Improved KNN Classification Algorithm by Dynamic Obtaining K

An Gong and Yanan Liu

College of Computer and Communication Engineering
China University of Petroleum (East China), Dongying, China
gongan0328@sina.com, jsjliuyanan@163.com

Abstract. KNN algorithm which is one of the best methods of text classifying in the vector space model (VSM) is a simple, example based and none-parameter method. But in the KNN algorithm, the fixed K value ignores the influence of the category and the document number of training text. So, selecting the correct K value can achieve better classification results. This paper proposes a kind of dynamic obtain k-valued for KNN classification algorithm, experimental results show that the dynamic obtain k-valued KNN classification algorithm with high performance.

Keywords: KNN classification algorithm, k-valued, dynamic obtain.

1 Instruction

KNN classification algorithm is widely used in the field of machine learning and data mining, and has been proved to be one of the best text classification methods under vector space model (VSM) [1]. However, KNN algorithm has inherent disadvantages [2]: (1) When the training sample set is too large or too many feature items, the algorithm's complexity of time and space is high, its time complexity is O (n*m) (n is the characteristic dimension, m is the sample set size), leading to the efficiency of KNN algorithm will be decreased; (2) There is not a strong basis for the selection of the value of K, when the value of K is small, the neighbors' interference is small, but following low accuracy; when the K value is large, the method has good accuracy, but the interference of neighbor is very large.

For the lacking of KNN classification algorithm, domestic and foreign scholars put forwards some improves, which can be divided into three categories: (1)Sample cutting;(2)Reducing the dimension of high-dimensional vectors, such as the method based on Latent Semantic Analysis (LSA) [3], the method based on feature vector polymerization [4]; (3)Improving the feature selection method.

In the KNN classification algorithm, many experimenters use a fixed K value. This value is an empirical value, which is the result of a large number of experiments and has no reliable theoretical basis. If the K value is too large, the text tends to belong to the class which contains more texts, classification performance is poor; If K value is too small, text has too few neighbors, this will reduce the classification accuracy. Almost all of the calculations have taken place in the classification stage, and the classification results depend on the K value. Therefore, how to select appropriate K values is critical

G. Shen and X. Huang (Eds.): ECWAC 2011, Part I, CCIS 143, pp. 320–324, 2011.

to improve the performance of KNN classification algorithm and it will be a research focus in the field of data mining. Then this paper proposes a kind of dynamic obtain k-valued for KNN classification algorithm.

2 Classical KNN Classification Algorithm

KNN classification algorithm is based on the examples and the parameters of text classification method, which is a relatively mature theory of simple machines for learning algorithm [5]. The method has simple idea: According to the traditional vector space model, the content of the text is formalized as the weighted feature vector in feature space, as $D=D(T_0, W_0; T_1, W_1; ...; T_n, W_n)$,the various dimensions of the various features used to corresponding to characterize properties of the document. Calculate the similarity that between the test text vector and the training set vector, then by sorting the similarity to select the K most similar to the test text. Accumulate the same type of text similarity; the test text belongs to that obtaining the maximum similarity. The algorithm steps are:

(1). Using the collection of features items to descript the training text vectors.
(2). Upon the test text's arrival, using the word segmentation to process the test text, determining its vector representation.
(3). Calculate the similarity between the test text vector and the training set vector, the formula as follows:

$$sim\left(d_i, d_j\right) = \frac{\sum_{k=1}^{n} a_{ik} \times a_{jk}}{\sqrt{\left(\sum_{k=1}^{n} a^2{}_{ik}\right)\left(\sum_{k=1}^{n} a^2{}_{jk}\right)}}$$

Where, d_i is the text feature vector under test, d_j is the text feature vector for the j type, a_{ik}, a_{jk} for the K-dimensional vector corresponding to the first.

(4). According to the similarity of text, select the K most similar to the test text. Calculate the weight of each class according to the following formula:

$$P(i, C_x) = \sum sim(d_i, d_j) y(d_j, C_x)$$

Where, $y(d_j, C_x)$ is class attribute function, if d_j belongs to class C_x, the function value is 1, otherwise the function values is 0.

(5). Comparison of the various weights, the test text was assigned to the class that has the maximum weight.

3 The Selection of K Value

The selection of K value has relationship with the class of texts and the number of texts in every class. Taking different categories with the different number of texts into account, this paper proposes a dynamic obtain K value.

Suppose the class of training text is N, the total number of the text is M. Each class contains different number of text. We use this training set to classify the test text X and

still use the Euclidean distance between the two calculation methods to calculate the similarity between X and other training texts.

$$sim(d_x, d_i) = \frac{\sum_{k=1}^{s} a_{xk} \times a_{ik}}{\sqrt{\left(\sum_{k=1}^{s} a^2{}_{xk}\right) \times \left(\sum_{k=1}^{s} a^2{}_{ik}\right)}} \tag{1}$$

Where, d_x is the text feature vector for X, d_i is the text feature vector for the i type, a_{xk}, a_{ik} for the K-dimensional vector corresponding to the first. S is the number of dimension.

Taking different categories with the different number of texts into account, the paper proposes using the number of various types of training set and the average similarity to fix the selection of K value. Can avoid large different numbers of training texts leads to the results of classification for test text tend to the class which has the larger number of training texts.

We use formula (1) to calculate the similarity between X and other training texts. After the similarity calculation with one class, we use the number of the training texts to fix the total value of similarity. The formula as follows:

$$\text{Sim(j)} = \left(1 - \frac{n}{M}\right) \sum_{i=1}^{n} sim(d_x, d_i) \tag{2}$$

Where, j is the type of training set, n is number of training texts for j types, M is the total number of the text. In formula (2), the class containing a larger number of training text, the value of n / M is larger, and the value of is smaller. Using (1-n/M) to fix the j type's similarity values, can eliminate the classification results are not accurate because of the number of training text is too large or too small.

From the formula (2), we can get the amendment similarity of the various types of training text. Then cumulate the amendment similarity, we can get an average of SIM (avg), as shown in formula (3), where N is the class of training text.

$$\text{SIM(avg)} = \frac{1}{N} \sum_{j=1}^{N} \text{Sim(j)} \tag{3}$$

To SIM(avg) as a standard, we can divided the similarity which between X and other training texts into two parts. One sim is greater than SIM(avg), the other part's sim is less than SIM(avg). We reserve part which has larger values. We calculate the number of texts in this part, and then assigned the value of this number to K. Therefore, we obtain the dynamic K value.

4 Experimental Results and Analysis

4.1 Experimental Data

This experimental data is provided by Tan Songbo who is the doctor form CAS Institute of Computing. This corpus is divided into two layers, the first layer of 12 categories, and the second layer of 60 categories which collects 14,150 documents. We select 8 categories of document data which from the first player to test. There are 1647

documents, including finance, geography, computers, property, education, automobile, health, entertainment. The training set has 896 documents and the test set has 751 documents (Table 1).

Table 1. Experimental data

Category	The training set	The test set
Finance	120	95
Geography	80	70
Computers	130	110
Property	95	80
Education	125	90
Automobile	100	93
Health	156	138
Entertainment	90	75

4.2 Experimental Results

This paper uses two kind of KNN classification algorithm to classify the selected documents, one is the fixed K value KNN classification algorithm, the other is the dynamic obtain K value KNN classification algorithm. For the first method, we assign 8, 20 to K, because the large number of experiments show that assign 8 or 20 to K, the classification can get better accuracy.

Internationally accepted the basis of the evaluation for the classification results are the recall (R), precision (P) and the value of F1 [6].

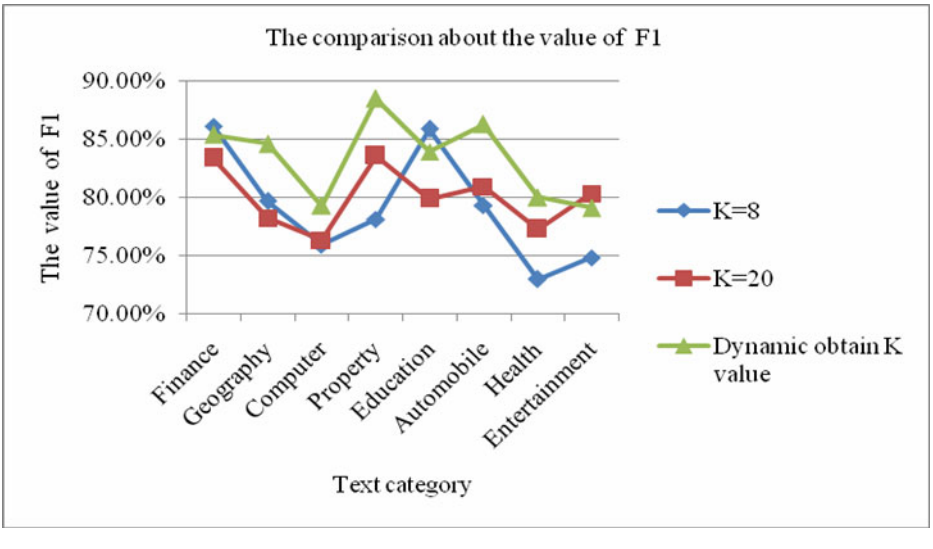

Fig. 2.

Fig.2 shows the comparison of F1, which from two kinds of KNN classification algorithm, and one of the K value is 8 or 20, and the other's k value is obtained by a dynamic method. From the line chart, in the KNN classification algorithm, the F1 value of the dynamic K value is higher than that of the fixed K value. From table1, we know that, the geography's number of training set and test set is about twice more than that of health category .As shown in chart4.3, when K=8 or k=20, the F1 values of the two categories are varied. For geography category, when k=8 there is better classification results, but for health category, it will get better classification results when k=20.However, the dynamic obtain k value algorithm can get the best classification results. For property category and education category, there is little difference in their numbers of training set and test set. Using the fixed K value algorithm, the F1 values of the two categories are varied, too and the dynamic obtain k value algorithm can get the better classification results. Therefore, we conclude that the dynamic K value for the classification algorithm can effectively avoid the impact of the size of the training text set and the test text set and has high classification efficiency.

5 Conclusion

Based on the analysis and studying classical KNN classification algorithm, this paper proposes a kind of dynamic obtain k value for KNN classification algorithm. In this new method, after the similarity calculation with one class, we use the number of the training texts to fix the total value of similarity, then cumulative the amendment similarity of each class; we can get an average of SIM (avg). We assign the number of the text whose value of similarity is larger than the SIM (avg) to K. Experimental results show that the dynamic obtain k-valued KNN classification algorithm can effectively avoid the impact of the size of the training text set and the test text set and can get high classification efficiency.

References

[1] Pang, J., Bu, D., Bai, S.: Based on vector space model the automatic Text Classification System and implementation. Computer Applications 18(9), 23–26 (2001)
[2] Han, E.H., George, K., Vipin, K.: Text categorization using weight adjusted k-nearest neighbor classification: Technical Report. University of Minnesota (2000)
[3] Deerwester, S., Dumas, S., Furnas, G., Landauer, T., Harsrtian, R.: Indexing by Laent Semantic Analysis. Journal of the American Society for Information Science 41(6), 391–417 (1994)
[4] Zhang, X., Li, Y., Wang, H.: Using Characteristics of polymerization to improve the KNN algorithm for Chinese Text Classification. Northeastern University (Natural Science) (3), 229–232 (2003)
[5] Liu, B., Yang, L., Yuan, F.: Improved KNN method and its application in Chinese text categorization. West China University of Technology (Natural Science) 27(2), 33 (2008)
[6] Song, F., Gao, L.: The evaluation about text classification performance. Computer Engineering 30(13), 107–109 (2004)