An *k*NN Model-based Approach and Its Application in Text Categorization

Gongde Guo[1], Hui Wang[1], David Bell [2], Yaxin Bi [2], and Kieran Greer [1]

[1] School of Computing and Mathematics, University of Ulster
Newtownabbey, BT37 0QB, Northern Ireland, UK
{G.Guo, H.Wang, Krc.Greer} @ulst.ac.uk

[2] School of Computer Science, Queen's University Belfast
Belfast, BT7 1NN, UK
{DA.Bell, Y.Bi} @qub.ac.uk

Abstract An investigation has been conducted on two well known similarity-based learning approaches to text categorization. This includes the *k*-nearest neighbor (*k*-NN) classifier and the Rocchio classifier. After identifying the weakness and strength of each technique, we propose a new classifier called the *k*NN model-based classifier by unifying the strengths of *k*-NN and Rocchio classifier and adapting to characteristics of text categorization problems.

A text categorization prototypes system has been implemented and then evaluated on two common document corpora, namely, the 20-newsgroup collection and the ModApte version of the Reuters-21578 collection of news stories. The experimental results show that the *k*NN model-based approach outperforms the *k*-NN, Rocchio classifier.

1 Introduction

Text categorization (TC) is the task of assigning a number of appropriate categories to a text document. This categorization process has many applications such as document routing, document management, or document dissemination [1]. Traditionally each incoming document is analyzed and categorized manually by domain experts based on the content of the document. A large amount of human resources have to be spent on carrying out such a task. To facilitate the process of text categorization, automatic categorization schemes are required. The goal of text categorization is to learn such categorization schemes that can be used to classify text documents automatically.

There are many categorization schemes addressed for this automatic text categorization task in text categorization literature. This includes Naïve Bayes (NB) probabilistic classifiers [2], Decision Tree classifiers [3], Decision Rules [4], regression methods [5], Neural Network [6], *k*-NN classifiers [5, 7], Support Vector Machine (SVMs) [8, 9], and Rocchio classifiers [10, 11] etc. In many applications, such as web mining for a large repository, the efficiency of those schemes is often the key element to be considered. Sebastiani has pointed this out in his survey on text categorization [12].

k-NN and Rocchio are two classifiers frequently used for TC, and they are both similarity based. The *k*-NN algorithm directly uses the training examples as a basis for computing similarity. For a data record t to be classified, its k nearest neighbors