Improving kNN Text Categorization by Removing Outliers from Training Set*

Kwangcheol Shin, Ajith Abraham, and Sang Yong Han**

School of Computer Science and Engineering, Chung-Ang University,
221, Heukseok-dong, Dongjak-gu, Seoul 156-756, Korea
kcshin@archi.cse.cau.ac.kr,
ajith.abraham@ieee.org, hansy@cau.ac.kr

Abstract. We show that excluding outliers from the training data significantly improves kNN classifier, which in this case performs about 10% better than the best know method—Centroid-based classifier. Outliers are the elements whose similarity to the centroid of the corresponding category is below a threshold.

1 Introduction

Since late 1990s, the explosive growth of Internet resulted in a huge quantity of documents available on-line. Technologies for efficient management of these documents are being developed continually. One of representative tasks for efficient document management is text categorization, called also classification: given a set of training examples assigned each one to some categories, to assign new documents to a suitable category.

A well-known text categorization method is kNN [1]; other popular methods are Naive Bayesian [3], C4.5 [4], and SVM [5]. Han and Karypis [2] proposed the Centroid-based classifier and showed that it gives better results than other known methods.

In this paper we show that removing outliers from the training categories significantly improves the classification results obtained with kNN method. Our experiments show that the new method gives better results than the Centroid-based classifier.

2 Related Work

Document representation. In both categorization techniques considered below, documents are represented as keyword vectors according to the standard vector space model with *tf-idf* term weighting [6, 7]. Namely, let the document collection contains in total *N* different keywords. A document *d* is represented as an *N*-dimensional vector of term weight *t* with coordinates

* Work supported by the MIC (Ministry of Information and Communication), Korea, under the Chung-Ang University HNRC-ITRC (Home Network Research Center) support program supervised by the IITA (Institute of Information Technology Assessment).

** Corresponding author.

A. Gelbukh (Ed.): CICLing 2006, LNCS 3878, pp. 563 – 566, 2006.

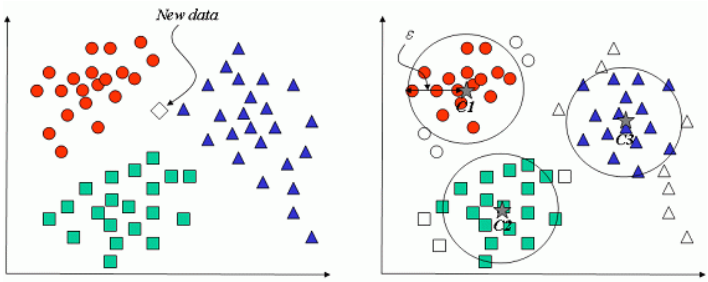

Fig. 1. Example of classification

$$w_{td} = \frac{f_{td}}{\max_t f_{td}} \log \frac{n_t}{N}, \tag{1}$$

where f_{td} is the frequency of the term t in the document d and n_t is the number of the documents where the term t occurs. The similarity between two vectors d_i and d_j is measured as the cosine of the angle between them:

$$s(d_i, d_j) = \cos\left(\theta(x_i, x_j)\right) = \frac{d_i^T d_j}{\| d_i \| \| d_j \|}, \tag{2}$$

where θ is the angle between the two vectors and $\| d \|$ is the length of the vector.

kNN classifier [1]**.** For a new data item, k most similar elements of the training data set are determined, and the category is chosen to which a greater number of elements among those k ones belong; see Figure 1, left.

Centroid-based classifier [2]**.** Given a set S_i of documents—the i-th training category, its center is defined as its average vector:

$$\vec{C}_i = \frac{1}{|S_i|} \sum_{\vec{d} \in S_i} \vec{d} \tag{3}$$

where $|S_i|$ is the number of documents in the category. For a new data idem the category is chosen that maximizes the similarity between the new item and the centers of each category. This was reported as the best known classifier so far [2].

3 Proposed Method

We observed that the training data items that are far away from the center of its training category reduce the accuracy of classification. Our hypothesis is that those items represent noise and not useful training examples and thus decrease the classification accuracy. Thus we exclude them from consideration; see Figure 1, right. Specifically, at the training stage we calculate the center C_i of each category S_i using (2). Then we form new categories by discarding outliers:

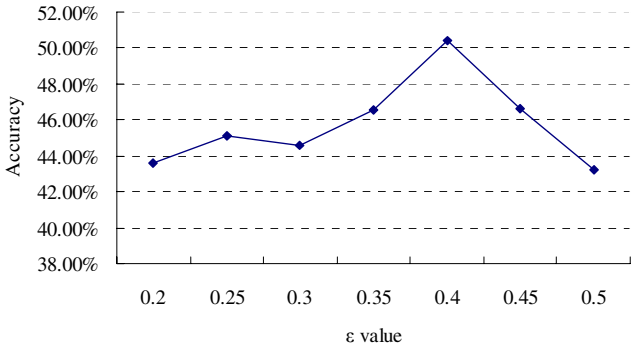

Fig. 2. Different accuracy according to ε value at using 80% of test dataset

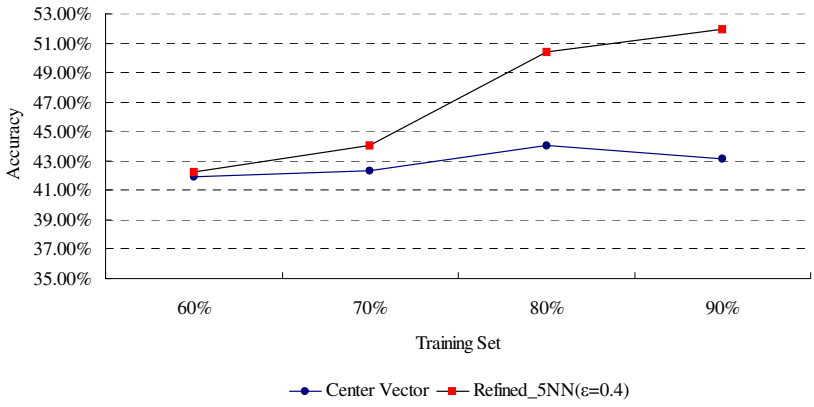

Fig. 3. Test Results

$$S_i' = \{d \in S_i : Sim(d, \vec{C}_i) > \varepsilon\}; \quad (4)$$

in the next section we discuss the choice of the threshold ε. Finally, we apply the kNN classifier using these modified categories.

4 Experimental Results

We used the 20-newsgroup dataset to evaluate the performance of the proposed method. The dataset consists of 20 classes of roughly 1000 documents each. We used MC [8] program to build the document vectors. We implemented our modified kNN method with $k = 5$ and compared it with the Centroid-based classification. As Figure 2 shows, our method provides the best performance with $\varepsilon \approx 0.4$. Figure 3 shows how the classification accuracy depends on the percentage of training dataset over total dataset. We obtain 9.93% improvement over the original Centroid-based classification.

5 Conclusion

We have presented an improved kNN classifier, combining it with the idea of the Centroid-based method. The improvement consists in removing outliers from the categories of the training dataset. Our method shows almost 10% better accuracy than the original Centroid-based classifier, which was reported in [2] as the most accurate text categorization method. In the future, automatic choice of the threshold value ε is to be considered.

References

1. W. W. Cohen and H. Hirsh. Joins that generalize: Text Classification using WHIRL. In Proc. of the Fourth Int'l Conference on Knowledge Discovery and Data Mining, 1998.
2. E. Han and G. Karypis. Centroid-Based Document Classification: Analysis & Experimental Results. Principles of Data Mining and Knowledge Discovery, p. 424–431, 2000.
3. D. Lewis and W. Gale. A sequential algorithm for training text classifiers. SIGIR-94, 1994.
4. J. Ross Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.
5. V. Vapnik. The Nature of Statistical Learning Theory. Springer, 1995.
6. G. Salton and. M. J. McGill, *Introduction to Modern Retrieval*. McGraw-Hill, 1983.
7. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.
8. Dhillon I. S., Fan J., and Guan Y. Efficient Clustering of Very Large Document Collections. *Data Mining for Scientific and Engineering Applications*, Kluwer, 2001.