision tree algorithm analysis and improvement of [D]. Lanzhou University, 2010

15.Ding Wenbin. Network anomaly detection and filtering based on decision tree classification [D]. Electronic Science and Technology University, 2013

16.Chen Jian, Zhang Guanghua, Lin Qiang, Wang Juan, Jiang Suyong. Decision Tree Mining Journal of research and application of technology in the surgical diagnosis [J]. Foshan Institute of science and Technology (NATURAL SCIENCE EDITION), 2014,06:46-50.

17. main Fu Yang. The design and implementation of electronic medical record system [D]. University of Electronic Science and technology, 2013
18. Wang Xiong. Research and design of electronic medical record system based on intelligent medical record editor [D]. Hunan University, 2013
19.Teng Shuhua. Rough set theory of uncertainty measure and attribute reduction method of [D]. based on the National Defense University of science and technology, 2010
20. Wu Jun. Based on rough set theory for preliminary diagnosis and decision support of disease [D]. Anhui University of Technology, 2013
21. Zhang Hehua, Sun Yongqiang, Zhaoyu Hong. Evidence based medicine in intensive care informatization development [a]. The Chinese Medical Association of Chinese Medical Association, Chinese medicine will branch of medical informatics. Chinese medicine will be the twenty-first national medical information academic conference papers sink series [J]. The Chinese Medical Association, Chinese Medical Association, Chinese medicine will Medical Informatics Association: 2015:3.
22. Yao yuan. The design and application of the chronic disease diagnosis and treatment information system of military hospital [D]. Chinese people's Liberation Army Military Medical Science Academy of the PLA, 2014