Text Classifiers By Support Vector Machine And RBF Kernal

Meenakshi Sharma
S.S.C.E.T, Badhani
Pathankot, India.
Sharma.minaxi@gmail.com

ShivangiSharma,
S.S.C.E.T,Badhani
Pathankot,India.
Shivangisharma15391@gmail.com

***ABSTRACT*:One of the major topic which support text mining is text representation i.e,it search for the appropriate terms to transfer documents in to numeral vectors.at present there are many efforts have been invested on this topic to magnify the text representation using the vector space model (VSM) which helps in improving the performances of text mining techniques such as text classification and text clustering..one term is used for text classification is text categorization(TC),which helps in automatically classifying a set of text documents into different classes from a already defined set.if there is a document which belongs to exactly one class i.e,called single-label classification task ,otherwise it is a multilabel classification task .text categorigation uses some tools of machine learning. In this paper our main concern is on improving the performance of text mining with the help of support vector machine (SVM) with RBF kernel.**

General Terms:Text Mining,support vector machine(SVM),RBF kernel.

I. INTRODUCTION

Text mining is also recognized as data mining, refers to the procedure of deriving high worth information from text.The principle of data mining is progression of raw and unstructured information; take out meaningful information from text. It generally involves the method of structuring the contribution text, deriving patterns contained by the structured data, and in conclusion evaluation and analysis of the output.

In text mining process, initializing with the gathering of documents, a tool extract the particular information or document and preprocess it. Then it comes to the next phase i.e. text analysis phase, where sometimes processes used is repeated until some useful information is extracted. Text mining and data mining are similar except the data mining tools which organize the structured data from the storage. On the other hand text mining extract information from semi-structured and structured datasets such as HTML file etc. text mining is better way to organize online data for organizations. Text Mining tasks consist of:Text clustering, Concept/Entity Extraction, Text Summarization,Text classification[1][2].text classification

Text classification is always an important research topic. .Nowdays,we have to deal with very large amount of text document. In text classification topic includes are text classification and text genre-based classification . text can be classified in any form. It can be in the form of a scientific articles, newsreports, reviews and advertisements. Genre of the text dependent on way the text was edited ,the language in which it is edited &also the way sin which it is edited .In this most of data is collected from the web ,through newspapers, through articles and broadcasted news.

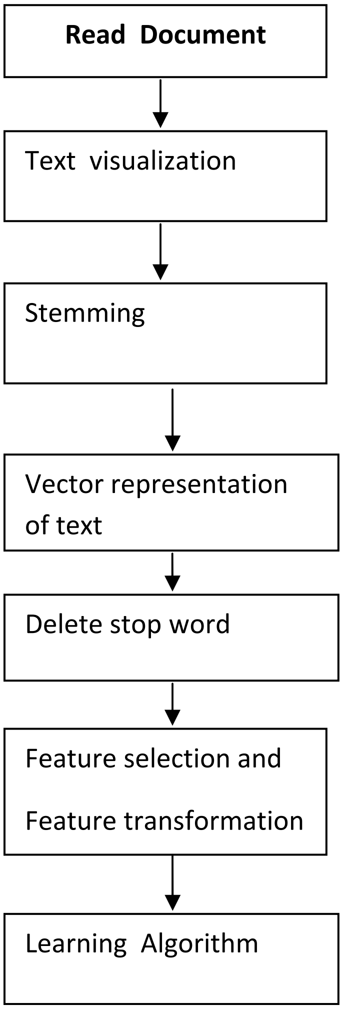
Read Document
Text visualization
Stemming
Vector representation of text
Delete stop word
Feature selection and Feature transformation
Learning Algorithm

Figure 1: Text classification process

All the data which we are using for the genre classifications are multisource and they have different formats, have different vocabularies and also the different writing styles for the documents of different genre,but the data is different.

Text classification is the way in which we can arrange a document under already defined or predefined category .it means we have different categories for documents. For every machine learning task , a dataset is needed . we can assign more than one category to a document[3].

II. METHODOLOGY

(i) Text Classifier Methods:

Support Vector Machine (SVM): SVM Method:- A Support Vector Machine is prejudic classifier formally defined by a separating hyper plane. For a linear independent set of 2D-points which fit in to one of two classes, find a separate straight line.

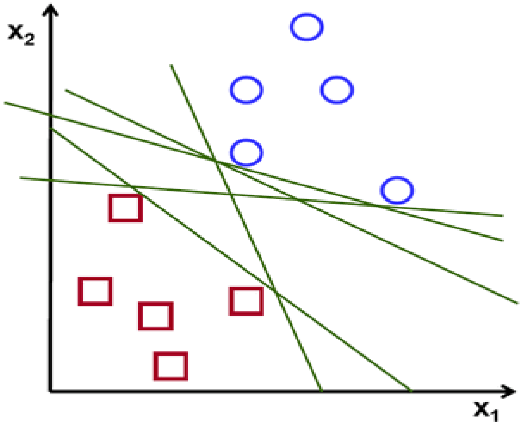

Figure 2:Hyper-plane separating two classes

Above given figure shows that there exist multiple lines that given a solution to a problem. A line is said to be bad if it is passed too close through the points due to the noise sensitivity. Therefore, it is very necessary to find the line that should passed through points as far as possible.

Then, the methodology of the support vector machine algorithm is based on finding the hyper plane that offer the smallest distance to the training examples. Again, the smallest distance receives the significant name of boundary within SVM's theory.

Therefore, the optimal solution for separation of hyper plane increases the boundary values of the training data.

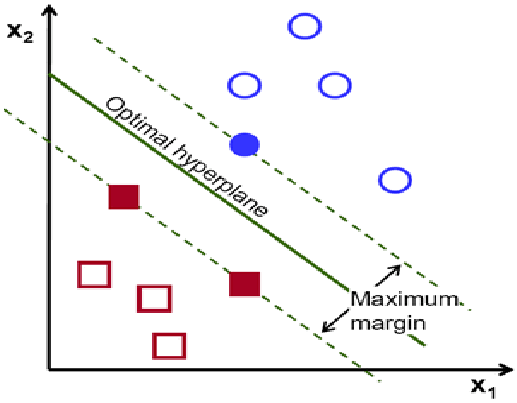

Figure 3: Optimal Hyper plane

How is the optimal hyper plane computed for SVM?

The given notation is used to describe a hyper plane:

$$f(x) = \beta_0 + \beta^T x,$$

Where β = *Weight vector* & β_0 = Bias

The numerator is equal to one and the distance to the support vectors, for the canonical hyper plane, is

$$distance_{support\ vector} = \frac{|\beta_0 + \beta^T x|}{||\beta||} = \frac{1}{||\beta||}$$

Remember that the boundary values introduced in the above figure, is represented as M here, and is twice the distance to the nearby examples:

$$M = \frac{2}{||\beta||}$$

Finally, the problem of maximizing M is equivalent to the problem of minimizing a function $L(\beta)$ subject to some constraints. The constraints model the requirement for the hyper plane to classify correctly all the training examples. Formally,

$$distance_{support\ vector} = \frac{|\beta_0 + \beta^T x|}{||\beta||} = \frac{1}{||\beta||}$$

The advantages of support vector machines are:

- Efficient in high and elevated dimensional spaces.
    - SVM is also known as memory effective as it uses the subset of training data points in support vectors (also called decision function).
    - Support Vector Machine is effective and efficient in cases where the no of samples is smaller than the no of dimensions.
    - Versatile: Different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels

The disadvantages of support vector machines include:

- Support vector machine doesn't perform well if the no of features are too much than the no of samples.
- Support vector machine are calculated using cross validation and it doesn't directly provide probability estimates.

(ii) K-mean Classifier:

K-mean clustering is a method of vector quantization for cluster analysis in data mining. K-mean clustering aims to partition n observations into k clusters in which each partition belongs to the cluster with the nearest mean, serving as a prototype of the cluster. K-means clustering tends to find clusters of comparable degree, while the expectation-maximization mechanism allows clusters to have different shapes.

The algorithm should not be confused with k-nearest neighbor(KNN), another popular machine learning technique. Given a set of observations $(x_1, x_2 \dots, x_n)$, where each observation is a *d*-dimensional real vector, k-means clustering aims to partition the *n* observations into *k* (≤ *n*) sets $S = \{S_1, S_2 \dots, S_n\}$ so as to minimize the within-cluster sum of squares (WCSS). In other words, its objective is to find

$$\arg\min \sum_{i=1}^{k} \sum_{x \in Si} \| k - \mu i \|$$

where μi is the mean of points in S

The aim of K Means is to partition the objects in such a way that the intra cluster similarity is high but inter cluster similarity is comparatively low. A set of n objects are classified into k clusters by accepting the input parameter k. All the data must beavailable in advance for the classification.

III.RESULT

(i)Precision: Precision is defined as the fraction of retrieved documents that are relevant to the find.

$$\text{Precision} = \frac{\{\text{relevant cocument}\}\cap\{\text{retrieved documents}\}\,|}{|\,\{\text{retrieved document}\}\,|}$$

Precision takes all retrieved documents into account, but it can also be evaluated at a given cut-off rank, considering only the topmost results returned by the system. This measure is called precision

(ii)Recall:

Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$\text{Precision} = \frac{\{\text{relevant cocument}\}\cap\{\text{retrieved documents}\}\,|}{|\,\{\text{relevant document}\}\,|}$$

For example for text search on a set of documents recall is the number of correct results divided by the number of results that should have been returned.

(iii)Accuracy:

in the fields of science ,engineering, industry and statistics**,** the accuracy of a measurement system is the degree of closeness of measurements of a quantity to that quantity's actual (true) value. In our thesis work 6 different algorithms are introduced in class to implement spam Detection in Python, and their performance is compared. The algorithms we used were:

- SVM (Support Vector Machine)
- K mean classifiers

False positive ratio is interesting because detection out a ham (non-spam) message is a bad thing, worse than letting a spam message get through creation of training set and test set. We wrote a Python script to process these messages and create a feature vector out of each message. In the sequel it is described how the feature vectors look like. The script divides the feature vectors into training set and test set, while preserving the ham-spam ratio in each set. Actually, the script randomly creates 90 different pairs of training set and test set as follows:

- We used 9 different "training fractions", i.e. the percentage of training set size out of the entire dataset.
- The training fractions we used were: 0.1, 0.2... 0.9.

- For each training fraction, we randomly created 10 different pairs of training set and test set, so we can examine the performance as an average of 10 runs.

In the following tables the various parameters for various machine learning algorithms has been shown. Here all results are with training fraction of 0.5 except for SVM for which the training fraction is 0.4 (which was the maximum we could test) .

We see that the best two algorithms are K-mean classifier and SVM. Accuracy, Precision, Recall of different classifiers are given in table for different machine learning algorithm.

.

Table1:- showing results for k-mean classifier

Training set	precision	Recall	Accuracy
0.60	90.75	91.25	95.25
0.75	91	91.75	95.75
0.85	91.5	91.25	96
0.90	91.25	91.5	95.5
0.95	90	90.25	95.5

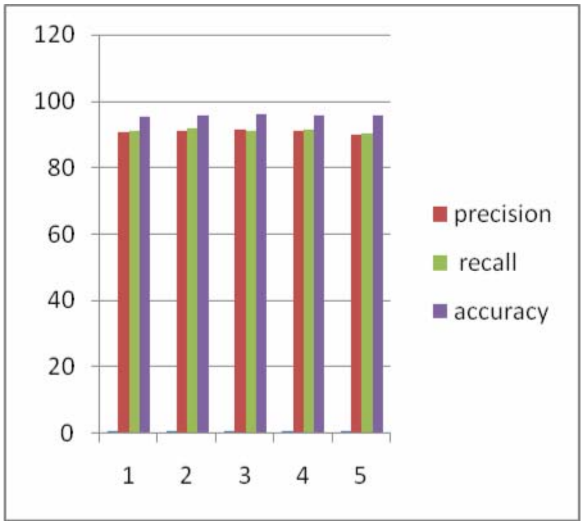

Figure 4: Result analysis graph for precision, recall, accuracy by K-mean classifier.

Table 2:-Showing results for SVM classifier .

Training data	precision	Recall	Accuracy
0.60	87.75	90.75	93.5
0.75	90.5	93	95
0.85	93.5	94.75	96.5
0.90	95	95.75	97.5
0.95	97.5	97.75	99

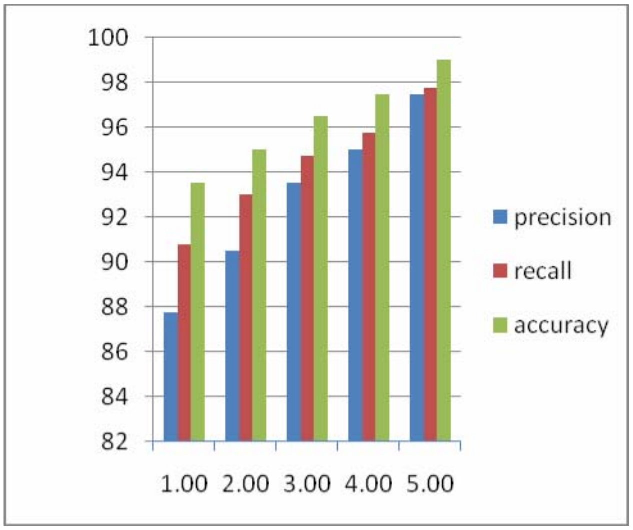

Figure 5: Result analysis graph for prcession , racall, accuracy of SVM

IV. CONCLUSION

This paper shows the best result for best two algorithms SVM and k-mean classifier. Considering the low false positive ratio K-mean classifier performs well as it is easier to implement and has low running time but has less accuracy than Linear SVC and k-mean. Hence we conclude that optimization methods perform well and show better results than other classifiers.We enhance this work by using Graphical model like conditional random field, hidden markov field which shows the dependency between the features in text classification. We can also use Kernal function for reducing the processing time and error of overlapping information in text classification.

References:-

1. Zhang, Wen, Taketoshi Yoshida, and Xijin Tang. "A comparative study of TF* IDF, LSI and multi-words for text classification." *Expert Systems with Applications* 38.3 (2011): 2758-2765.

2. Bijalwan, Vishwanath, et al. "KNN based machine learning approach for text and document mining." *International Journal of Database Theory and Application* 7.1 (2014): 61-70.

3. Ikonomakis, M., S. Kotsiantis, and V. Tampakas. "Text classification using machine learning techniques." *WSEAS Transactions on Computers* 4.8 (2005): 966-974.

4. Pilászy, István. "Text categorization and support vector machines." *The proceedings of the 6th international symposium of Hungarian researchers on computational intelligence*. 2005.

5. Joachims, Thorsten. *Text categorization with support vector machines: Learning with many relevant features*. Springer Berlin Heidelberg, 1998.

6. Joachims, Thorsten. "Transductive inference for text classification using support vector machines." *ICML*. Vol. 99. 1999.

7. Lodhi, Huma, et al. "Text classification using string kernels." *The Journal of Machine Learning Research* 2 (2002): 419-444.

8. Chen, Mengen, Xiaoming Jin, and Dou Shen. "Short text classification improved by learning multi-granularity topics." *IJCAI*. 2011.

9. Basu, Tulika, and C. A. Murthy. "Effective text classification by a supervised feature selection

approach." *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*. IEEE, 2012.

10. Ko, Youngjoong. "A study of term weighting schemes using class information for text classification." *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2012.

11. Kiritchenko, Svetlana, and Stan Matwin. "Email classification with co-training." *Proceedings of the 2011 Conference of the Center for Advanced Studies on Collaborative Research*. IBM Corp., 2011.

12. Ektefa, Mohammadreza, et al. "Intrusion detection using data mining techniques." *Information Retrieval & Knowledge Management,(CAMP), 2010 International Conference on*. IEEE, 2010.

13. http://en.wikipedia.org/wiki/Text_Classification

14. Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. "Evaluation measures for ordinal regression." *Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference on*. IEEE, 2009.

15. Xu, Qian, et al. "Sms spam detection using noncontent features." *IEEE Intelligent Systems* 6 (2012): 44-51.

16. Mangalindan, Mylene. "For bulk E-mailer, pestering millions offers path to profit." *Wall Street Journal* 13 (2002).