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.

A. Gelbukh (Ed.): CICLing 2005, LNCS 3406, pp. 715 – 718, 2005.

Step 2: Remove the words that have the value W below a threshold 0.9. W is an elimination factor that is calculated as follows.

W_{EF} = Number of occurrence in its own context / Total number of occurrences in all contexts

We chose the value of ***W*** empirically. Thus, the words that has W below the threshold 0.9 do not participate in evaluation

Step 3: Choose the remaining words as the features for document classification.

Step 4: Create one pattern vector for each document with the features selected in step 3. The numeric value for each component of such vector would be the number of occurrence for the particular word corresponding to that component in the given document.

Note: Step 2 eliminates all those words that have almost no discriminatory significance so far classification is concerned. For instance, the word "play" can be expected to occur frequently in all the documents of various faculties of games. Thus it will have a relatively lower value of W_{EF} and accordingly it will not be considered as a feature to form the pattern vector.

Clustering is an unsupervised technique used in discovering inherent structure present in the set of objects [1]. Let the set of patterns be $S = \{x_1, x_2, ..., x_n\} \subseteq \Re^m$, where x_i is the *i*-th pattern vector corresponding to *i*-th document, *n* is the total number of documents in a given set of text documents and *m* is the dimensionality of the feature space. Note that the value of *m* for a given set of texts is determined in Step 3 as stated above. Let the number of clusters be *K*. The value of *K* may or may not be known *a priori*. In the present work, the value of *K* is computed automatically by the proposed method. If the clusters are represented by C_1, C_2, ..., C_K then we assume:

1. $C_i \neq \varnothing$ for $i = 1, 2, ..., K$
2. $C_i \cap C_j = \varnothing$ for $i \neq j$ and
3. $\cup_{i=1}^{K} C_i = S$ where $\varnothing$ represents null set.

The next section describes the proposed method which uses Kohonen's Self Organizing Network to detect the groups present in a given set of text documents. Then the groups are labeled on the basis of frequency of the class titles found in the documents of each such group.

3 The Proposed Method

We have used Kohonen's self-organizing network with *m* input nodes (since *m* is the number of features in the pattern vector) and 16 output ($p = 16$) nodes being arranged in a two-dimensional 4 x 4 grid. After the algorithm (to produce self organizing feature map) has converged, the locations of weight vectors in feature space almost correspond to mean vectors of *p* possible groups of the given data. In other words, after convergence, each node of the Kohonen's self organizing network represents a local best representative (seed) point associated with a particular high density region of the feature space.

The number of seed points p and their location in feature space are obtained by using Kohonen's self-organizing network. Then the data set is divided into p groups using the standard minimum squared Euclidean distance classifier concept. Later the groups are merged using a threshold h_n on the minimum interpoint distance between the groups. Here the threshold h_n for the cluster separation is taken to be equal to $h_n = l_n / n$, where l_n is the sum of the edge weights (edge weight is taken to be the Euclidean distance) of minimal spanning tree of S. The process of merging decreases the number of groups in the data set. The process terminates when no further merging is required and we get the desired clustering.

Then we compute the frequency of the possible class titles such as cricket, football, chess, athletics etc. in each of the groups obtained by the above algorithms. Then each group is labeled a specific class title which has the largest frequency in that particular group. For instance, a group is labeled as "cricket" if the number of occurrence of the word "cricket" in all the documents of that group is the largest than that found for other groups. Note that text classification should not be done on the basis of the frequency of class titles only, because a document of a specific class may not contain its class title at all. However, when we have a group of documents of a specific class, the group can be labeled on the basis of largest frequency of all the possible class titles.

4 Experimental Results and Conclusion

We have considered text documents from various faculties of games and sports such as cricket, football, hockey, basketball, swimming, lawn tennis, chess and atheletics. We have used three sets of data for experimentation. Each data set consists of a number of documents or articles published in three leading English daily newspaper of our country, namely *The Telegraph*, *The Times of India* and *The Statesman*. All these articles belong to any one of the said eight faculties of games. Each data set is constructed with unequal number of documents from different classes. This is done to incorporate variability in the size of the clusters. The total number of articles for each of the three data sets is also taken in the increasing order so that we can detect the effect of the size of the data on the performance of the proposed method.

Experiment 1: The data consists with 400 articles, where 115, 110, 95 and 80 articles are taken from the classes cricket, football, basketball and hockey athletics respectively. Here the proposed method has provided a success rate of 94.2 %.

Experiment 2: Here 500 news articles, out of which 128, 110, 90,98 and 74 articles are taken from cricket, football, chess, athletics and swimming respectively. In this case, the success rate provided by the proposed method is 96.8 %.

Experiment 3: The data consists with 600 articles where 132, 130, 82, 72, 94 and 90 articles are taken from cricket, football, lawn tennis, hockey, basketball and athletics respectively. The success rate of the proposed method in this experiment is 98.3%.

It is observed that the performance of the proposed method improves as the size of the given data set increases. It seems that the mentioned method would be able to provide a success rate of nearly 100 % when the size of the data set is very large.

Table 1. Results of experiments by the proposed method

Expt. No.	Size of data	No. of groups in the data	Performance of the Proposed Method		
			No. of groups detected by the proposed method	No. of correct classification	% of correct classification
1	400	4	4	377	94.2
2	500	5	5	484	96.8
3	600	6	6	590	98.3

References

1. M. R. Anderberg, *Cluster Analysis for Application*, Academic Press, Inc, New York, 1973.
2. Michael W. Berry, *Survey of Text Mining: Clustering, Classification, and Retrieval*, Amazon, 2003.
3. A. Gelbukh, G. Sidorov, A. Guzman-Arenas. Use of a weighted topic hierarchy for text retrieval and classification. *Lecture Notes in Artificial Intelligence*, N 1692, Springer, 1999.
4. P. G. J. Lisboa, *Neural Networks: Current applications* . Chapman and Hall, London, 1992.