ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 2, March 2016

# A Novel Approach for Support Vector Machine in Text Classification using Dimensionality Reduction

C.MAHALAKSHMI[#1],

[#] Assistant Professor, Department of Computer Science and Engineering,
Bharathiyar Institute of Engineering for Women,
Deviyakurichi(po),Attur(tk),Salem, Tamilnadu, India.
[1] maha.it2008@gmail.com

***Abstract* - Feature clustering is the way to reduce the dimensionality of features presents in the text documents and it is highly important for text categorization problems. The performance of the text classification is degraded when the dimensionality of input text is huge .Feature clustering is a powerful alternative to feature reduction approaches. The first task is to calculate the word patterns for each feature present in the text document. The second task is to calculate the membership function by the mean and deviation of the word patterns .The third one is to generate the clusters based on the membership function. Evaluation results for these tasks show that the proposed methodology obtains reliable performance for text classification tasks.**

*Keywords* - **Clustering, Text Classification, Dimensionality Reduction, Membership Function, Feature Reduction.**

## I. INTRODUCTION

Given a set of documents and their associated class labels, text classification is the problem of finding the true class label of a new document. There are several algorithms used for text classification, ranging from the simple but effective Naïve Bayes algorithm to the more computationally demanding Support Vector Machines. A common characteristic of text data is its extremely high dimensionality. A standard procedure to reduce feature dimensionality is feature selection and feature extraction. An alternative approach is to reduce feature dimensionality by grouping "similar" words into a much smaller number of word-clusters, and use these clusters as features. In the first stage clusters that capture the information about the set of documents are extracted as features, and in the second stage these features are used for clustering the documents in an unsupervised manner. The results clearly showed that this representation of the documents, significantly improved the accuracy of unsupervised document classification. In general, feature extraction approaches are more effective than feature selection techniques, but are more computationally expensive. Feature clustering is a powerful method to reduce the dimensionality of feature vectors for text classification.

Feature clustering algorithm is an incremental, self-constructing learning approach. Word patterns are considered one by one. The user does not need to have any idea about the number of clusters in advance. No clusters exist at the beginning, and clusters can be created if necessary. For each word pattern, similarity of this word pattern to each existing cluster is calculated to decide whether it is combined into an existing cluster or a new cluster is created. Existing feature clustering methods are referred as hard clustering, the words are exactly belongs to a particular cluster. But the proposed algorithm is referred as soft clustering since the features have the similarity to more than one cluster.

Once a new cluster is created, the corresponding membership function should be initialized. The membership function is calculated using a measure called fuzzy similarity measure with mean and deviation. Evaluation results for these tasks show that this fuzzy clustering obtains reliable performance for text classification tasks.

## II.TEXT PREPROCESSING

Text preprocessing is the basic step needed to prepare the text for data mining process. Preprocessing is to remove the stop words and stem words present in the text document.

### *A. Prop Bank Generation*

The prop bank generation is used to generate proverbs in the database. In prop bank generation text preprocessing can be done. This process contains separating the sentences, label the terms.

### *B. Stop Bank Generation*

It is a two step process where the stop and stem words removed to obtain the concepts of a text document.

#### *1) Stop Words Removal:*

The words like *and, in, to, the, is* ...are called stop words. These words should be removed.

#### *2) Stemming Process:*

The stemming process is to reduce the word into the root form of the word. Eg: dropping, dropped are reduced into the root form drop.

## III.CALCULATE THE WORD PATTERNS

The word patterns are calculated by using the frequency of a certain feature in the text document and the class label information. In this way, more significant patterns will be fed in first and likely become the core of the underlying cluster.

$$xi=<x_{i1},\ x_{i2}\ ,\dots,\ x_{in}>$$

$$=<P(c1|wi),P(c2|wi),\dots\dots,P(cp|wi)>$$

Where,

$$P(cj|wi)=\frac{\sum dqi*\delta qj}{\sum dqi} \qquad (1)$$

For 1 < j <p. Note that $d_{qi}$ indicates the number of occurrences of $w_i$ in the document.

Also, δqj is defined as

$$\delta_{qj}=\begin{cases}1, document\ belongs\ to\ a\ class\\ 0, otherwise\end{cases}$$

## IV. CALCULATING THE MEMBERSHIP FUNCTION

The mean, deviation of a word pattern is calculated as follows,

$$m_i=\frac{\sum_{j=1}^{q} xji}{|G|} \qquad (2)$$

$$\sigma i=\frac{\sqrt{\sum(x-m)}}{\sqrt{|G|}} \qquad (3)$$

The fuzzy similarity of a word pattern <x1………….. up> to cluster G is defined by the following membership function,

$$\mu_G(x)=\prod_{i=1}^{p} exp[-((xi-mi))/\sigma i)^\wedge 2] \qquad (4)$$

## V. FEATURE CLUSTERING

Word patterns are considered one by one. No clusters exist at the beginning, and clusters can be created if necessary. For each word pattern, the similarity of this word pattern to each existing cluster is calculated to decide whether it is combined into an existing cluster or a new cluster is created. When the word pattern is combined into an existing cluster, the membership function of that cluster should be updated accordingly. Feature clustering methods are generally hard clustering where each word of the original features belongs to exactly one word cluster. Fuzzy clustering are defined as soft clustering where each feature has more than one similarity.

## VI. SVM BASED TEXT CLASSIFICATION

Assume that k clusters are obtained for the words in the feature vector W. Then we find the weighting matrix T and convert D to $\mathbf{D'}$ by Using $\mathbf{D'}$ as training data, a classifier based on support vector machines (SVM) is built. Note that any classifying technique other than SVM can be applied. SVM is better than other methods for text categorization. To make the method more flexible and robust, some patterns need not be correctly classified by the hyper plane, but the misclassified patterns should be penalized. d is an unknown document. We first convert d to d' by,

$$D'=DT \qquad (5)$$

Then we feed d' to the classifier. We get p values, one from each SVM. Then d belongs to those classes with 1, appearing at the outputs of their corresponding SVMs. A support vector machine (SVM) is a concept in computer science for a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. The standard SVM takes a set of input data and predicts, for each given input, which of two possible classes the input is a member of, which makes the SVM a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

## VII. PERFORMANCE COMPARISON

To compare the classification of each method, adopt the performance measures in terms of microaveraged precision (MicroP), microaveraged recall (MicroR), microaveraged F1 (MicroF1), and microaveraged accuracy (MicroAcc). Data mining tool kits are used to perform the classification. Performance Comparison is done on text classification with feature reduction and text classification without feature reduction.

$$\text{MicroP}=\frac{\sum_{i=1}^{p} TPi}{\sum_{i=1}^{p}(TPi+FPi)} \qquad (6)$$

$$\text{MicroR}=\frac{\sum_{i=1}^{p} TPi}{\sum_{i=1}^{p}(TPi+FNi)} \qquad (7)$$

$$\text{MicroF1}=\frac{2MicroP*MicroR}{MicroP+MicroR} \qquad (8)$$

$$\text{MicroAcc}=\frac{\sum_{i=1}^{p}(TPi+TNi)}{\sum_{i=1}^{p}(TPi+FPi+TNi+FNi)} \qquad (9)$$

## VIII. EXPERIMENT

In the preprocessing step, can eliminate the stop words and stem words to exploit the text concepts. In fig.1 first remove the stop words generated in the input document.

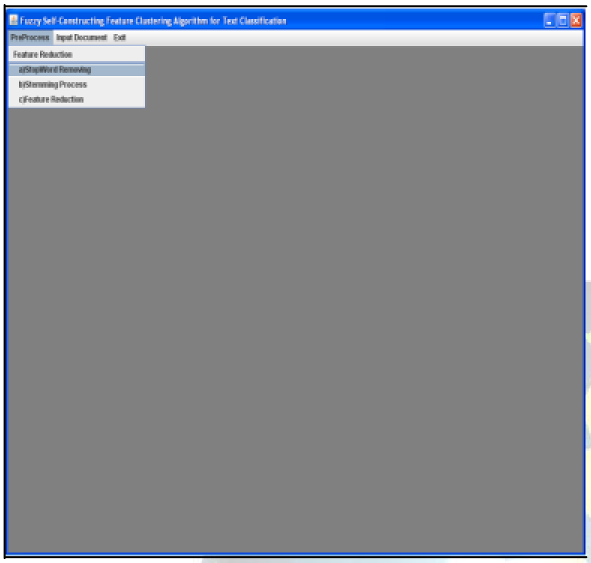
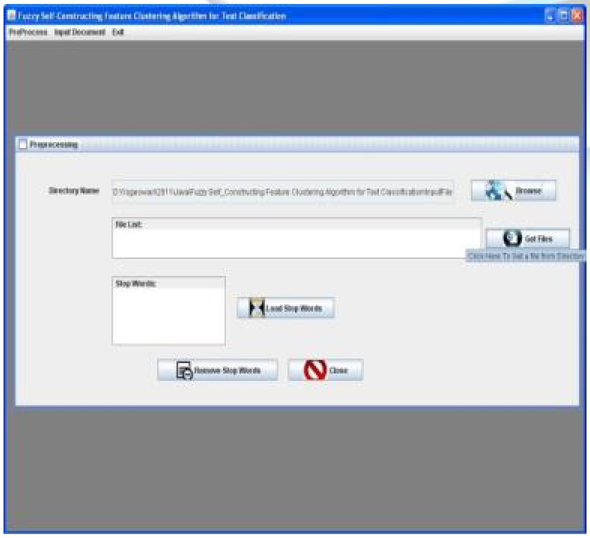


Fig.1 Text Preprocessing

In Fig.2 load the file in which text preprocessing is done and add the stop words to be removed. Then the file is subjected to stemming process and the input document is used to calculate word patterns. The word patterns are used for clustering. Stop words and stem words could be added in a file in order to remove those from the given text.



Directory Name D:\Yogeswari0281\Java\Fuzzy Self_Constructing Feature Clustering Algorithm for Text Classification\inputFile

Fig.2 Selecting a file for preprocessing and remove stop words

## IX. CONCLUSION

The tasks presented here are the concept extraction, clustering, selecting a model for text classification. When a document set is transformed to a collection of word patterns, the relevance among word patterns can be measured, and the word patterns can be grouped by applying similarity based clustering algorithm. First use the soft clustering approaches to cluster among relevant features. Then use the data mining tool or else by use an algorithm for text classification with performance measures. This novel approach minimizes the trial-and-error for finding the appropriate number of features. In this clusters are formed instantly and there is no need to specify the number of clusters in advance. Evaluation on various text documents show the proposed work gives good performance over the existing works. This soft clustering approach is also used in segmenting the images, web mining tasks.

## REFERENCES

[1] Jung-yi Jiang and Ren-Jia Liou and Shie-Jue Lee (2011) 'A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification', IEEE Transactions on Knowledge and Data Engineering,vol 23.

[2] Dhillon, I.S and Mallela, S. and Kumar, R.(2003) 'A Divisive Information-Theoretic Feature Clustering Algorithm for Text Classification', Machine Learning Research, vol. 3, pp. 1265 - 1287.

[3] Baker , L.D. and McCallum, A. (1998) 'Distributional Clustering of Words for Text Classification', Proc. ACM SIGIR, pp. 96-103.

[4] Joachims, T (1998) 'Text Categorization with Support Vector Machine: Learning with Many Relevant Features', Technical Report LS-8- 23, Univ. of Dortmund.

[5] Slonim, N.and Tishby, N. (2001)'The Power of Word Clusters for Text Classification', Proc. 23rd European Information Retrieval Research (ECIR).

[6] Li,H. and Jiang, T and Zang, K. (2004) 'Efficient and Robust Feature Extraction by Maximum Margin Criterion', Advances in Neural Information Processing System, pp. 97-104, Springer.

[7] " TextClassification"en.wikipdia.org/wiki/text classification.