Enhanced Centroid-Based Classification Technique by Filtering Outliers

Kwangcheol Shin, Ajith Abraham, and SangYong 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. Document clustering or unsupervised document classification has been used to enhance information retrieval. Recently this has become an intense area of research due to its practical importance. Outliers are the elements whose similarity to the centroid of the corresponding category is below some threshold value. In this paper, we show that excluding outliers from the noisy training data significantly improves the performance of the centroid-based classifier which is the best known method. The proposed method performs about 10% better than the centroid-based classifier.

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 continuously. One of representative tasks for efficient document management is text categorization, also called as classification. Given a set of training examples assigned each one to some categories, the task is to assign new documents to a suitable category. A fixed collection of text is clustered into groups or clusters that have similar contents. The similarity between documents is usually measured with the associative coefficients from the vector space model, e.g., the cosine coefficient.

A well-known text categorization method is kNN [1]; other popular methods are Naïve Bayesian [3], C4.5 [4], genetic programming [10], self organizing maps [11] artificial neural networks [9] 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 by using the Centroid-based method. Our experiments show that the new method gives better results than the Centroid-based classifier.

The paper is organized as follows. In Section 2, some related work is presented followed by the details of the proposed method in Section 3. Experiment results are presented in Section 4 and some Conclusions are given towards the end.

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*

[*] Corresponding author.

Petr Sojka, Ivan Kopeček and Karel Pala (Eds.): TSD 2006, LNAI 4188, pp. 159–163, 2006.

term weighting [6,7]. For definition purposes, let the document collection contains total N different keywords. A document d is represented as an N-dimensional vector of term weight t with coordinates

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

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 documents d_i and d_j is measured using the cosine measure widely used in information retrieval—the cosine of the angle between them:

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

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.

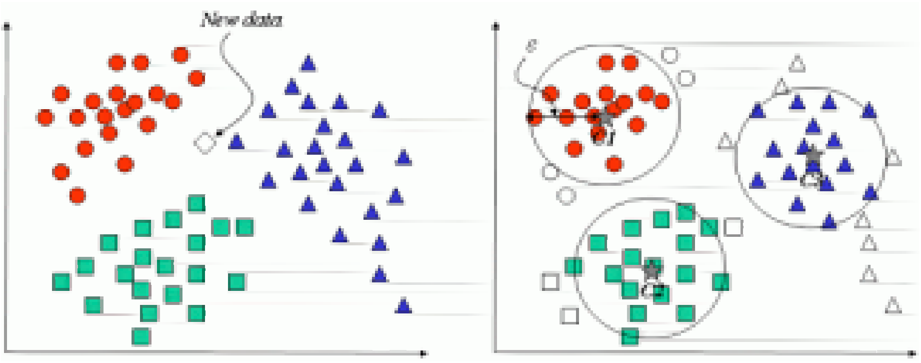

Fig. 1. Example of classification

Centroid-based classifier [2]: Given a set Si 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 $|Si|$ is the number of documents in the category. For a new data item the category is chosen that maximizes the similarity between the new item and the center 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 tend to reduce the accuracy of classification. Our hypothesis is that those items merely represent noise and not provide any 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 Ci of each category Si using (2). Then we form new categories by discarding the outliers:

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

in the next section we discuss the choice of the threshold ε. After refining training data, we recalculate the center of each category:

$$\vec{C}_i' = \frac{1}{|S_i'|} \sum_{\vec{d} \in S_i'} \vec{d} \tag{5}$$

And finally the Centroid-based classifier is applied to get the results.

4 Experimental Results

To evaluate the efficiency of the proposed method, we used two different datasets. First one is the 20-newsgroup dataset which has many noisy data and the other is the popular Reuter-21578 R10 dataset which doesn't have noisy data.

4.1 20-Newsgroup Dataset

At first, we use the 20-newsgroup dataset to evaluate performance of the proposed method. The dataset has 20 categories of roughly 1000 documents each. We used MC [8] program to build the document vectors. We implemented our modified Centroid-based classification and compared with the Centroid-based classification and kNN method with $k = 5$. As illustrated

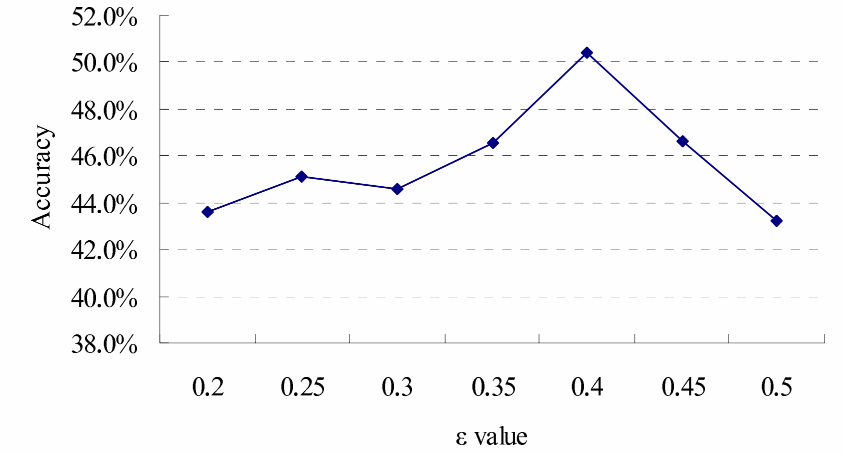

Fig. 2. Different accuracy according to ε value for 80% of 20-newsgroup data

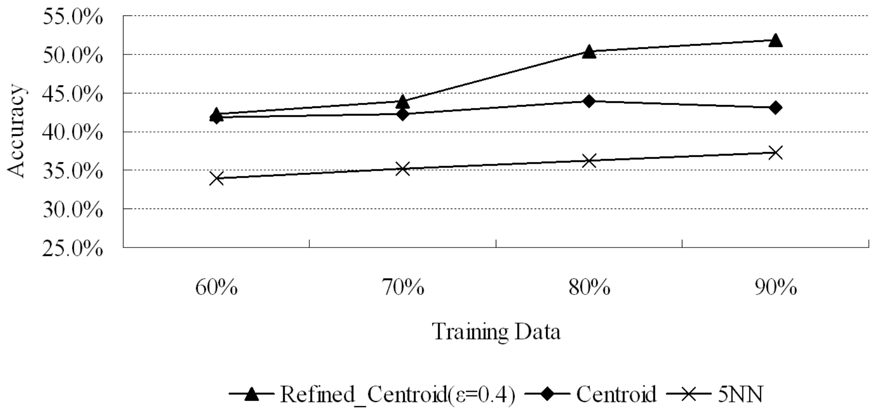

Fig. 3. Test results

in Figure 2, the proposed method provides the best performance with $\varepsilon = g.4$. Figure 3 shows how the classification accuracy is affected due to the percentage of training dataset over total dataset. We obtained 9.93% improvement over the original Centroid-based classification and 32.11% over 5NN. Improvements clearly show that the proposed method worked very well for the noisy dataset.

4.2 Reuter-21578 R10 Dataset

We also applied our method to the popular Reuter-21578 R10 dataset, which has 10 categories and each category has different number of data. Because of being categorized by human indexers, it doesn't have noise data. Figure 4 illustrates that for noiseless dataset, like Reuter-

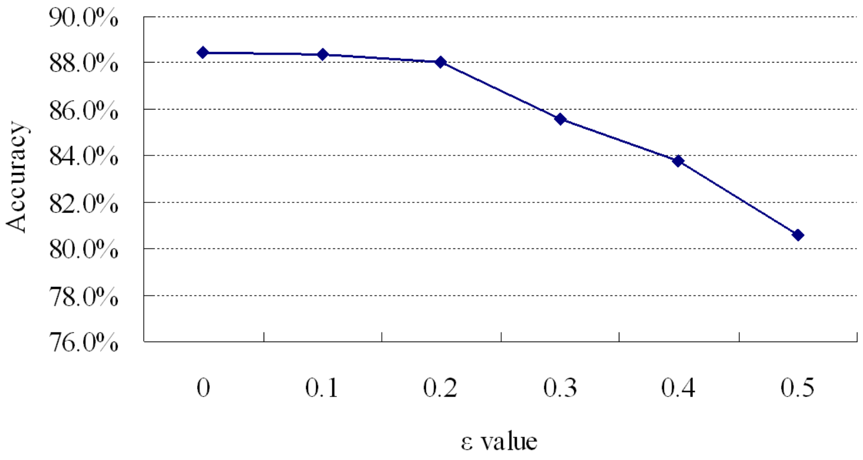

Fig. 4. Performance (accuracy) of the algorithm for different ε values for 60% of Reuter-21578 R10 dataset

21578 R10, the performance is not improved when compared to the Centroid-based classification approach (note: it is same with Centroid-based classification when we set 0 to ε).

5 Conclusion

We have presented an improved Centroid-based classifier. The improvement consists in removing outliers from the categories of the training dataset. Our method shows almost 10% better accuracy for noisy dataset 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

Acknowledgments. 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).

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 and 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, In SIGIR-94, 1994.
4. J. Ross Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA, 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.
9. MacLeod, K. *An application specific neural model for document clustering.* Proceedings of the Fourth Annual Parallel Processing Symposium, vol. 1, p. 5–16, 1990.
10. Svingen, B. *Using genetic programming for document classification.* FLAIRS-98. Proceedings of the Eleventh International Florida Artificial Intelligence Research, p. 63–67, 1998.
11. Hyotyniemi, H. *Text document classification with self-organizing maps.* STeP '96 - Genes, Nets and Symbols. Finnish Artificial Intelligence Conference, p. 64–72, 1996.
12. Lam, Wai and Low, Kon-Fan *Automatic document classification based on probabilistic reasoning: Model and performance analysis.* Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Vol. 3, p. 2719–2723, 1997.