Machine Learning Approach to Document Classification using Concept based Features

C.Saranya Jothi
Department of Computer Science and Engineering
SSN College of Engineering
Chennai, India

D.Thenmozhi
Department of Computer Science and Engineering
SSN College of Engineering
Chennai, India

ABSTRACT
Text mining refers to the process of deriving high-quality information from text. Text processing involves in search and replace in electronic format of text. A number of approaches have been developed to represent and classify text documents. Most of the approach tries to attain good classification performance while taking a document only by words. We propose a concept based methodology instead of terms. It represents the meaning of text to reduce the features. Support Vector Machine (SVM) algorithm is applied for document classification. Then the performance measure is compared with document classification using original features and concept based features. This methodology enhances the document classification accuracy.

Keywords
Text classification, Support Vector Machine, Feature Selection.

1. INTRODUCTION
Text mining also referred to as data mining. It contains the process of structuring the input as a text. It derives high-quality information from text. High quality information refers to word integrity. Text mining tasks include text classification and clustering.

In document classification bag-of-words model is commonly used. Each document represented as a vector. The vector can accept only a numerical value. The value indicate a frequency of occurrence in the document, each word represent as a feature. Generally, the dimensionality of a document space is scarce, i.e., most of the values in the vector are zero. In this case, this high dimensionality of feature space is a major problem in Text categorization (TC).

Text categorization is the task of assigning predefined categories to a text documents. It can provide theoretical views of document collections and has important applications in the real world. Feature selection for TC helps in reducing dimensionality of feature space and to improve the classification effectiveness. Feature selection is the process of selecting a subset of relevant features in the document. In this paper, we use concept based feature selection method to reduce the high dimensional space to fewer spaces. It represents the meaning of text in the document. For classification accuracy we use this method. The document is classified using SVM classifier. Then the performance measure is compared using original features and reduced features.

The rest of the paper is organized as follows. Section 2 discusses related work that describes the various classification techniques. Section 3 deals with the system description of the proposed system. Experimental results are discussed in Section 4 and Section 5 concludes the paper.

2. RELATED WORKS
The work accomplish by other analyzer that are related to feature extraction and different classification techniques presented here. Each technique have been attempted to prove the effectiveness of the approach.

2.1 Feature Extraction
Lin et. al introduced the SMTP similarity measure, to measure the similarity between two document sets with respect to a feature. The effectiveness of the measure is evaluated on several real-world data sets for text classification and clustering problems [3]. For classify a classes they used k-NN classifier with different measure. They proved that similarity between two sets of documents is symmetric measure. The results have shown that the performance is better than other measures. But the dimensional space is sparse, so time taken for each and every process is huge. Peng et. al proposed an Chinese text processing based on concept similarity [4]. It is based on concept similarity between words or sentences. They apply the algorithm for text classification using web news dataset. In this research the result is better than k-NN method's based on vector space model. Basu et. al proposed an supervised feature selection approach, it develops a similarity between a term and a class [1] . In this they use score based on their similarity with all the classes and then all the terms will be ranked accordingly. The observation results are presented on several TREC and Reuter data sets using k-NN classifier.

In feature selection, similarity between a terms and class are used, the dimensionality of the feature vector is very large. Our work presented here is concept based feature extraction. It represents the meaning of the texts in a high dimensional space to fewer dimensional spaces.

2.2 Classification
Gayathri K. and Marimuthu A. have used k-NN and SVM classifiers on a Reuters-21578 dataset, using document classification based on their contents [2]. They found that both k-NN classifier and SVM classifier have been widely implemented in many real world applications. In this, they prove that well-performing k-NN classifier may suffer from less accurate than the SVM classifier. Wang et. al have applied optimal SVM algorithm for text classification, to

determine a given document belongs to which of the predefined categories [5]. SVM is a powerful supervised learning model. A large number of techniques have been developed for text classification, including Naive Bayes, Nearest Neighbor, neural networks, regression, rule induction, and Support Vector Machines. Among them SVM has been recognized as one of the most effective text classification methods.

Among all the classifiers, SVM is identified to be the best classifier. The time taken to proceed for each process is less while compare to other classifier.

3. PROPOSED FRAMEWORK

A concept based feature selection method for text classification will be discussed here to improve the performance of the classifier. The system architecture is described in the Figure 1. The proposed model is to acquire highly consistent result by applying a classification algorithm.

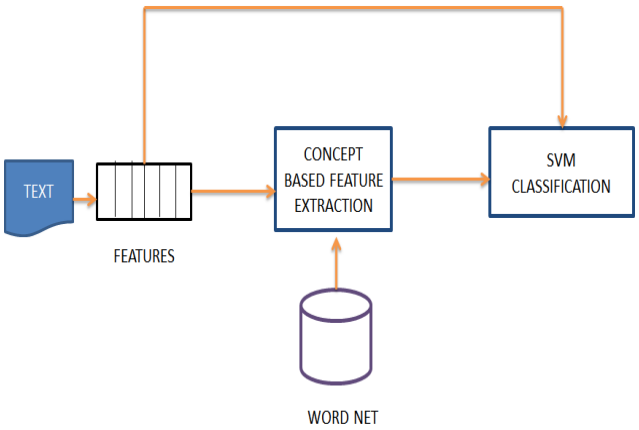

Figure 1: System Architecture

In this architecture, the input is given as a form of text which is a collection of webpages. Here the text data is a structured one and the web pages are collected by the world wide knowledge base. In the input data, each and every line indicates a document. From the documents, features where extracted. Concept based methodology is used to reduce the dimensionality of the feature space using these extracted features. SVM algorithm is applied for document classification. Finally compare the performance with original features and concept based features for classification accuracy.

3.1 Feature Extraction

Feature selection is usually employed for reducing the size of the feature space. After removing the stop words, unique features are extracted. The extracted features are selected for categories a document. Then the document is represented as a vector. Each attributes indicates the value of the corresponding feature in the document. The feature value is a term frequency. In-order to reduce the dimensionality of the feature space concept based method is proposed.

Example:

After remove all the stop words unique features are extracted. Assume that we use word count as feature values.

d1:

Health is good for children.

Youngster will play with kid

Kid is well.

d2:

Apple is good for health

Wellness for good

We consider 9 features which are,

Apple, children, good, health, kid, play, well, wellness, youngster.

Feature Vector:

Apple children good health kid play well wellness youngster

0 1 1 1 1 1 1 0 1

d1= (0,1,1,0,1,1,1,0,1)

d2= (1,0,2,1,0,0,0,1,0)

Therefore original feature vector is like this, some of the features are semantically same in the text document.

3.2 Concept Based Feature Extraction

It represents the meaning of texts in a high-dimensional space of concepts derived from Word Net [8]. By using the concepts, the transformation of the data in the high-dimensional space is reduced to fewer dimensions. The example for concept based feature extraction is shown in Figure 2.

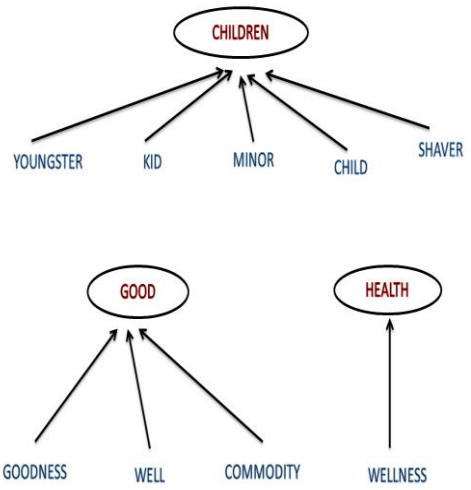

Figure 2: Concept Based Feature Extraction example

Original Feature Vector:

Apple children good health kid play well wellness Youngster

0 1 1 1 1 1 1 0 1

Reduced Feature Vector:

Apple	children	good	health	play	wellness
0	3	2	0	1	0

d1=(0,3,2,0,1,0)

Here the dimensionality of the vector space is reduced from 9 columns to 6 columns by using the concepts.

4. IMPLEMENTATION

The publicly available dataset of the WebKB dataset is obtained from [6], used to evaluate our methodology of document classification. The dataset description is given in Table 1. The documents in the WebKB data set are webpages collected by the World Wide Knowledge Base. The documents are manually classified into four classes. The number of features involved is 7786.

Table 1: Data set

Class	Training documents	Testing documents	Subtotal of documents
Project	336	168	504
Course	620	310	930
Faculty	750	374	1124
Student	1097	544	1641
Total	2803	1396	4199

Through the dataset, features will be extracted from each document. Each word in the document represents a feature. Feature is a unique word. Feature value indicates the word occurrence in the document. Each document represents as a vector. Each word indicates the value of the corresponding feature in the document. The feature vector will be formed as a dimensionality of the feature space.

4.1 Text classification with SVM

Support Vector Machine is a powerful supervised learning models that analyze data used for classification. It is the maximum margin classifier. SVMs can only do binary classification. It controls complexity and overfitting issued, so it works well on a wide range of practical problems. It works very fast in training and testing. So that it become an achievable option for large data set. Large margin gives better generalization ability and less over-fitting. It classify a correct classes based on class labels. Multi class labels are shown in the figure 3

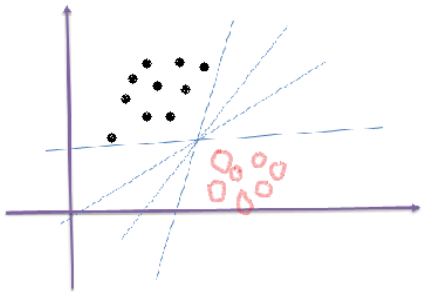

Figure 3: Multi-class classification using SVM

4.1.1 *Data Pre-processing*

The publicly available datasets of the WebKB datasets are downloaded and convert in to .ARFF (Attribute-Relation File Format) files [7]. The file is loaded into the WEKA data mining tool after which a verification of data correctness is done.

4.1.2 Train and Test a Classifier with Cross Validation

Cross validation is commonly used method to tune a classifier. It avoids overlapping test sets. Data is split into k subsets of equal size. Sets number of folds for cross-validation. It takes 10% for testing data and the remainder for training. Train the data using SVM algorithm. After finding the parameters, now apply the same classifier on the test set.

5. EXPERIMENTAL RESULTS

The comparative performance measures in terms of classification accuracy, of the Original data sets and reduced data sets are tabulated in Table 2

Table 2: Result Table

	Accuracy by Original data using SVM	Accuracy by Reduced data using SVM
Training and Test set	0.611	0.644
Cross-Validation	0.681	0.767

The classification accuracy is shown in True Positive (TP) rate (instances correctly classified as a given class) and False Positive (FP) rate (instances falsely classified as a given class). A confusion matrix contains information about actual and predicted classifications done by a classification system. The accuracy is taken in to true positive value. The result table shows, that the accuracy of reduced data using SVM is better than accuracy of original data.

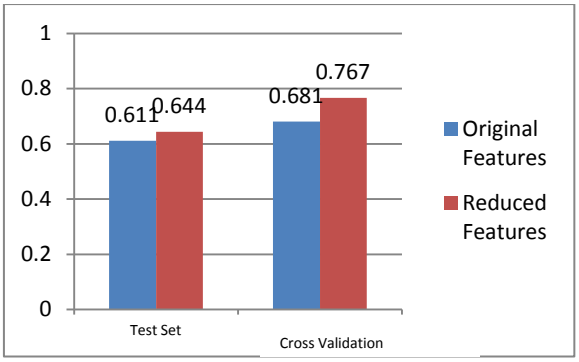

Figure 4: Analysis

In the analysis, the cross validation using original data and reduced data is far better than test set is shown in the figure 4. The time taken for reduced data is less than original data.

6. CONCLUSION

The proposed work for feature selection in which, concept based features are extracted from original features by using Word Net. The high dimensional space is reduced to fewer spaces by improving the classification accuracy. Support Vector Machine (SVM) algorithm is applied for document classification. Finally, compare the performance with original features and reduced features. The results have shown that the performance obtained by the reduced data is better than that original data. To improve the classification efficiency, use SMTP similarity measure [3] and try different classification approach in further enhancement.

7. REFERENCES

[1] Basu T. and Murthy C. (2012). Effective text classification by a supervised feature selection approach. In Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on, pages 918–925. IEEE.

[2] Gayathri K. and Marimuthu A. (2013). Text document pre-processing with the knn for classification using the svm. In Intelligent Systems and Control (ISCO), 2013 7th International Conference on, pages 453–457. IEEE.

[3] Lin Y.S., Jiang J.Y., and Lee S.J. (2013). A similarity measure for text classification and clustering. IEEE Transactions on Knowledge and Data Engineering, page 1.

[4] Peng J., Yang D.q., Tang S.W., Gao J., Zhang P.y., and Fu Y. (2007). A concept similarity based text classification algorithm. In Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery-Volume 01, pages 535–539. IEEE Computer Society.

[5] Wang Z.Q., Sun X., Zhang D.X., and Li X. (2006). An optimal svm based text classification algorithm. In 2005 International Conference on Machine Learning and Cybernetics, pages 1378–1381.

[6] Datasets for single-label text categorization: http://web.ist.utl.pt/~acardoso/datasets/

[7] WEKA, classpath: http://weka.wikispaces.com/classpath

[8] WordNet 2.1. http://www.brothersoft.com/wordnet-236667.html.