The Improved Text Classification Method Based on Bayesian and k-NN

Wang Tao, Huo Liang, and Yang Liu

Department of Information Management and Engineering Hebei Finance University China
Experimental Center of Economics and Management Hebei University China

Abstract. In order to improve the classification accuracy and speed, classification of the structure of this paper has been improved, is proposes a combination of Bayesian and k-nearest neighbor classifier model, which combines Bayesian classification method of classification rate fast and k-nearest neighbor method with higher classification accuracy advantages. Experimental results show that the method to ensure the classification rate under the premise of effectively improving the classification accuracy.

Keywords: Bayesian classification, k-NN, Text classification.

1 Introduction

Text classification is a classification algorithm using a known sample set to learn, to train a classifier, using the classification of unknown samples were automatically classified category. Commonly used classification algorithms are Bayesian methods, k-NN method, the center vector method, decision tree methods, support vector machine methods. The classification algorithm has advantages and disadvantages, so the use of a combination of strategy, these algorithms together, whichever benefits to its shortcomings, is a classification algorithm to optimize a variety of ways. Combination of strategies used in two forms: horizontal combinations and vertical combinations. Each horizontal layer which contains only the combination is a classifier, the classification level on the forecast with all properties as initial input to the next level, until the final layer gives the classification results. Vertical combination is the first layer classifier for the different categories of independent samples, the second layer of the results of these classifiers are combined according to some strategy[1] [2].

This paper presents a cross classification method using a central vector as the first layer classifier, Bayesian classifier as the second layer, k-NN classification as a third layer classifier. Experimental results show that the classification of the classification accuracy is higher than the single classifier accuracy.

2 Combination Classification Model

First, the training sample pretreatment. Text that is currently used mainly vector space model, the text that point to a vector space.

Z. Du (Ed.): Proceedings of the 2012 International Conference of MCSA, AISC 191, pp. 57–61.
springerlink.com

(1) First, the text word by word as a vector of these dimensions to represent text, which will match the text message said the problem with the vector into a vector space representation and matching.
(2) Then the text vector dimensionality reduction algorithm, scanning the document vector vocabulary, with synonyms, delete the word frequency is very small and the emergence of frequent words.
(3) The use of TF-IDF representation of lexical weights for processing.
(4) Use of mutual information (MI) method for extracting feature items, select a category in a high probability, select other categories of low-probability words as feature words.

2.1 The First Layer Classifier

The first layer classifier of this paper choose the simple algorithm, fast center vector method, the basic idea is, according to belong to a class of all training text vector calculation of the category of the center vector, and then calculate the text to be classified with each class centre vector similarity, and put it into the largest category similarity.

(1) The text of the test word, the formation of new test vectors.
(2) The characteristics of each type of query word contains the number of the vector.
(3) certain number of test vectors contain more text and description of the test in this category, the more similar, so the test text test vector containing the number assigned to a class of most, if category number up to more than one class, then the test text is not classified as an input to the next level of classification.

2.2 The Second Layer Classifier

The second layer of a Bayesian classifier selection method is faster but the classification accuracy rate is low, so the text in the text of the test, when added to forecast the probability of a threshold control, encountered a probability below this threshold value, will enter the next level, so you can improve the performance of Bayesian methods.

Specific steps are as follows:

(1) According to Equation 1 for each category feature words t_k is the probability vector w_k.

$$w(t_k \mid C_j) = \frac{1 + \sum_{i=1}^{|D|} tf(t_k, d_i)}{M + \sum_{s=1}^{M} \sum_{i=1}^{|D|} tf(t_s, d_i)}$$

Formula 1

Among them, the $w(t_k \mid c_j)$ is the probability of t_k appear in c_j, c_j is a category, d_i is an unknown type of text, t_k appear in the d_i feature

items, $tf(t_k, d_i)$ as the number of occurrences in the d_i ,M is the general characteristics of the training set number of words, |D| is the number of such training text[3].

(2) In the new text arrives, the new text word, and then the text in accordance with the formula 2 to calculate the probability of belonging to the class:

$$P(C_j \mid d_i) = \frac{P(C_j)\prod_{k=1}^{M} P(t_k \mid C_j)^{tf(t_k, d_i)}}{\sum_{r=1}^{|C|} P(C_r)\prod_{k=1}^{M} P(t_k \mid C_r)^{tf(t_k, d_i)}}$$

Formula 2

Which $$p(c_j) = \frac{\text{The number of training documents of } c_j}{\text{The totalnumber of training documents}}$$

$P(C_r)$ is the similar meaning, $|C|$ is the total number of categories.

(3) To compare the new text belongs to all classes of probability, the probability of finding the maximum value, if the value is greater than the threshold value, the text assigned to this category, if less than the threshold, then the next layer of classification.

2.3 The Third Layer Classification

Consider these two main categories of classification rate, the third layer selected a better classification performance k-NN method.[4]

$$S_{ij} = CosD_{ij} = \frac{\overrightarrow{D_i}\,\overrightarrow{D_j}}{\left|\overrightarrow{D_i}\right|\left|\overrightarrow{D_j}\right|} = \frac{\sum_{k=1}^{M} w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^{M} (w_{ik})^2 \cdot \sum_{k=1}^{M} (w_{jk})^2}}$$

Formula 3

(1) According to Formula 3 with the training text and the test text on each similarity;

(2) According to text similarity, text set in the training and testing of selected text in the k most similar texts;

(3) Test the k neighbors in the text, followed by calculating the weight of each class is calculated as Formula 4.

$$p(\vec{x}, C_j) = \begin{cases} 1 & if \sum_{\vec{d}_i \in KNN} Sim(\vec{x}, \vec{d}_i) y(\vec{d}_i, C_j) - b \geq 0 \\ 0 & others \end{cases}$$

Formula 4

Among them, $\vec{x}$ the eigenvector for the new text, $Sim(\vec{x}, \vec{d}_i)$ for the similarity calculation formula 4, b is the threshold, this value is a value chosen to be optimized, while the value of $y(\vec{d}_i, C_j)$ is 1 or 0, $\vec{d}_i$ Belongs to class C_j ,then the function value is 1, otherwise 0.

(4) Compare the weight class, the text assigned to the heaviest weight on that category.

3 Experimental Results and Analysis

In this paper, achieve the above types of Chinese text classification system on the Windows XP operating system, the database is SQL SERVER2000.

Corpus used in this article from the website "Chinese natural language processing open platform" provided by the Fudan University, Dr. Li Ronglu upload, repeat the text of which initial treatment and damage to documents, training documents 9586, test text 9044, is divided into 20 categories, including training and test set ratio of 1:1. First, preliminary processing, training text with the second scan of the sub-word method, feature selection using mutual information method is the test text has been over the word and feature selection process, and on this basis for classification. After statistical classification of the time contains only the time, does not include the initial processing time[5].

3.1 The Results

Table 1. The Accuracy of Classification

classification method	Center vector	Bayesian classification	k-NN (k=5)	Combined classification
Accuracy rate rate	65.72%	66.03%	71.51%	78.69%
Time(s) consuming(s)	437	492	3636	1311

The experimental results obtained, the center's fastest vector classification method, and only 437 seconds, and the Bayesian method than the classification accuracy increased by 0.5%. Bayesian methods and improved methods of combining k-nearest neighbor method of classification accuracy rate is highest, the classification accuracy rate higher than the Bayesian method 19.7%, higher than the k-nearest neighbor method 10%; it takes is a simple Bayesian 2.66 times approach, but only k-nearest neighbor method of 36.06%[6].

3.2 Experimental Results

Because K - nearest neighbor method each decision-making need to all training samples for comparison, calculation are deferred to the classification process, so the classification speed is slow. The Bayesian classifier, only needs to calculate the product to the test documents are classified, so the test time is short.

For improved Bayesian classification method, only calculating the containing characteristic word number, so than naive Bayesian classification method is faster, at the same time it with some small probability event, by naive Bayesian can classify text, can be correctly classified, so the classification accuracy rate than the naive Bayesian method is improved.

For improved Bayesian method and K - nearest neighbor method combining method, combines Bayesian classification speed and K - nearest neighbor classification accuracy rate high, the experimental results compared with the ideal.

4 Conclusion

This article discusses two is generally considered good classification, naive Bayesian method and K - nearest neighbor method, and the Bayesian method was improved, finally proposed combining Bayesian method and K - nearest neighbor method a new method. In this paper, to achieve the above four methods in the same corpus -- "Chinese natural language processing open platform", From the experimental results, the improved Bayesian classification method is faster, suitable for larger data sets and on-line real-time classification; Bayesian method and K - nearest neighbor method a new method combining classification accuracy was highest, applicable in high accuracy, at the same time its classification speed compared with the pure K - nearest neighbor method raise, can also handle a larger sample set of data or for real time processing.

References

[1] Langley, P., Iba, W., Thompson, K.: An analysis of Bayesian classifiers. In: Proceedings of the Tenth National Conference on Artificial Intelligence, pp. 223–228. AAAI Press, Menlo Park (1992)

[2] Geiger, F.N., Goldszmidt, D.: Bayesian network classifiers. Machine Learning 29(2/3), 131–163 (1997)

[3] Ramoni, M., Sebastiani, P.: Robust Bayes classifiers. Artificial Intelligence 125(122), 209–226 (2001)

[4] Cheng, J., Greiner, R.: Comparing Bayesian network classifiers. In: Laskey, K.B., Prade, H. (eds.) Proc. of the 15th Conf. on Uncertainty in Artificial Intelligence, pp. 101–108. Morgan Kaufmann Publishers, San Francisco (1999)

[5] Susumu, T.: A study on multi relation coefficient among variables. Proceedings of the School of Information Technology and Electronics of Tokai University 4(1), 67–72 (2004)

[6] Bocchieri, E., Mark, B.: Subspace Distribution clustering hidden Markov model. IEEE Transactions on Speech and Audio Processing 9(3), 264–275 (2001)