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

* 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.