Unsupervised Text Classification using Kohonen's Self Organizing Network

Nirmalya Chowdhury and Diganta Saha

Department of Computer Science and Engineering,
Jadavpur University, Kolkata – 700 032, India.
nir63@vsnl.net

Abstract. A text classification method using Kohonen's Self Organizing Network is presented here. The proposed method can classify a set of text documents into a number of classes depending on their contents where the number of such classes is not known a priori. Text documents from various faculties of games are considered for experimentation. The method is found to provide satisfactory results for large size of data.

1 Introduction

Text classification research and practice [2] have exploded in the past decade. There are different methods of text classification such as methods based on ontologies and key words [3] or machine learning techniques [4]. We present here an unsupervised text classification technique that uses a special type of neural network called Kohonen's Self Organizing Network. The novelty of the method is that it automatically detects the number of classes present in the given set of text documents and then it places each document in its appropriate class. The method initially uses the Kohonen's Self Organizing Network to explore the location of possible groups in the feature space. Then it checks whether some of these groups can be merged on the basis of a suitable threshold to result in desirable clustering. These clusters represent the various groups or classes of texts present in the set of given text documents. Then these groups are labeled on the basis of frequency of the class titles found in the documents of each group. The proposed method needs no *a priori* knowledge about the number of classes present in a given set of text documents.

The next section presents the steps involved in the formation of the pattern vector for each document for clustering followed by the statement of the clustering problem.

2 Statement of the Problem

Given a set of text documents, the steps adopted to extract the features and form the pattern vectors for all the documents in the given set of text documents are as follows.

Step 1: Remove all stop words such as 'the', 'a', 'an' etc., and also all functional words such as adverbs, preposition, conjunction etc, from the text of all the documents in the given set.