International Journal of Internet, Broadcasting and Communication Vol.13 No.3 92-103 (2021)
http://dx.doi.org/10.7236/IJIBC.2021.13.3.92

IJIBC 21-3-12

Design and Implementation of Intelligent Medical Service System Based on Classification Algorithm

Yu Linjun[1], Yun-Jeong Kang[2], Dong-Oun Choi[3*]

[1]*Assistant Professor, School of Electronic Commerce, Jiujiang University, Jiangxi, China*
[2]*Assistant Professor, College of Convergence Liberal Arts, Wonkwang University, Republic of Korea*
[3] *Professor, Department of Computer Software Engineering, Wonkwang University, Republic of Korea*
E-mail ： ec@jju.edu.cn, yjkang66@wku.ac.kr, cdo209@wku.ac.kr

Abstract

With the continuous acceleration of economic and social development, people gradually pay attention to their health, improve their living environment, diet, strengthen exercise, and even conduct regular health examination, to ensure that they always understand the health status. Even so, people still face many health problems, and the number of chronic diseases is increasing. Recently, COVID-19 has also reminded people that public health problems are also facing severe challenges. With the development of artificial intelligence equipment and technology, medical diagnosis expert systems based on big data have become a topic of concern to many researchers. At present, there are many algorithms that can help computers initially diagnose diseases for patients, but they want to improve the accuracy of diagnosis. And taking into account the pathology that varies from person to person, the health diagnosis expert system urgently needs a new algorithm to improve accuracy. Through the understanding of classic algorithms, this paper has optimized it, and finally proved through experiments that the combined classification algorithm improved by latent factors can meet the needs of medical intelligent diagnosis.

Keywords: *Expert system, Medical diagnosis, Latent factor, Association rule algorithm*

1. Related theories and technologies of smart medical service system

1.1 Expert system

The expert system contains a large amount of expert-level knowledge, based on knowledge and rules for reasoning, simulating the thinking process of human experts, so as to solve problems in the professional field. The structure is shown in Figure 1.

Manuscript Received: June. 23, 2021 / Revised: June. 29, 2021 / Accepted: July. 3, 2021
Corresponding Author: cdo209@wku.ac.kr
Tel:***－****－****
Professor, Department of Computer Software Engineering, Wonkwang University, Korea

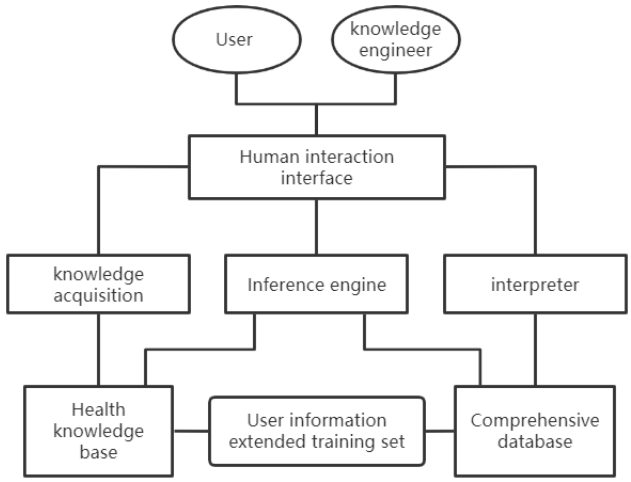

Figure 1. The structure of the expert system

The classic expert system structure consists of six modules: human-computer interaction interface, knowledge base, inference engine, interpreter, comprehensive database and knowledge acquisition. Its core is the knowledge base and reasoning engine. The knowledge base is used to store the knowledge provided by the experts. The quality of the knowledge base and the maintenance and upgrade after the system runs largely determine the performance of the entire expert system. The inference engine combines the current input with the middleware trained by the knowledge in the knowledge base to perform inference and solution, and finally gives an answer that is similar to or even surpasses human experts. Expert system is by no means an ordinary computer application. It has the ability to reason and learn to evolve that ordinary computer applications do not have. The biggest difference is that ordinary computer programs will pass established computer logic programs when running under exactly the same conditions. Obtain accurate and fixed results, but the process of obtaining the results after the expert system through knowledge inference is more similar to the results obtained by the human thinking process is not necessarily correct.

1.2 Knowledge Base and inference engine

The core that truly supports the operation of the expert system is the knowledge base and the inference engine: the knowledge base is used to store the database of knowledge provided by experts; the inference engine imitates the process of human experts solving problems, and combines the current input with the knowledge trained in the knowledge base The middleware solution gives an answer that is similar to or even surpasses human experts. The core of this article is to study these two modules of a medical diagnosis expert system. Start with the knowledge base, and build a unique knowledge base based on the characteristics of medical data. Then use the association rule mining to research and apply the knowledge discovery of the system. Then, based on the classification algorithm, the inference engine of the medical diagnosis expert system is studied, and the combined classification algorithm is improved based on the latent factor and decision tree, and the experimental data is analyzed and compared. The knowledge base is one of the most important modules of the expert system. As the name suggests, it is a database used to store knowledge provided by experts. Its quality and the maintenance and upgrades after system operation largely determine the performance of the entire expert system. The data sources in the knowledge base must be authentic and reliable in order to conduct valuable research.

The reasoning machine imitates the thinking process of human experts in solving problems, playing the brain of the system. The current input combined with the middleware trained by the knowledge in the

knowledge base will give an answer that is similar to or even surpasses human experts.

The core process of medical diagnosis is to assign the most appropriate classification to disease data records after corresponding analysis. Based on this central idea, in 2.2, a classification algorithm is designed for the inference engine part of the medical diagnosis expert system module and its effect is tested.

1.3 Data mining technology

1.3.1 Association analysis

The occurrence of certain events in natural phenomena, economic phenomena and social phenomena is often not independent, and they are often accompanied by the occurrence of other events behind them. The hidden certain or even multiple relationships between these events are called associations between events, or associations for short. Discovering hidden and potential associations from massive data is the mining of association rules, referred to as association analysis for short.

For measuring the accuracy of association rules, confidence is generally used, and for measuring the applicability of association rules, support is generally used. There are various algorithms for mining association rules, such as Apriori and FP-Growth algorithms. Regardless of the purpose of the algorithm, it is to mine the rules that meet the minimum confidence threshold and the minimum support threshold in the data set.

1.3.2 Data mining technology based on classification analysis

Classification is based on the characteristics of the existing data set to establish a data mining technology used to map the data in the data set to the classification of a certain category. Classification plays an extremely important role in the field of data mining. Analyze the data to extract the model of the data class. This described model is called a classifier, which is specifically used to predict the class of the subsequent data classification and gives a unique class identification number. Classification technology can give samples of unknown categories to their categories, thereby providing a powerful reference for decision-making. Therefore, classification technology is widely used in systems with diagnostic and rating functions.

To build a classification model, you must first use a reliable data set with a certain scale as the training set to train the classifier. Through the analysis of the characteristics of the training set, a unique and accurate judgment description is found for each class to classify the subsequent test data. It is used to evaluate whether a classification model prediction is accurate, and the most widely recognized indicator in the industry is to comprehensively consider recall and accuracy. The classic classification algorithms include the nearest neighbor algorithm (KNN), decision tree, and neural network.

1.3.3 Latent factor algorithm

Latent factor algorithm is an award-winning algorithm in Netflix's recommendation algorithm competition, which was first used in movie recommendation. The user potential factor matrix Q represents the preference of different users for different elements. Potential factor - Movie matrix P, which indicates that each movie contains various elements. According to the formula (1).Each user's rating matrix $\breve{R}$for each movie can be obtained，So as to make recommendations.

$$\breve{R} = QP^{T} \tag{1}$$

After that, the user's interest score matrix for all movie items is obtained, and the scores of each item are sorted to preferentially recommend the items with high interest scores.

In medical diagnosis, the impact of individual patient differences is very large, mainly in the following two points: (1) The same disease may have different symptoms for different individuals; (2) The same symptoms may correspond to different symptoms for different individuals. disease. The influence of individual

differences caused by factors such as gender, age, geographic location, season, etc. on diagnosis cannot be ignored. Considering these influences in the algorithm can be better applied to medical diagnosis scenarios. The best way to embody them is to use latent factors. Latent factors are mainly derived from clinical medical data and the experience summary of experts' domain knowledge. The association rule analysis of a large number of cases can also be used as a source of hidden potential factors for various diseases.

2. The role of different algorithms in the medical diagnosis system

In this part, we studied different algorithms, combined them with the structural principles of the expert system, and flexibly used different algorithms to provide support for the main modules of the expert system.

2.1 Association rule mining and knowledge discovery

The expert system will not be static. Like humans, knowledge needs to be updated and expanded, and old knowledge may be inefficient or even wrong, so expert systems also need to have the ability to acquire knowledge. As shown in Figure 2, it is a general architecture diagram of knowledge acquisition.

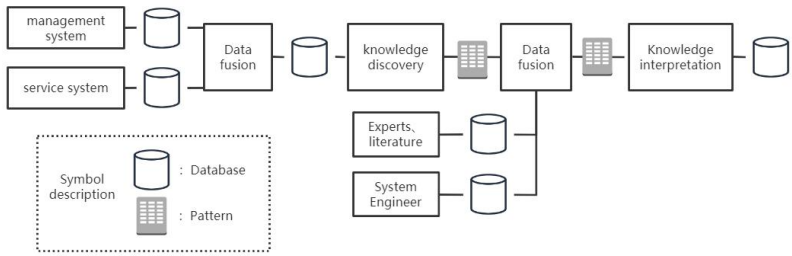

Figure 2. General architecture diagram of knowledge acquisition

In figure 2, the way of general expert system to acquire knowledge is described. Macroscopically, it is mainly divided into two ways: the system passively accepts human input and actively extracts data from literature. In 1997, usama.m.fayyad gave a widely accepted definition: knowledge discovery and data mining (KDD) refers to the identification of innovative and effective data from a large number of data High level processing of potentially valuable and understandable patterns1. In the process of KDD, the most important technology is data mining technology. Mining association rules for massive data is the most common way to implement KDD.

There are many existing algorithms for mining association rules. among them, Apriori algorithm is the most widely used association analysis algorithm.

However, the Apriori algorithm is very inefficient in generating frequent itemsets. The data file must be read every time a candidate item set is generated. This is unacceptable for the overhead in the case of massive data and will cause the read time to be much longer than the running time of the algorithm itself. . Therefore, optimization is needed. The optimization ideas are given in reference [1], which greatly reduces the computational overhead and improves the efficiency of the algorithm.

2.2 Classical classification algorithm

As mentioned in 1.2, classification algorithms can play an important role in medical diagnosis expert systems. When applying a classification algorithm, if you want to build a classification model, you must first

[1] U. M. Fayyad. Knowledge Discovery in Databases: An Overview[C]. International Workshop on Inductive Logic Programming. Springer-Verlag, 1997, 3-16

use a reliable data set with a certain amount of scale as a training set to train the classification. Through the analysis of the characteristics of the training set, a unique and accurate judgment description is found for each class to classify the subsequent test data.

The data sources in the knowledge base must be authentic and reliable in order to conduct valuable research. In the experimental part of the third subsection, the data used are all from the medical records of a Chinese hospital. As shown in Figure 3, part of the records randomly selected from more than 30,000 cases of dermatology records in 2019, out of privacy protection respect and the requirements of the data-providing institutions, the private patient information contained in it has been mosaic-processed.

Date of visit	Gender	Age	Tel.	Add.	ID card	Date of onset	Diagnosis (disease name)	First visit	Follow up visit	Handle
2019/7/23	M	51				2019/7/18	Dermatitis	√		
2019/7/23	F	28				2019/7/12	Acute urticaria	√		
2019/7/23	M	52				2019/6/20	Acute urticaria		√	
2019/7/24	M	54				2019/7/18	Herpes simplex	√		
2019/7/24	M	36				2019/7/18	Onychomycosis	√		
2019/7/24	F	39				2019/7/18	Onychia lateralis	√		
2019/7/24	F	19				2019/7/18	Papular urticaria	√		
2019/7/24	M	51				2019/7/18	Dermatitis	√		
2019/7/24	F	40				2019/7/18	Onychia lateralis		√	
2019/7/24	M	44				2019/7/18	Chronic urticaria	√		
2019/7/25	M	41				2019/7/18	Dermatitis	√		
2019/7/25	F	59				2019/7/18	Chronic urticaria	√		
2019/7/25	F	36				2019/7/18	Dermatitis	√		
2019/7/25	F	45				2019/7/18	Onychomycosis	√		
2019/7/25	F	61				2019/7/18	Dermatitis		√	
2019/7/25	M	33				2019/7/18	Onychia lateralis	√		
2019/7/25	F	28				2019/7/18	Dermatitis	√		
2019/7/25	M	52				2019/7/18	Chronic urticaria	√		
2019/7/25	F	57				2019/7/18	Herpes simplex	√	√	
2019/7/25	F	46				2019/7/18	Acute urticaria	√		
2019/7/25	F	43				2019/7/18	Dermatitis	√		
2019/7/25	M	44				2019/7/18	Papular urticaria	√		

Figure 3. Data source sampling

2.3 Mining potential factors for medical diagnosis

Although, the classic classification algorithm can find some qualified rules in a large amount of data.But,in medical diagnosis, individual differences of patients have a great influence, mainly in the following two aspects

(1) The symptoms of the same disease may be different for different individuals;

(2) The same symptoms may correspond to different diseases for different individuals.

The influence of individual differences caused by factors such as gender, age, geographical location and season on diagnosis can not be ignored. Considering these influences in the algorithm can be better applied in the scene of medical diagnosis. And the best way to reflect them is to use potential factors.

Potential factors mainly come from clinical data and the experience of experts in the field of knowledge. The association rules analysis of a large number of cases can also be used as the source of hidden potential factors of various diseases. The method in 2.2 is still used to analyze the association rules based on the diagnosis records of dermatology experts in a hospital for one year, in which support means support degree and confidence means confidence degree. Due to the variety of diseases and huge data set of the actual outpatient patients, the support degree is very low. In order to mine the rules, we tolerate the low support degree and require high confidence degree, For example:

{papular urticaria}=>{children,autumn} (support=0.004186,confidence=0.557241)

{pityriasis variegata]} = > {children} (support=0.002093 ,confidence= 0.888889)

{dermatitis of cryptoptera = > autumn} (support=0.003663 ,confidence= 0.650000)

{acne} = > {teenagers} (support=0.003663 ,confidence= 0.650000)

So we think that the potential factors of papular urticaria are autumn and children, the potential factors of paederus dermatitis are autumn, tinea versicolor is more common in children, acne is more common in adolescents, which coincides with the knowledge of experts and the description of medical data. Association rules can be obtained by mining and analyzing a large number of real medical data. Combined with the reliable professional knowledge of human experts, valuable knowledge of potential factors of diseases can be obtained.

The knowledge of these potential factors is stored in the knowledge base of expert system. These potential factor data records in the knowledge base will be used to improve the classification algorithm in section 2.2.

2.4 Combined classification algorithm based on latent factor

2.4.1 Training the data set of the classifier

The application of big data system requires accuracy, consistency and credibility. At the same time, high quality data should also have good interpretability. Therefore, the design of data objects is particularly critical: using the most classic attribute vector in data mining to store data objects. The attribute vector in this system includes illness-symptom vector, user late factor vector and illness-symptom vector. illness-symptom vector, which means disease symptom attribute vector, hereinafter referred to as I-S vector, is composed of numerical attribute and nominal attribute. Each symptom of a disease is regarded as a numerical attribute, and the last nominal attribute is the name of the disease itself: structure of I-S vector:

For example, laryngitis has the following common symptoms: hoarseness, moderate cough, itchy throat, sore throat, dry throat and mild lymphadenopathy. Using a large number of real cases combined with expert knowledge and literature to extract I-S vector as the research data set, covering more than 10 departments and thousands of diseases. The source can guarantee its accuracy, consistency and credibility. Some examples of symptom attributes are shown in Figure 4.

```
@Attribute throat dryness  @ Attribute sore throat  @ Itchy throat  @ Attribute sneezing
@ Attribute nasal obstruction  @ Attribute has a runny nose  @ Attribute hearing loss
@ Attribute lymph node  @ Attribute fever  @ Menstrual disorder  @ Attribute nausea and
vomiting  @ Attribute syncope  @ Attribute is sore all over the body  @ Attribute hemoptysis
```

Figure 4. Some symptom attributes

The expert selects symptom description and quantifies it according to the severity to get the numerical attribute of I-S vector. Finally, the diagnosis result is taken as the nominal attribute to get I-S vector, such as: {1100210031001111001270151O,42, "Laryngitis"}. Some I-S vector sets are shown in Figure 5.

```
@DATA
{0 100, 1 100, 2 100, 3 100, 4 100, 5 100, 6 100 , 7 100, 8 100, 9 10, 10 100, 11 50, 12 100, 13 100, 14 10, " throat dryness"}
{11 100, 12 60, 14 90, 19 80, " bronchitis"}
{3 100, 4 100, 17 100, 23 100, 24 100, 29 10, " throat dryness"}
{9 100, 3 20, 2 70, 3 100, 4 100, 5 100, 6 100 , 7 100, 8 100, 9 10, 10 100, 11 50, 12 100, 13 100, 14 10, " tuberculosis"}
{9 100, 13 100, 20 100, 31 70, 42 50," influenza"}
```

Figure 5. Example of I-S vector set

The massive data objects in the data set lay the foundation for training the classifier.

2.4.2 Comparison with related classification algorithms

In order to select the most suitable classifier for the system, the classical classification algorithm is used to classify the experimental disease data set. This data set contains a variety of common diseases in internal medicine. Experts and researchers extract I-S vector set from real cases. There are 500 data sets for testing.

For each algorithm, 10 fold cross validation is used to divide the original data into 10 sub samples, of which 9 samples are used to train the classifier and 1 sample is used to test. The cross validation is repeated 10 times, and each sub sample is verified once, averaging 10 times. Finally, a single estimated test result is obtained, Using 10 fold cross validation can make the model more general. Using 2nn, 3NN, Bayesian network, naive Bayes, C4.5, SMO algorithm to compare and analyze the above test data sets. The test results as shown in Figure 6.

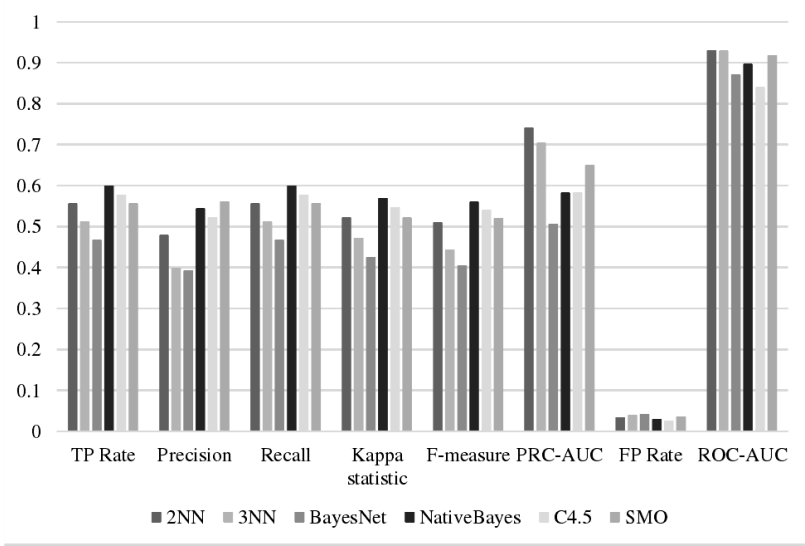

Figure 6. Evaluation histogram of various classification algorithms

Considering the decision tree algorithm, naive or Bayesian algorithm are suitable for this system. Considering the improvement of classification algorithm by combined classification method, and a large number of theory and experiments show that the optimization of decision tree is more convenient to improve the over fitting problem than Bayesian algorithm. Therefore, the system adopts decision tree algorithm as the basic algorithm for improvement.

3. Algorithm improvement

3.1 Improved classification based on combination

Combined classifier is a compound classification model composed of multiple base classifiers. The combined classifier constructs a composite classifier M^* by combining k base classifiers $M_1, M_2, \ldots M_K$ which are trained respectively. The data set D_i is used to create the classifier MI. Given a data tuple to be classified, each base classifier is input, and the final result is returned after voting according to the results of each base classifier. For example, bagging, ascension and random forest all use the idea of combination.

Each classifier of bagging idea is equal weight, and the training set is sampled with put back, so that each training set D_i has a self-help sample, and the final result is majority voting. The idea of promotion is to update the weight of M_i iteratively, so that the next classifier M_{i+1} focuses on the training tuple of M_i misclassification. Compared with bagging, lifting is affected too much by over fitting. Random forest is a complex process that considers the results of multiple base classifiers: each base classifier is generated based on the value of an independent set of random samples. Multiple subsets are sampled from the training set to generate decision trees. In the face of large data, random forest has faster speed and better performance than bagging and lifting, and can solve the over fitting problem well. However, the performance of processing small data and low dimensional data is not necessarily good, and the speed of random forest is far slower than that of single decision tree. In the actual medical system diagnosis with high real-time requirements, the cost of optimization brought by random forest is unacceptable. The actual medical diagnosis expert system must adopt other more rapid and acceptable optimization methods.

3.2 Improved combination classification algorithm based on latent factor

As described in 1.3.3, the influence of potential factors on improving the accuracy of medical diagnosis system can not be ignored. Combined with the idea of combination classification in 3.1, the combination classification

algorithm suitable for inference engine of medical diagnosis expert system can be changed. The system maintains the user latent factor matrix, denoted as U matrix, with space limitation,The selected parts are shown in Table 1.

Table 1. Example of user / potential factor matrix

	Male	female	children	youth	aged	Spring	summer	autumn
A	1	0	0.8	0	0	0	0.1	0.9
B	1	0	0	0	0.7	0	0.1	0.9
C	0	1	0.1	0.9	0	0	0.1	0.9

As shown in Table 1, various factor values of users A、B and C are described. For example, in late August 2019, B was a 60 year old male. The information describing B is (male: 1, female: 0, children: 0, youth: 0, old age: 0, spring: 0, summer: 0.1, autumn: 0.9, winter: 1). With time and user specific behavior matrix will change: changes in user's own information, including changes in age, height, weight and other indicators, geographic migration, major diseases, vaccination and other behaviors that seriously affect the diagnosis will change the user's potential factor vector. There are hundreds of potential factors in the system, which will not be described one by one.

According to 3.1, the association rules obtained by mining a large number of data analysis and combining with expert knowledge, the disease latent factor matrix is also maintained in the knowledge base, which is recorded as I matrix. The space limit is shown in Table 2.

Table 2. Example of disease potential factor matrix

	Male	female	children	youth	aged	Spring	summer	autumn
Urticaria	0.5	0.5	0.6	0.4	0.1	0.7	0.4	0.8
Acne	0.6	0.5	0.1	0.8	0.1	0.5	0.6	0.5
Paederus dermatitis	0.5	0.7	0.2	0.8	0.2	0.3	0.8	0.8

Refer to 3.1, the diagnostic result matrix D can be obtained by $D=UI^T$, which indicates the potential preference of users for different diseases. In fact, because there are many factors, and for a single disease, they are not related to most of the factors, so the matrix is very sparse. In order to avoid the time-consuming decomposition and dimension reduction, the single user searches the U matrix to get the vector R, and the I matrix is filtered by the results of the sub classifiers of the combined classifier to get the sub matrix P of a very small i-matrix, thus simplifying the previous operation of $D=UI^T$ to the operation of a single vector R and matrix P: $\widetilde{D}=RP^T$. Because the matrix P is a submatrix containing the results of each sub classifier, its scale is very small, and it will not get a sparse matrix, so there is no need to decompose this process of low efficiency and high time consumption. Then a diagnosis vector is obtained, and the values of each dimension are sorted in descending order. The diagnosis results sorted in the front have higher confidence.

For example, in the example in this section, for 60 year old male B in late August 2019: (paederus dermatitis: 1.44, papular urticaria: 1.33, acne: 1.18,), in fact, older men are more likely to suffer from paederus dermatitis in early autumn than papular urticaria and acne (commonly known as acne). For the 18-year-old female C in late August 2019: (paederus dermatitis: 2.24, acne: 1.74, papular urticaria: 1.68), in fact, young women are highly likely to suffer from paederus dermatitis and acne in early autumn, which coincides with the expert experience and general cognition. This section shows that it is feasible to use potential factors to optimize the combined classification results.

3.3 The flow of the improved algorithm is described in detail

Firstly, the basic classifier algorithm is described. Base classifier algorithm adopts C4.5 decision tree theory. The basic definitions are as follows:

Let s be the set of data samples, the capacity be set to s, m different types of C, S_i be the number of samples of C_i, and p_i be the probability that any sample belongs to C_i, then the entropy required for a given sample classification is given by formula (2)

$$Entropy\ (s_1, s_2, \ldots, s_m) = -\sum_{i=l}^{m} p_i \log_2 p_i \quad (2)$$

Let the value set $\{a_1,a_2,\cdots,a_v\}$ of non class attribute A, and divide S into v subsets $\{S_1,S_2,\cdots,S_v\}$ according to A, where S_j is the sample whose value is a_j on a. S_{ij} is the sample number of C_i in S_j, then the entropy of A dividing S is given by formula (3)

$$Entropy(A) = -\sum_{j=1}^{v} \frac{s_{1j}+\ldots+s_{mj}}{s}\ Entropy(s_{1j}+\ldots+s_{mj})(i = 1,2,\ldots,m) \quad (3)$$

The definition of information gain ratio is shown in formula (4):

$$GainRatio(A,S) = \frac{Gain(A,S)}{SplitInformation(A,S)} \quad (4)$$

Where *Gain(A,S)* is the information gain, The definition is given by formula (5):

$$Gain(A,S) = Entropy\ (s_1,\ s_2,\ \ldots s_m) - Entropy\ (A) \quad (5)$$

SplitInformation(A,S) is split information, which indicates the uniformity of attribute split data,The pseudo code of the classification process of the trained base classifier is as follows:

$$SplitInformation(A,S) = \sum_{i=1}^{c} \frac{|S_i|}{|S|} log_2 \frac{|S_i|}{|S|} \quad (6)$$

The pseudo code of the classification process of the trained base classifier is as follows:

Input: disease attribute I, trained decision tree root

Process: function baseclassifier(*I,roo*t)

(1) All eigenvalues of *I* are obtained;

(2) Judge the eigenvalue monitored by the root node *root*, select the branch and enter the branch node *current_node*;

(3) If *current_node* is a leaf node, the class label of *current_node* is returned as the classification result; otherwise, *baseClassifier*(*I*,current_node) is iterated;

Output: classification results

The pseudo code of the training process of the base classifier is as follows:

Input: training set $D=\{(x_1,y_2),(x_1,y_2),\ldots,(x_1,y_m)\}$

attribute set $A=\{a_1,a_2,\ldots,a_n\}$.

Procedure: function *treeGenerate*(*D,A*)

(1) Generate node *node*;

(2) If all the samples in *D* belong to the same class *C*, the node is marked as class *C* leaf node;

(3) If a is empty or all samples in d have the same value on *A*, the node is marked as leaf node and its category is marked as the most frequent class in *D*;

(4) Choose the best partition attribute *a** from a: make formula (4) get the maximum value;

(5) Traverse *a**: generate a branch for each *a*_v* node; D_v is the subset of samples in *D* set whose value is *a*_v* on *a**; If D_v is empty, the branch node is marked as leaf node, and its class is marked as the most frequent class in *D*; Otherwise, take *treeGenerate*(D_v,*A*¥{*a**}) as the branch node;

Output: single decision tree

The following describes the combination classifier algorithm of voting result weight based on potential factor optimization.

Its implementation function name *improvedClassifier(A,I,U,N)* is also the external call interface. The interface of the function to be called is as follows: baseClassifier(I,root): the interface of the model trained by

treeGenerate(D, A), whose input is symptom vector I and decision tree root, and returns the classification result of the classifier. In order to achieve the effect of voting, we use different training sets to train a large number of classifiers. For a major disease classification *A*, there are more than *N* combined base classifiers to call. Based on the conclusion of 3.2 and according to the theoretical analysis and design in 3.1, the pseudo code of the improved algorithm is as follows:

Input: disease classification *A* disease attribute I user attribute *U*

Number of combined base classifiers *N*

Procedure: function *improvedClassifier(A,I,U,N)*

(1) If *A*||*D*||*U*||*N* is empty, it is determined as illegal input;

(2) According to the trained classifier root of *A* index, *N baseClassifier(I, root)* are randomly called, each root is different, and the returned results of *N* times of classification are stored as set *C*;

(3) Traverse *C*: search *illness-latentfactor* matrix with each *C* in *C* as index, and save all search results of *C*, that is, submatrix of *illness-latentfactor* matrix, as *P* matrix;

(4) Search the *user-latentfactor* matrix with u as index, and save the result as vector *R*;

(5) Substitution $\widetilde{D}$=RP^{T};

(6) The values of each dimension of the vector are sorted in descending order from large to small;

Output: the disease corresponding to the maximum value is the classification result

---

4. Comparison and analysis of results before and after improvement

A comparative experiment was conducted on a large data set, using 10000 dermatological patient data as the experimental data set, 3000 samples were randomly selected each time, and a base classifier was trained based on *treegenerate (D, A)* in 3.2, a total of 10 base classifiers were trained. Then 500 data records were randomly selected from the data records as the test input, and the C4.5 *basedClassifier (I, root)* and improved classifier (*Dermatology, I, U, 10*) were used for classification diagnosis. The comparison results of the indicators before and after the improvement are shown in Table 3.

It can be found that the improved indicators have significantly improved. Because the data set used in the training is the data recognized by human experts, they are all positive samples. Therefore, in the case of uneven distribution of positive and negative samples, there is the problem of data label tilt. In practical engineering problems, it is more meaningful to use F-measure and recall to evaluate the classifier performance than *ROC-AUC*. The results of the Recall indicator of the 10 basic classifiers and the improved Recall indicator are shown in the figure below.

The experimental results show that the improved algorithm has better performance than decision tree algorithm C4.5 under the condition of large data set: recall rate of the improved classifier reaches 76.5%, which is greatly improved by 16.04 percentage points compared with the average value of 60.46% of 10 base classifiers based on decision tree algorithm; F-measure also increased from 0.556 to 0.703, with a relative increase of 26.4%; PRC-AUC increased from 0.6634 to 0.833, with a relative increase of 25.57%.

It can be seen that the improved algorithm has nearly 76.5% diagnostic accuracy in medical diagnosis, which has reached the standard of practical application, and can be comparable with the ordinary level of human doctors. Due to the small scale of i-matrix and the fact that each diagnosis is aimed at a single user vector R, each$\widetilde{D} = RP^{T}$operation is very easy under the computing power of today's computer. In the future, each base classifier can use multithreading technology to get the results of parallel calculation in the implementation of actual software system, Therefore, the optimization cost of the combined classification

algorithm improved by the potential factor is completely acceptable.

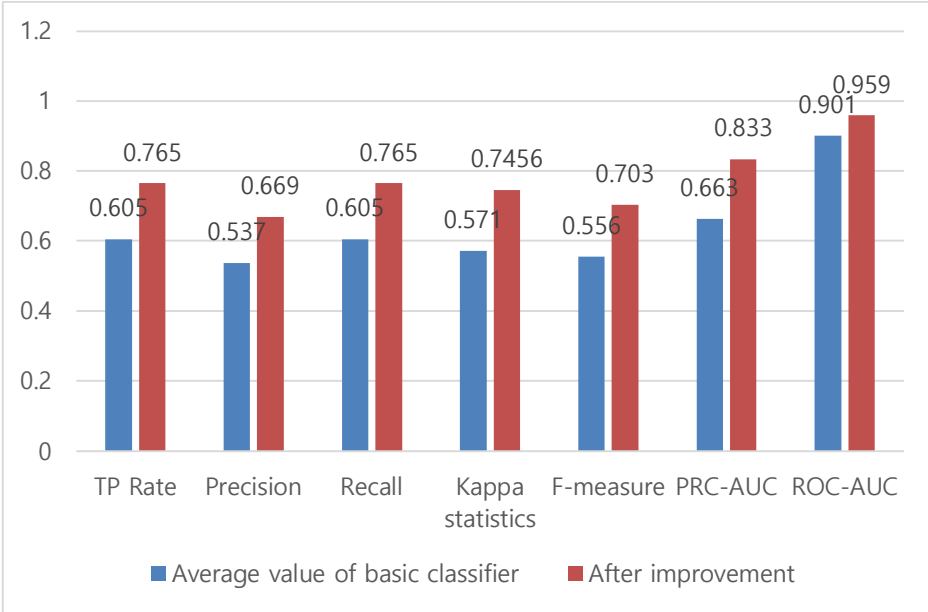

Figure 7. Each index of the improved algorithm

In order to fully illustrate the advantages of the improved algorithm compared with other classification algorithms, this paper takes the recall rate as the index, and carries on the experimental comparison in different scale data sets to verify the effect of the improvement. As shown in Figure 8.

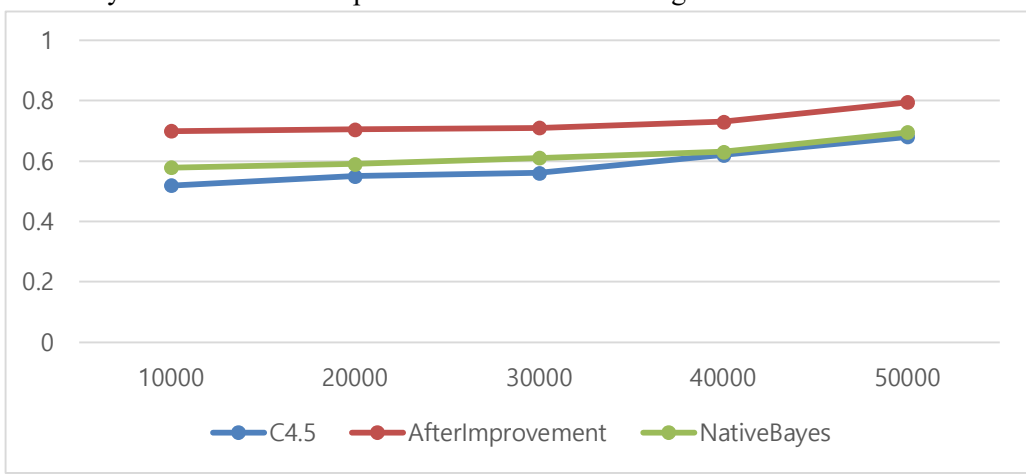

Figure 8. Comparison of recall rate between the improved algorithm and other classification algorithms under different data set sizes

It can be seen from Figure 8 that the improved classification algorithm has a higher recall rate than C4.5 and native Bayes algorithm in both small-scale data sets and big data conditions. When the scale of data set is below 500, it has little advantage, but when the scale of data set reaches 5000, it has obvious advantage. When the scale of data set is beyond 10000, the gap is very obvious. Combined with the above, the improved algorithm is very suitable for the medical intelligent diagnosis and classification business in the era of big data.

5. Conclusion

This paper studies the core intelligent diagnosis function requirements in the medical diagnosis expert system. In 1.1, 1.2 and 2.1, the core module knowledge base and inference engine of the medical diagnosis expert system are explained. The core is knowledge acquisition through association rule mining technology, and the classic classification algorithm and association rule algorithm are analyzed. The shortcomings of the algorithm have been optimized. Afterwards, a mining experiment was carried out on real medical big data. In the third subsection, based on the above results, and drawing on the idea of potential factors in the recommended field, the association analysis is used for the mining of potential factors in medical diagnosis. Then in 3, the core process based on medical diagnosis is to give the most appropriate classification to the central idea of the disease data record after corresponding analysis, and design the classification algorithm for the inference engine part of the medical diagnosis expert system module. An improved combination classification algorithm based on the potential factors obtained by mining association rules is proposed to solve the overfitting problem of a single decision tree. At the same time, the strong influence of individual differences on the diagnostic classification is used to combine the classifiers through potential factors. The results are weighted and optimized, and after the final ranking, the classification result with the highest confidence is obtained. It can be seen that the combined classification algorithm improved by latent factors can meet the needs of medical intelligent diagnosis, realize the diagnosis of common diseases, have great reference significance for real medical diagnosis, and have stable performance and strong robustness.

Acknowledgememt

This paper was supported by the research grant of the Wonkwang University in 2021

References

[1] U. M. Fayyad. Knowledge Discovery in Databases: An Overview[C]. International Workshop on Inductive Logic Programming. Springer-Verlag, 1997, 3-16

[2] Liu Peiqi. Development technology and application of a new generation of expert system[M]. Xi'an: Xidian University Press, 2014, 51-90[41].

[3] Emken B A, Li M, Thatte G, et al. Recognition of physical activities in overweight hispanic youth using KNOWME Net-works[J]. Journal of Physical Activity & Health,2012, 9(3):432-441.

[4] Rodriguez-Villegas E, Chen G, Radcliffe J, et al.A pilot study of a wearable apnoea detection device[J]. BMJ Open,2014,4(10) : e005299.

[5] Gozani S N. Fixed-site high-frequency transcutaneous electrical nerve stimulation for treatment of chronic low back and lower extremity pain[J]. Journal of Pain Research,2016,9(3):469-479.

[6] Russell S J, El-Khatib F H, Sinha M, et al. Outpatient glycemic control with a bionic pancreas in type 1diabetes[J]. New England Journal of Medicine, 2014,371(4):313-325.

[7] Valentina K, Bianco N, Szymkiewicz S, et al. First clinical experience with the wearable cardioverter defibrillator in left ventricular assist device patients[J]. Europace,2016,18(1):128.

[8] Ridler C. Stroke: Wearable robot aids walking after stroke[J]. Nature Reviews Neurology,2017,13(10):576-577.

[9] Paul G, Irvine J. Privacy implications of wearable health devices[C]//International Conference on Security of Information and Networks. Glasgow, Scotland,UK,2014. ACM,2014:117.