The Automatic Classification Research of Regional Medical Imaging Disease Based on Support Vector Machine[1]

He BiShi[1, a], NI HangJian[2,b] and Xu Zhe[3,c*]

[1] Hangzhou Dianzi University, China, 310018

[2] Hangzhou Dianzi University, China, 310018

[3] Hangzhou Dianzi University, China, 310018

[a] hebs@hdu.edu.cn, [b]nihangjian0518@163.com, [c]xuzhe@hdu.edu.cn

Keywords: support vector machine; disease classification; k-means algorithm

Abstract:This paper proposed an optimized support vector machine method to realize the automatic diseases classification of medical imaging results, which overcome the traditional Bayesian Classification problems. It improved the classification accuracy effectively, and was used to analyze the medical characteristics of the two hospitals in the same region.

INTRODUCTION

With medical data explosive growth, research on clustering and classification of medical data is significant to knowledge of medical information and management decision-makings. Currently, disease classification of medical imaging results usually classified artificially by imaging diagnosis doctors. However the doctors usually are too busy to do this work. Therefore, it has the application value to realize automatic classification. Through the literature retrieval, a method is found in the paper "The Automatic Classification Research to Medical Imaging Disease Based on Naïve Nayesion Algorithm" written Huo Hongbo, etc[1]. But the Naive Bayesian method has its shortcomings: (1)Each categorical attributes are assumed independent, but in fact this assumption is often not set up, which certainly impact the classification accuracy; (2)The assumption is to classification with infinite training samples, however the training sample is limited in the actual situation.

For those above problems, we applied Optimized SVM method to solve the problems of the small sample training, high-dimensional, nonlinear and low classification accuracy etc. And try to apply this method into practice.

K-MEANS CLUSTERING ANALYSIS

In the situation of there are no classified samples, the text clustering must be completed firstly. Paper selected the most classic K-means clustering algorithm[2], and seventy thousand inspection records from RIS database are clustered into ten classes based on international disease standard code ICD-10. *A. K-means Text Clustering*

First the extracted data was pretreated, and the Chinese word segmentation tool ICTCLAS are used to segement, then the data is converted into feature model. After the text converted into VSM, K-means clustering will be did. K-means text clustering can be described as[3]: (1)In a given text

[1]This paper is supported by Zhejiang Key Enterprise Institute Program and Zhejiang Smart City Regional Collaborative Innovation Center Project

set，d is VSM space model; (2) the number of generated clusters is determined, k=10; (3) K initial cluster centers is generated order by the principles of randomly generated; (4) the similarity of the initial cluster centers for the each of the D text is calculated, the similarity can be expressed as:

$$sim(d_i, s_j) = \frac{\sum_{k=1}^{n} w_{1k} * w_{2k}}{\sqrt{\sum_{k=1}^{n} w_{1k}^2 \sum_{k=1}^{n} w_{2k}^2}} \quad (1)$$

(5) The biggest similarity from all of cosine similarity is picked, then classified text d into class s; (6) The fourth or fifth steps are repeated until all texts have been completedclustering.

Results of K-means text clustering

After completing clustering in k-means clustering algorithm, the clustering results need to be tested. Index of clustering results commonly used precision, recall and F value, however the test used by these indicators must be based on the situation of divided class. Because of the huge number of clusters, it's difficult to judge the clustering result through the divided class by artificial classification. So the clustering results are used in next research directly.

ESTABLISHMENT AND OPTIMIZATION OF SUPPORT VECTOR MACHINE

After the k-means clustering are completed, then 300 samples per class are selected consisting 3000 training set from clustered text.

Automatic classification process includes text preprocessing, feature selection, weight calculation, text representation and text classifier training. The flow chart of automatic diseaseclassification is shown in Fig 1.

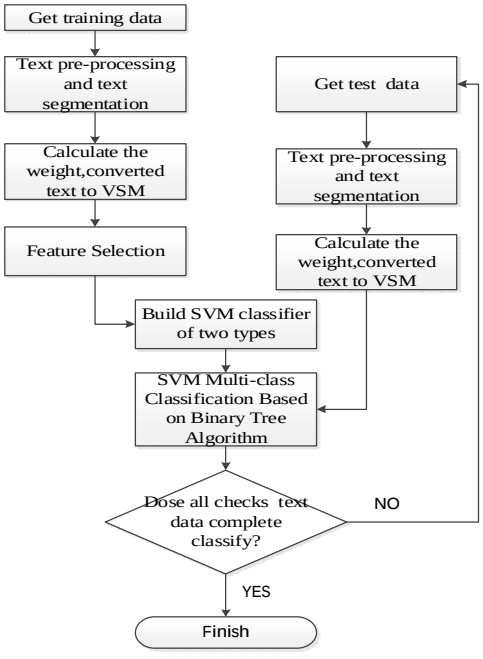

Fig.1 Text classification flowchart

Text Pre-processing

Text pre-processing involves the following modules. (1)Text segmentation: use ICTCLAS Chinese segmentation system after text pre-processing. (2)The text is converted into VSM after completing segmentation. The thought of the VSM is: *D* initial documents and a group of *n* words are assumed, we can regard each document as independent term group. And the feature is given a certain weight, as the corresponding coordinate values. VSM can be expressed as: TF-IDF is used to

calculate weighted forms, we can obtain the following formula:

$$W_{ik} = \frac{tf_{ik} \times \log(N / n_k + 0.01)}{\sqrt{\sum_{k=1}^{p} (tf_{ik})^2 \times \log^2(N / n_k + 0.01)}} \quad (2)$$

Where is the times of the words appeared in document, N is the number of all training documents, n_k is the number of the training documents in which the word appears.

(3)Feature selection to reduce the dimensionality. Information Gain (IG) is selected as feature selection method. IG reflect the text information by calculating the average of a feature in the text[4]. The main formula is as below:

$$IG(t) = P(t)P(C_i \mid t)\log\frac{P(C_i \mid t)}{\log P(C_i)} + P(\bar{t})P(C \mid \bar{t})\log\frac{P(C_i \mid \bar{t})}{\log P(C_i)} \quad (3)$$

Where $P(C_i)$ represents the probability of C_i type text appears in the overall text; $P(t)$ represents the probability that t feature items in total text; $P(C_i|t)$ represents the probability that the text contains t feature items belonging to the class of C_i text; $P(C_i|\bar{t})$ represents the probability that the text not contains t feature items but belonging to the class of C_j text.

SVM Text Classification

Support Vector Machine (SVM) algorithm is developed based on statistical learning theory[5]. The key of construct SVM classifier is to find a hyperplane. Certain types classified linear equations of medical imaging text is assumed $g(x) = w \cdot x + b$, it was normalized so as to satisfy the condition, we can obtain the classification interval is $|g(x) \geq 1|$ when the condition is equal to 1. To make the largest class interval, it's equivalent to find the minimum value of $\|w\|$, and it was transformed into solve the minimum of $\varphi(w) = \frac{1}{2}\|w\|^2 = \frac{1}{2}(w \cdot w)$. Lagrange function is introduced to solve the constrained optimization problem, so we can obtain the following formula:

$$\min L(w, b, a_i) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} a_i[y_i(w\ x_i + b) - 1] \quad (4)$$

Then we can obtain the optimal solution of w^* and b^* by Lagrange minimum theorem: $w^* = \sum_{i=1}^{n} a_i^* y_i x_i$; $b^* = \frac{1}{|s|}\sum_{i \in s}(y_i - w^* x_i)$ $(y_i \in \{+1, -1\})$. Finally, we can get the optimal classification function of two categories:

$$f(x) = \mathrm{sgn}(w^* \cdot x + b^*) = \mathrm{sgn}(\sum_{i=1}^{n} \alpha_i^* y_i x_i \cdot x + b^*) \quad (5)$$

If f(x) is greater than zero indicates that the text belongs category, and less than 0 indicates not belongcategory.

We need use kernel function to convert medical imaging text into a linear when it is nonlinear. In this paper, we use linear kernel function, and use it to replace the x value. Finally, we can get the optimal classification function as below:

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{n} a_i^* y_i k(x_i, x_j) + b^*) \qquad (6)$$

SVM Multi-class Classification Based on Binary Tree Algorithm

Medical imaging text classification is a multi-classification problem because of it has ten categories, but SVM can only solve two classification problems, so it must be achieved by building multiple classifiers. In this paper, we use binary tree algorithm to build multi-classifier[6].

The process of establishing medical imaging text multi-class SVM classifier can use the following example to illustrate: (1)The first, training data of respiratory disease category as +1 training data, while the remaining training data of nine categories as -1 training data. So SVM classification decision function of respiratory disease was built. (2)Training data of circulatory system diseases as +1 training data, while the remaining training data of eight categories except respiratory disease category as -1 training data. SVM classification decision function of circulatory system disease was built. (3)At last, classification functions of several other categories are built according to the above-mentioned two-steps.

After that, we can set test sample data X for example. The X begins respiratory diseases classification, if it belongs to this class, then we stop classifying, else we use the second, third and the last SVM classifier to finish classifying.

Performance Test of Multi-class SVM Classifier

The test sample need to performance test after build the multi-class SVM classifier. The classification results shown in Table 1.

	Recall(%)	Precision(%)	F1 (%)
Respiratory system disease	93.799	97.005	95.375
Pregnancy, childbirth and the puerperium diseases	75.706	91.881	83.013
Circulatory system disease	85.323	53.908	66.071
Genitourinary system disease	94.034	91.828	92.918
Alimentary system disease	99.904	88.115	93.639
Skin and subcutaneous tissue disease	88.134	93.066	90.533
Nervous system disease	94.363	98.421	96.35
Ear and mastoid disease	77.061	67.895	72.188
The musculoskeletal system and connective tissue diseases	90.772	97.103	93.831
Tumor	84.089	81.127	82.581
Macro mean	88.318	86.035	86.649

Table1 Classification results

Improving SVM Classifier

From Table 1, the average precision of text classification can reach 86%, the classification precision than naive Bayes classifier by Huo Hongbo has improved a lot. But we can also view the precision and recall of circulatory system diseases,ear and mastoid disease are poor. So we can

obtain classification confusion matrix based on the classification results, the results are shown in Table 2.(Disease name is called for short in table2)

	Respiratory	Pregnancy	Circulatory	Genitourinary	Alimentary	Skin	Nervous	Ear	Musculoskeletal	Tumor
Respiratory	11239	20	24	66	26	0	7	52	64	484
Pregnancy	1	10162	2894	219	2	9	0	62	57	17
Circulatory	1	528	4325	5	0	0	20	0	21	169
Genitourinary	39	184	108	8417	6	34	21	34	22	86
Alimentary	1	0	0	0	1038	0	0	0	0	0
Skin	30	9	1	51	0	765	3	2	6	1
Nervous	156	0	12	5	0	0	4302	2	51	31
Ear	28	19	96	17	4	0	0	645	22	6
Musculoskeletal	90	136	6	349	26	14	17	143	8145	47
Tumor	1	2	557	37	76	0	1	10	0	3615

Table2 Confusion matrix table of Classification

In table, the row represents the number of the category for original was assigned to other categories, and vertical column represents the medical imaging text which not belongs to this category. We also find the misclassified texts of pregnancy category, circulatory system and tumor are so much. We found that the main reason of misclassification is SVM classifier training sample is too narrow, it don’t contain a representative medical imaging text.

So we need to optimize the SVM. Since the training data contains more beneficial text for SVM classifier, so we can constitute new training sample includes the misclassified text and the original training samples. The new composition of training sample has a stronger ability to learn, and last the SVM classifier was trained again. The optimization can be described by the following two graphs. Figure 1, the black sample represents the misclassified data by SVM, we can see that the classification is not the best optimized hyperplane. Figure 2 represents the result after optimization, the optimized SVM can correctly classified the misclassified text. It effective solution the situation because of the training sample is too narrow to classify.

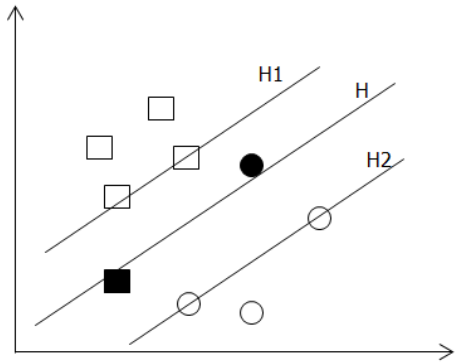

Fig.2 Misclassified sample

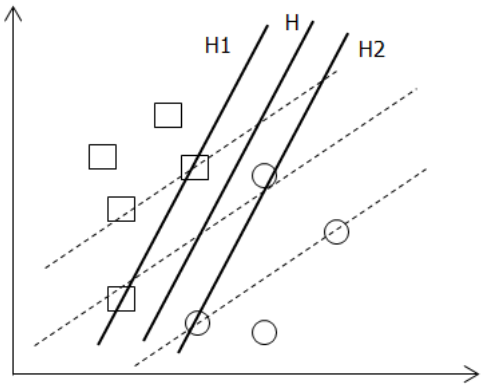

Fig.3 Optimized SVM

Performance Test of Optimized SVM Classifier

Training in accordance with the above optimization mode, and then the training samples are classified again. Then we can take the results before optimized compared with the results after optimized, its comparative results shown in Table 3.

	Before optimization（%）			After optimization（%）		
	Recall	Precision	F1	Recall	Precision	F1
Respiratory system disease	93.799	97.005	95.375	99.169	99.414	99.291
Pregnancy, childbirth and the puerperium diseases	75.706	91.881	83.013	85.682	94.903	90.057
Circulatory system disease	85.323	53.908	66.071	93.68	88.738	91.142
Genitourinary system disease	94.034	91.828	92.918	92.794	96.595	94.656
Alimentary system disease	99.904	88.115	93.639	97.845	85.019	90.982
Skin and subcutaneous tissue disease	88.134	93.066	90.533	92.946	95.522	94.216
Nervous system disease	94.363	98.421	96.35	98.136	96.704	97.415
Ear and mastoid disease	77.061	67.895	72.188	97.137	87.585	92.114
The musculoskeletal system and connective tissue diseases	90.772	97.103	93.831	97.891	96.426	97.153
Tumor	84.089	81.127	82.581	89.453	85.757	87.566
Macro mean	88.318	86.035	86.649	94.473	92.666	93.459

Table3 Performance comparison table of SVM before and after optimization

The table shows classification results of optimized SVM classifier significantly increased, the precision of circulatory system from 53.908% increased to 88.738%, and precision of Ear and mastoid class from 67.895% increased to 87.585%. And the macro precision also increased by about nearly 7%, it effective proved practicability of the optimization method.

THE APPLICATION OF SVM CLASSIFIER IN REGIONAL MEDICAL IMAGINE TEXT

After the SVM text classifier optimized training was completed, the optimized SVM classifier is applied into the automatic classification of Regional Medical imaging text. Then we deal with classified research for A, B two medical institutions in the same region. It facilitates analysis of disease trends of different medical institutions in regional, and it also facilitates relevant departments to expand preventive measures and arrange staff. And we can judge the specialties in the field of different medical institutions according to the results.

We extract check text from September 2013 to September 2014, and then classify these texts. We can obtain following two disease trends after counting data. From figure we can find that disease trends of A and B agency generally maintain stable trend. But it can obviously be found in the region with a high incidence of disease, for example urogenital diseases and respiratory diseases. At the same time, we can also find that the featured areas of A, B agency is different. The featured area of A agency is pregnancy category and genitourinary diseases; and the featured area of B agency is respiratory, genitourinary and bone diseases.

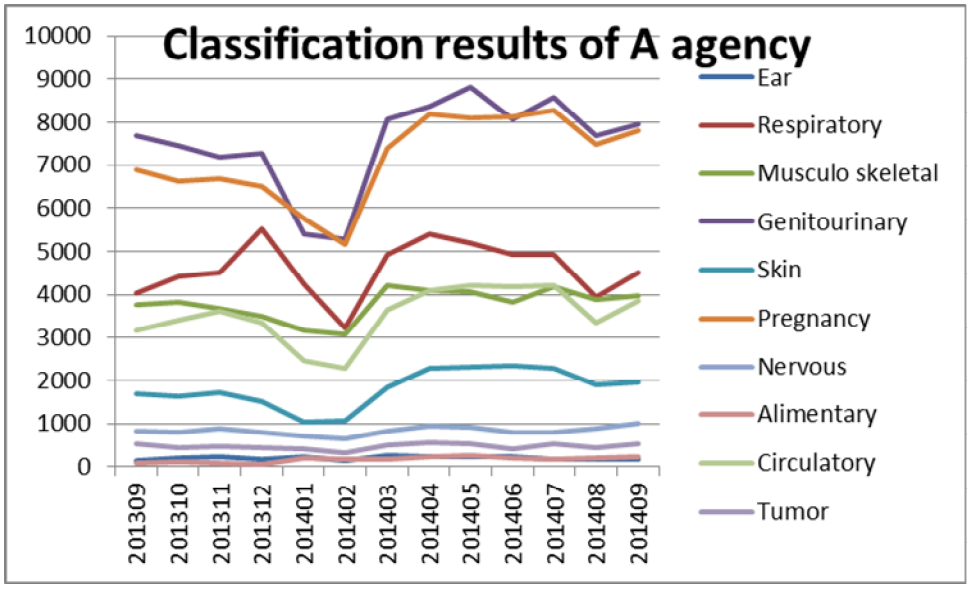

Fig.4 Disease classification map of A agency

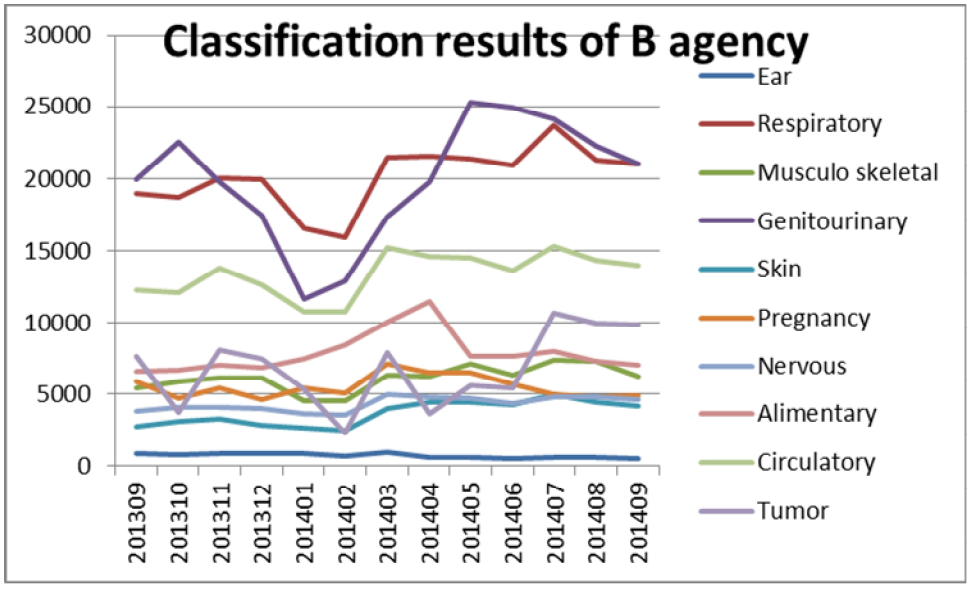

Fig.5 Disease classification map of B agency

CONCLUSION

In this paper, we apply SVM classifier to classify the medical imaging text, and optimize the training method to improve the classification accuracy. Then we analyze the medical characterstics of two hospitals in the same region.

REFERENCE

[1] Luoyong N, Jiaming H, Hongbo H, et al. The Automatic Classification Research to Medical Imaging Disease Based on Naïve NayesionAlgorithm[C]//Computational Intelligence and Security (CIS), 2014 Tenth International Conference on. IEEE, 2014: 308-311.

[2] Ahmad A, Dey L. A k-mean clustering algorithm for mixed numeric and categorical data[J]. Data & Knowledge Engineering, 2007, 63(2): 503-527.

[3] Xinwu L. Research on text clustering algorithm based on improved K-means[C]//Computer Design and Applications (ICCDA), 2010 International Conference on. IEEE, 2010, 4: V4-573-V4-576.

[4] Zhang H, Ren Y, Yang X. Research on Text Feature Selection Algorithm Based on Information Gain and Feature Relation Tree[C]//Web Information System and Application Conference (WISA), 2013 10th. IEEE, 2013: 446-449.

[5] Zhang M, Zhang D. Trained SVMs based rules extraction method for text classification[C]//IT in Medicine and Education, 2008. ITME 2008. IEEE International Symposium on. IEEE, 2008: 16-19.

[6] Weifa Z. A SVM Text Classification Approch Based on Binary Tree[C]//Computer Science-Technology and Applications, 2009. IFCSTA'09. International Forum on. IEEE, 2009, 3: 455-458.