ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 153

International Conference on Applied Mathematics, Modeling and Simulation (AMMS 2017)

Research on Assistant Diagnostic Method of TCM Based on Multi Classifier Integration

Yonghong Xie[1,2],Yuyang Yan[1,2], Jianyuan Li[3] and Dezheng Zhang[1,2,*]

[1]School of Computer and Communication Engineering,University of Science and Technology Beijing, Beijing 100083,China

[2]Beijing Key Laboratory of Knowledge Engineering for Materials Science,Beijing, 100083,China

[3]School of Computer and Information, Hohai University, Nanjing 210098,China

[*]Corresponding author

***Abstract*—TCM diagnosis is difficult because of the variety of syndromes and the lack of uniform norms. The traditional Chinese medicine auxiliary diagnosis and treatment system refers to the computer aided system which uses computer modeling technology to assist TCM doctors in recording diseases, prompt diagnosis, assisting prescriptions, and performing some telemedicine and teaching. In this paper, hypertension is taken as an example to study the auxiliary diagnosis of TCM, Based on the classification of symptoms and syndrome elements, a method of TCM assistant diagnosis based on multi classifier ensemble is proposed. This paper studies four classification algorithms: Naive Bayes, Weighted bipartite graph, SVM and ProSVM. To take full advantages of the diversity of different algorithms, a intelligent diagnosis procedure is proposed which could provide technical support for hypertension diagnosis. The effect of the integration was better than that of single classifier, with the average precision increased by 10% to 20%.**

Keywords-TCM diagnosis; weighted bipartite graph; ProSVM; classifier integration

I. INTRODUCTION

With the development of global informationization and internationalization of Chinese medicine, the digitalization and standardization of TCM are becoming more and more thorough, computer-aided diagnosis of Chinese medicine has become the focus of research.TCM diagnosis procedure is very clear, the first need for doctors to conduct a general examination of patients, after to determine the clinical manifestations of patients with disease and disease development of the law; finally come to a concise and effective analysis of the results. The doctor's analysis of the results known as syndromes, the analysis process to become dialectical, record the analysis of the process is called the medical case. We get the general medical cases can not be directly used to test, they need to Chinese segmentation, symptoms and a series of standardized text pretreatment, and then the suitable representation model to represent such data before we used to classify algorithms.

The main research of this paper is the diagnosis of traditional Chinese medicine (TCM). Transfer the diagnosis problem into symptom syndrome classification problem, which is the data mining classification algorithm in question .We first analyzes the classification algorithm based on statistical theory: Naive Bayes, Weighted bipartite graph[2-7], then introduces the improved algorithm of SVM[8][9] and SVM ProSVM[10].Finally, we combine the advantages of the four classification methods, and the integrated model is constructed to achieve the purpose of diagnosis. In this paper, based on the proposed algorithm to design and verify the classification of hypertension symptoms of the classification model for the diagnosis of hypertension to provide technical support.

II. DATA PRETREATMENT

A. The General Process and Frame of TCM Diagnosis

The medical records of TCM diagnosis generally includes preprocessing, feature representation, case processing, model selection and classifier training, the classification results of the evaluation process and its main functional modules are:

1) Medical records pretreatment: Standardize the original medical case, and transformed into a unified format, to facilitate unified follow-up treatment.

2) Feature representation: The medical records of symptoms and syndromes is decomposed into basic processing units, represented by the mathematical model, the model is mainly used in Boolean model and vector space model.

3) Feature dimensionality reduction: After finishing the data are sparse matrix, useless data a lot, will seriously affect the efficiency of the subsequent data processing, so the data processing, this operation will greatly improve the efficiency of classification.

4) Training and classification: Select the classification algorithm, train different learning models, and classify the text of the test. This is the core of traditional Chinese medicine diagnosis, the effect of classification directly determines the result of diagnosis.

5) Impact assessment: The classification results are analyzed, the effect of classification is evaluated, the appropriate classification and evaluation parameters are selected, and the advantages and disadvantages of each classification algorithm are analyzed in depth to find a better diagnostic method.

B. Symptom Normalization

First of all, we need to standardize the patient's symptoms and standardize the existing symptoms in the body so that the system can be processed. The specific specification process is shown in FIGURE Ⅰ.

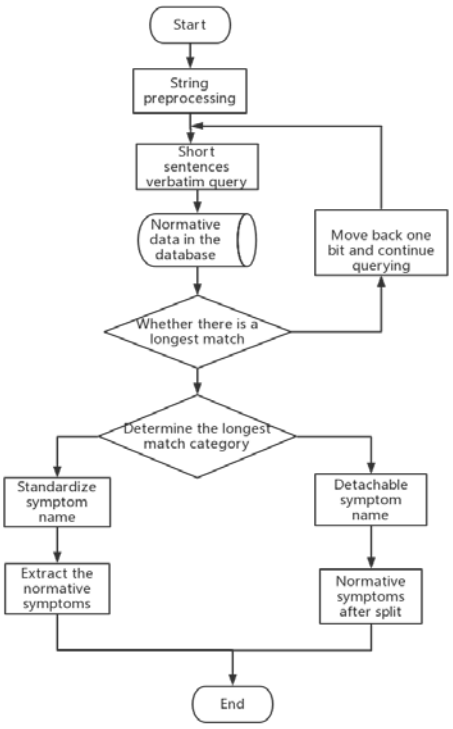

FIGURE I. SYMPTOM NORMALIZATION PROCESS

C. Medical Data Pretreatment Process

Medical case data preprocessing process mainly has a few processes:

1) First of all, we need to standardize the terminology related to Chinese medicine. These need to be done with the help of experts from the Chinese medicine. We can only extract the symptoms and syndromes in the case to study.

2) The next step is to standardize the symptoms, decompose the combination symptoms, and merge the similar symptoms to form the basic unit of the symptomatic terms.

3) The same name of syndrome elements were standardized.

4) Standardize the symptoms of each case and store it in the database.

5) The symptoms set split into individual symptoms, syndrome elements set split into individual syndrome elements.

Data preprocessing, medical records extracted from the medical records database. Normalize the symptoms, preprocess the data, and use the vector model to represent, this method to facilitate the subsequent classification of data research. The number of occurrences of each syndrome and the number of occurrences of the combination are counted by the processed data for comparison with the classification results to analyze the effect of the classification.

III. RESEARCH ON INTEGRATED DIAGNOSIS BASED ON MULTI CLASSIFICATION

A. Construction of a Single Classifier for Integration

1) Naive bayes

Using the Matlab2014a program, 80% of the data was selected as the training set for the naive Bayesian algorithm, and the remaining 20% was used as the test set to test the learning effect of Naive Bayes. The specific classification process is as follows:

a) Enter the medical record data, one by one scan each medical case d.

b) Read all the experimental data into the system, and then select a part of the case as a training set Train, the remaining cases as a test set Test, and then get training set symptoms Train_in, training set tags Train_tar and test set symptoms Test_in and test set tags Test_tar.

c) Classify the symptom set and the corresponding set of labels into the naive Bayesian classifier model.

d) Take the symptoms of the test set Test into the trained naive Bias classifier model, and then give a predictive value for each label that corresponds to the symptoms of the test set.

e) Choose a suitable threshold, so that the case to predict the correct number of the most, for each case can be found m the most likely evidence.

2) Weighted bipartite graph

A bipartite graph consisting of two different categories of nodes and the edges between two types of nodes. In the diagnosis of traditional Chinese medicine the main analysis is the relationship between symptoms and syndrome elements, Each symptom corresponds to a node that constitutes a symptom set Z. Each syndrome corresponds to a node that forms the set of syndromes S. If a symptom z_i is a factor that leads to a certain prime element s_j , an edge is linked between the symptoms and the syndrome, indicating that the two are interrelated so that we can construct a symptom - syndrome bipartite graph. Defining symptom sets as $Z=\{z_1,z_2,\cdots z_n\}$, Syndrome sets as $S=\{s_1,s_2,\cdots s_m\}$ Therefore, the network structure can be represented by n+m nodes, among which $1\le i\le n,1\le j\le m$,The specific example is shown in FIGURE Ⅱ.

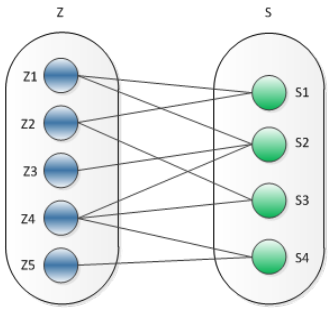

FIGURE II. BIPARTITE GRAPH RELATION MODEL

For this basic bipartite graph, if we add weight to the bar, it becomes a weighted dichotomy. Or to the above query and the results of the dichotomous map, for example, if we give the edge $\left(z_i, s_j\right)$ between a symptom z_i and its Syndrome s_j, plus the weight of the symptom relative to the prime, then the given bipartite graph is a weighted Partial chart.

Similarly, the relationship between syndrome and symptoms can be established. Based on these two diagrams, the relationship between symptoms and syndrome can be given. In this way, by means of a bipartite graph, which can analyze which symptom the patient may have.

3) SVM

The simplest and first proposed model of the support vector machine is the largest interval classifier, which is also the main module of SVM, which shows the key features of this kind of learning machine. When using support vector machines (SVM) for symptom syndrome classification, the specific classification process is as follows.

a) Enter the medical record data, one by one scan each medical case d.

b) Read all the experimental data into the system, and then select a part of the case as a training set Train, the remaining cases as a test set Test, and then get training set symptoms Train_in, training set tags Train_tar and test set symptoms Test_in and test set tags Test_tar.

c) Training set of symptom sets and corresponding tag sets are entered in the SVM model for learning.

d) The symptoms of the training set are brought into the SVM model trained above, and then the label corresponding to the symptom of the training set is obtained, and each tag has a predicted value.

e) Training Set Predicted Tags Train_pre Control Train Set Train_tar, select an optimal threshold for each case, and finally form a set of thresholds.

f) Enter the threshold set and the training set of symptoms into a multivariate linear regression model to learn a linear regression function line.

g) Select the test set Test_in, the test set of predictive tag set Test_pre into the function, you can get the training set of each case of the threshold, so that each of the cases can be sorted out with the most probable m.

4) PRO-SVM

In practice, an object can be associated with multiple tags at the same time. For a label of multiple tasks, an object is usually associated with a subset of the tags; we call these tags as related, while the rest is called irrelevant. The basic goal of multi-tag learning is usually to predict the label, that is, to predict which tags are relevant and which are irrelevant. However, in many applications, in addition to tag prediction, there is often another requirement that a good ranking of the relevant tags is predicted. Many of the optimization algorithms are designed to optimize the performance of some classifiers, such as BR is designed for Hamming loss; RankSVM is designed for Ranking losses; AdaBoost.MH and Adaboost.MR (BoosTexter's two implementations) Optimize HammingLoss and RankingLoss. This paper uses Zhou Zhihua proposed a new classification method. He proposed a new performance evaluation index ProLoss (Prediction and Relevance Order Loss) prediction and correlation ranking loss. Then the ProSVM classification method is proposed, which uses the method of alternating multiplier to effectively optimize the PRO loss. To further improve efficiency, we introduce an upper approximation, reducing the number of constraints from O (T2) to O (T), where T is the number of labels. Experiments show that their recommendations are better for multi-tag classification.

B. Classification Method Performance Analysis

The evaluation index of the traditional single-label classification problem includes accuracy rate, precision rate, recall rate and F1 value, but these are not suitable for the diagnosis of TCM diagnosis, because each medical case has a variety of evidence, Multi-tag problem, multi-label learning problems in the evaluation of the evaluation than the single-label learning a lot of complex. In order to evaluate the performance of multi-tag classification algorithm, the following five evaluation indexes are selected.

- Hamming Loss: This indicator evaluates the match error rate between the predicted marker set of each sample in the set of samples to be sorted and the actual set of markers.
- Error Rate (One-Error): This indicator evaluates the probability that the first marker in the actual tag set does not appear in the set of predictive markers.
- Average Precision: This indicator evaluates the average accuracy of the forecast marker.
- Coverage This indicator evaluates the average number of markers that need to be covered by the corresponding number of markers.
- Ranking Loss This indicator evaluates the extent to which the predicted marker set for each sample in the set of categories to be sorted is not the same as the actual marker set.

A good multi-tag learning algorithm should have a lower Hamming loss, a 1-error rate, and a relatively high average accuracy. Comparing the above-mentioned weighted dichotomous maps, naive Bayesian and ProSVM, which are

suitable for TCM diagnosis in this experiment. Specific data as shown in TABLE Ⅰ.

TABLE I. CLASSIFICATION METHOD PERFORMANCE COMPARISON

Classifier	Hamming Loss	RankingLoss	OneError	Coverage	Average_Precision
NBM	0.8278	0.5732	0.4286	21.5714	0.4832
WBG	0.8102	0.3940	0.5714	48.8571	0.5719
SVM	0.4236	0.4976	0.5906	30.6584	0.5883
ProSVM	0.3094	0.3875	0.6906	28.2857	0.6073

Naive Bayesian method Coverage value is relatively small, indicating Naive Bayesian method in predicting marker sample sort search all belong to the category of the sample coverage depth is relatively small, compared to Weighted bipartite graph.The One-Error index of naive Bayes is smaller, which indicates that the probability of the first tag in the actual tag set is not small in the prediction mark set. The accuracy of the Weighted bipartite graph is in the middle of ProSVM and Naive Bayesian. ProSVM in HammingLoss, RankingLoss these two indicators of the value of the indicators are relatively small, indicating that the corresponding relationship between symptoms and evidence, misplaced less, and the average error rate of marked sorting is also significantly reduced. Average_Precision accuracy improved a lot, the classification effect is better. These methods have their own advantages, we need to further analyze the relationship between these data.

Because it is not the average distribution of the label for TCM diagnosis and in this experiment, there are obvious class imbalance phenomenon, as shown in FIGUREⅢ, more like the number of blood stasis and phlegm appears, and the number of Yang deficiency and blood deficiency occur rarely. The accuracy of each method for predicting different syndrome factors may be different. Therefore, this paper calculates the accuracy of each classification method for each type of target prediction, as shown in TABLE Ⅱ.

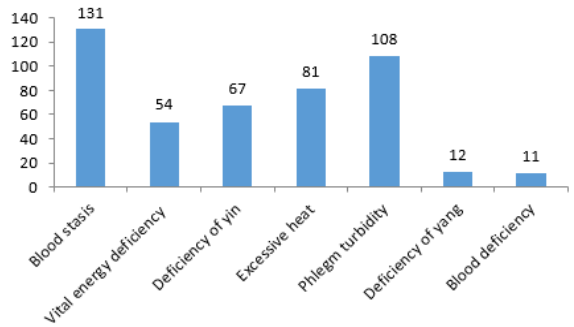

FIGURE III. SYNDROME ELEMENT DISTRIBUTION DIAGRAM

From TABLE Ⅱ, we can see that the Weighted bipartite graph. has better prediction effect on the syndrome of blood deficiency, phlegm and heat, and ProSVM has better prediction effect on the lower probability of symptom. Therefore, in order to achieve a better diagnostic effect, this paper finally uses the integrated method, these kinds of prediction methods as a base classifier to form an integrated classifier, integrated classifier combined with the classification of these classifiers to do the final results Forecast, the forecast will be better.

TABLE II. CLASSIFICATION METHOD PERFORMANCE COMPARISON

	Blood stasis	Phlegm turbidity	Excessive heat	Deficiency of yin	Vital energy deficiency	Deficiency of yang	Blood deficiency
WBG	0.7217	0.7999	0.7725	0.8284	0.8128	0.8351	0.6004
ProSVM	0.6646	0.7411	0.7729	0.8144	0.8126	0.9513	0.9567
NBM	0.7124	0.7843	0.7515	0.7973	0.7450	0.9416	0.8915
SVM	0.7145	0.7482	0.7763	0.8069	0.7763	0.9457	0.9362

C. Integrated Diagnostic Model

This paper chooses ProSVM, Weighted bipartite graph, naive Bayesian and SVM as four classifiers as base classifiers, and they have their own advantages. Integrated together to draw the advantages of each classifier. The main process is to randomly extract samples, train each base classifier with the same number of samples, train different classification models, and then integrate the predicted results of these classifiers. The results of the integrated classifier are better than the single classifier. The TCM diagnosis of hypertension is a multi label classification problem, and considering the previous experimental results, the final diagnosis result depends on the four base classifiers on each marker prediction results and the weights of the product is more suitable for the present situation of the integrated learning method.

D. Integration Strategy

There are many kinds of integrated strategies, In order to give full play to the advantages of each classifier, this paper uses the weighted integration strategy. To introduce the integration process, given the following definition. p_{ij} represents the accuracy of each syndrome classification, $i=\{1,2,3,4\}$, respectively represent the four base classifiers of ProSVM, SVM, NBM, weighted dichotomy $j=\{1,2,3,4,5,6,7\}$,they represent 7 syndromes: blood stasis, phlegm turbidity, heat excess, yin deficiency, Qi deficiency, Yang deficiency and blood deficiency. w_{ij} represents the weight of a sign; I_{ij} indicates the prediction result of each element, r_j represents the integration result of each card element, as shown in the following procedure.

1) Select a base classifier as a reference classifier to assign weights to other base classifiers. According to the

accuracy of each classifier, select Naive Bayes as a reference,that is $w_{3j}=1$.

2) The ratio of the accuracy of the other three classifiers to each of the NBM classes is taken as their respective weights,that is $w_{ij}=\frac{p_{ij}}{p_{3j}}$, The specific weight distribution is shown in TABLE Ⅲ.

3) Firstly, the four base classifiers of ProSVM, weighted dichotomy, naive Bayesian and SVM are used to predict the samples, and a prediction result is obtained. Then, the results of each classifier are calculated for each class.

TABLE III. THE WEIGHT OF EACH BASE CLASSIFIER FOR EACH CLASS

	Blood stasis	Phlegm turbidity	Excessive heat	Deficiency of yin	Vital energy deficiency	Deficiency of yang	Blood deficiency
ProSVM	0.933	0.945	1.028	1.021	1.091	1.004	1,038
SVM	1.003	0.954	1.033	1.012	1.042	.0998	1.016
NBM	1	1	1	1	1	1	1
WBG	1.013	1.020	1.066	1.039	0.880	0.881	0.652

From the TABLE Ⅲ can clearly see the advantages of each base classifier, SVM and ProSVM on the occurrence of fewer blood deficiency, Yang and other classification effect is better, so the weight will be given too large, and the number of occurrences Of the blood stasis, phlegm and other classification effect is weaker, given the weight will be too small. On the contrary, the Weighted bipartite graph the weighted dichotomy of the number of occur more blood stasis, phlegm classification effect is good, given the weight is too large, and the weaker fewer times of Yang deficiency and blood deficiency of the classification results, thus given small weight.

After the integration of the main consideration AveragePrecision this evaluation index, as shown in Table Ⅳ, the last line that is integrated learning after the results. The results of the classification after integration will be better than the individual classifiers.

TABLE IV. CLASSIFIER INTEGRATION RESULTS

Classifier	AveragePrecision（%）
NBM	48.32
WBG	57.19
ProSVM	60.83
SVM	58.863
Integrated classifier	68.20

IV. CONCLUSION

The main research of this paper is the diagnosis of traditional Chinese medicine (TCM). The data of hypertension in traditional Chinese medicine are analyzed. According to the current research status of traditional Chinese medicine, we can judge according to the "syndrome" of the symptoms, according to the syndrome elements, we can get the combination of patient's symptoms, the diagnosis problem into symptom syndrome classification problem, which is the data mining classification algorithm in question.

There are many kinds of classification algorithms based on data mining. In this paper, we first analyze the Naive Bayesian and the Weighted bipartite graph, and then introduce SVM and SVM-based improved algorithm ProSVM. Since this study is the symptom of TCM diagnosis, The problem of classification of syndromes can be classified as a set of indefinite number, which belongs to multi-tag classification problem. The ProSVM is designed specifically for the classification of the design method, so the test results will be better, more suitable for the diagnosis of Chinese medicine decision-making research.

At the end of this paper, an integrated classifier model based on Naive Bayesian, Weighted bipartite graph and ProSVM and SVM is proposed as the classification model of diagnosis, which provides technical support for hypertension diagnosis.

ACKNOWLEDGMENT

This work was partially supported by the National Key Research and Development Program of China under Grant 2017YFB1002304, and we would like to express our gratitude to Yao Jingjing for she help us conduct experiments.

REFERENCES

[1] Chen Shuhui. A Comparative Study on the Application of Classification in TCM Syndrome Differentiation Diagnosis (Doctoral Dissertation) [D].Guangzhou: Guangzhou University of Traditional Chinese Medicine, 2008.

[2] Cherman E A, Spolaôr N, Valverde-Rebaza J, et al. Lazy Multi-label Learning Algorithms Based on Mutuality Strategies[J]. Journal of Intelligent & Robotic Systems, 2014:1-16.

[3] Zhang M L, Zhou Z H. ML-KNN: A lazy learning approach to multi-label learning[J]. Pattern Recognition, 2007, 40(7):2038-2048.

[4] Li Changjun. Data Mining Based on Bayesian Network in Traditional Chinese Medicine (Master's Thesis) [D]. Xiamen University, 2008.

[5] Du Ruijie, Bias. Classifier and its application. (Ph. D. Thesis) [D]., Wang Hanxing. Guidance. Shanghai: Shanghai Univer, 2012.

[6] Zhu Lang. Query recommendation algorithm based on two diagram (Master Thesis) [D]. Zheng Cheng. Guidance: Anhui: Anhui University, 2014.

[7] Zheng Siyuan. Research and implementation of a hybrid recommendation system based on two diagram (Master Thesis) [D]. Beijing University of Posts and Telecommunications, 2015.

[8] Kim H C, Pang S, Je H M, et al. Constructing support vector machine ensemble[J]. Pattern Recognition, 2003, 36(12):2757-2767.

[9] Liu H X, Zhang R S, Luan F, et al. Diagnosing Breast Cancer Based on Support Vector Machines[J]. Journal of Chemical Information & Computer Sciences, 2003, 43(3):900-907.

[10] Miao Xu, YuFeng Li, ZhiHua Zhou. Multi-Label Learning with PRO Loss[C]. // Proceedings of AAAI Conference on Artificial Intelligence, 2013.

[11] Han J, Kamber M. Data mining concept and technology[J]. Th Annual International Symposium on Supply Chain Management, 2001.