**1.** **ABSTARCT**

k is the most important parameter in a text categorization system based on k-Nearest Neighbor algorithm (kNN).In the classification process, k nearest documents to the test one in the training set are determined firstly. Then, the predication can be made according to the category distribution among these k nearest neighbors. Generally speaking, the class distribution in the training set is uneven. Some classes may have more samples than others. Therefore, the system performance is very sensitive to the choice of the parameter k. And it is very likely that a fixed k value will result in a bias on large categories. To deal with these problems, we propose an improved kNN algorithm, which uses different numbers of nearest neighbors for different categories, rather than a fixed number across all categories. More samples (nearest neighbors) will be used for deciding whether a test document should be classified to a category, which has more samples in the training set. Preliminary experiments on Chinese text categorization show that our method is less sensitive to the parameter k than the traditional one, and it can properly classify documents belonging to smaller classes with a large k. The method is promising for some cases, where estimating the parameter k via cross-validation is not allowed.

**2.INTRODUCTION**

k-Nearest Neighbor is one of the most popular algorithms for text categorization. Many researchers have found that the kNN algorithm achieves very good performance in their experiments on different data sets

The idea behind k-Nearest Neighbor algorithm is quite straightforward. To classify a new document, the system finds the k nearest neighbors among the training documents, and uses the categories of the k nearest neighbors to weight the category candidates. One of the drawbacks of kNN algorithm is its efficiency, as it needs to compare a test document with all samples in the training set. In addition, the performance of this algorithm greatly depends on two factors, that is, a suitable similarity function and an appropriate value for the parameter k.

In this paper, we focus on the selection of the parameter k. In the traditional kNN algorithm, the value of k is fixed beforehand. If k is too large, big classes will overwhelm small ones. On the other hand, if k is too small, the advantage of kNN algorithm, which could make use of many experts, will not be exhibited. In practice, the value of k is usually optimalized by many trials on the training and validation sets. But this method is not feasible in some cases where we have no chance to do cross-validation, such as online classification. To deal with this problem, we propose a revised k-Nearest Neighbor algorithm, which uses different k values for different classes, rather than a fixed k value for all classes.