A Hybrid Text Classification Approach Using KNN and SVM

[1]R.Vinoth, [2]Anjana Jayachandran, [3]M.Balaji, [4]R.Srinivasan,
[1,3,4]Assistant Professor, [2]PG Scholar,
[1,2,3]Department of Electronics and Communication Engineering,
[4]Department of Electrical and Electronics Engineering,
[1,3,4]Muthayammal College of Engineering, Rasipuram, Tamilnadu
[2]Paavai College of Engineering, Pachal, Tamilnadu
[1]rvinothrathinam@gmail.com, [2]ajs3143@gmail.com, [3]balajilink@gmail.com, [4]srinihbk@gmail.com

***Abstract*— Text classification is the process of assigning text documents based on certain categories. A classifier is used to define the appropriate class for each text document based on the input algorithm used for classification. Due to the emerging trends in the field of internet and computers ,billions of text data are processed at a given time and so there is a need for organizing these data to provide easy storage and accessing .Many text classification approaches were developed for effectively solving the problem of identifying and classifying these data .In this project a new text document classifier is proposed by integrating the nearest neighbor classification approach with the support vector machine(SVM) training algorithm. The proposed SVM-NN approach aims to reduce the impact of parameters in classification accuracy. In the training stage, the SVM is utilized to reduce the training samples for each of the available categories to their support vectors (SVs).The SVs from different categories are used as the training data of nearest neighbor classification algorithm in which the nearest centroid distance function is used to calculate the average distance instead of Euclidean function, which reduce time consumption.**

***Index Terms*—Machine Learning, Text document classification, Nearest Neighbor, Support Vector Machine, nearest centroid.**

I. INTRODUCTION

Data mining is the process of extracting hidden predictive information from large databases. Data mining is a tool that predicts future trends and behaviours. The scope of data mining can generate new business opportunities by providing these capabilities as automated prediction of trends and behaviours and automated discovery of previously unknown patterns. Text mining also referred to as text data mining and equivalent to text analysis. It is the process of deriving high-quality information from text. Analysis involves Information retrieval and pattern recognition. Text mining also referred to as text data mining, roughly equivalent to text analytics. Text analytics refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning .Text mining usually involves the process of structuring the input text, usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database, deriving patterns within the structure data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance.

One of the most effective binary classification techniques is the support vector machines (SVMs).It has been demonstrated that the method performs superbly in binary discriminative text .As one of the discriminative classification methods, SVM classification has been shown to be more accurate than other classification approaches. The proposed hybridized algorithm was under in binary and multiclass classification of data. The results were compared to those obtained by single SVM and KNN. In this study, linear kernel function is included in the SVM and HKNNSVM procedure, so the SVM, KNN and HKNNSVM are linear process. It has been demonstrated that the proposed method is a useful tool for classification and the classification performance is stable. It has indicated that the proposed classifier is superior to some other classifier.

II. LITERATURE REVIEW

Zhen Mei, Qi Shen, Baoxian Ye (2008) proposed an efficient hybridized k-nearest neighbor (KNN) classifiers and SVM algorithm for multiclass classification of gene expression data. This KNN algorithm prunes training samples and combines with SVM to classify samples. Compared with SVM and KNN, the Misclassification rate of HKNNSVM for datasets were notably lower, which indicated that the classification performance of HKNNSVM was stable.

Tai Li, Shenghuo Zhu,Mistsunori Ogihara(2008) proposed an method for categorization of text document via discriminant analysis. Here problem of text categorization are optimized via discriminative analysis, then categorize the text by finding coordinate transformations that reflect similarity from data. By using generalized singular value decomposition" (GSVD), it's a transformation that reflects class structure indicated by singular values is identified. But the cost of the operation is extremely large in document analysis.

Yiming Yang and Xin Liu proposed an effective re-examination of text categorization approaches of statistical test using five categorization method as KNN, SVM, NNet (Neural Network), NB (Naive Bayes) classifier , LLSF (LinearLeast square Fit). Among them SVM, KNN, LLSF outperform the results than the NB, NNet when the number of positive training set per categories are small.

Euihong (sam) Han,George karypis , and Vipin Kumar (1999) proposed an text categorization of documents by using K nearest neighbor. The KNN learn the importance of discriminating words by using techniques as mutual information and weight adjustment. Using WAKNN for categorization facing a problem as how to avoid local minima and solution to local minima lead to changing weights of multiple words at a time.

III. KNN AND SUPPORT VECTOR MACHINE APPROACH

The KNN (k-nearest neighbor) method is said to efficient and provides good results in classification, they are performs as lazy learning method which keeps the entire training samples until classification time. Text classification using single approach is not much effective output. No hybrid approach is used to predict the text classification. There is not much preference for identifying the short, stem, and stop words for the text classification. The similarity measure for the nearest neighbor doesn't work much accuracy values to generate the text classification .The main drawback is accuracy of classification is not too high area.

An SVM constructs a hyper plane or set of hyper planes in a high –or infinite –dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training data point of any class and so- called functional margin, since in general the larger the margin the lower the generalization error of the classifier. The vector space is finded by:

-(1)

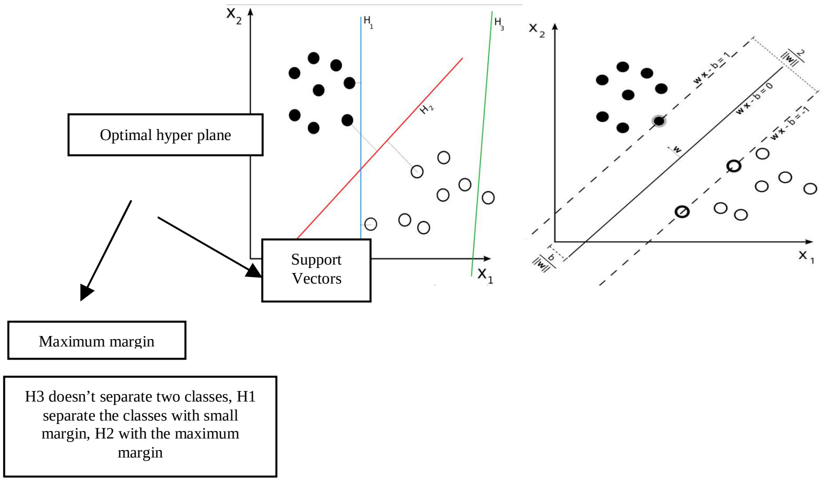
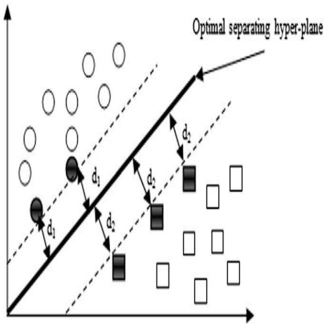


Fig:1 Maximal margin and Vetor Space

IV. OUR CONTRIBUTION

Applied a hybrid approach for the text classification. In the training stage, the SVM is utilized to reduce the training samples for each of the available categories to their support vectors (SVs). The nearest centroid classifier approach is combined with the SVM. NC-SVM provides the accurate text classification as more good performance than K-nearest neighbor. Even though the nearest neighbor and support vector machine involves high effective classification at individual works they combined to produce more accurate text.



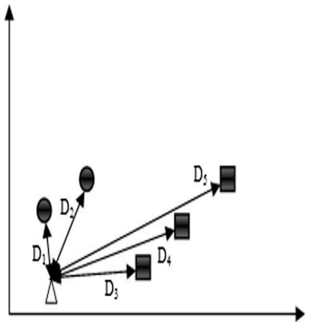


Fig:2 Optimal hyper-plane **Fig:3 Distance Estimation**

There are two categories of data points, which have been mapped into the vector space, represented by "circle" and "square" respectively. Based on the optimal separating hyper-plane can be constructed by maximizing the margin of d1 + d2. After identifying the SVs of each of the categories, the rest of the training data points could be eliminated. In the classification stage, the optimal separating hyper-plane is discarded since its role in making the classification decision has been replaced by the nearest centroid function instead of Euclidean distance function.

$$D = \sqrt{(\sum_{i=1}^{n} (p_i - q_i)^2}$$

In the classification stage ,the input data points from the testing data and SV from the SVM are applied for the distance estimation function are used. Nearest centroid distance, function is also stated as nearest prototype classifier. During classification, the label for the training class is assign by the centroid (mean) of the closest observations.

V. FLOW GRAPH

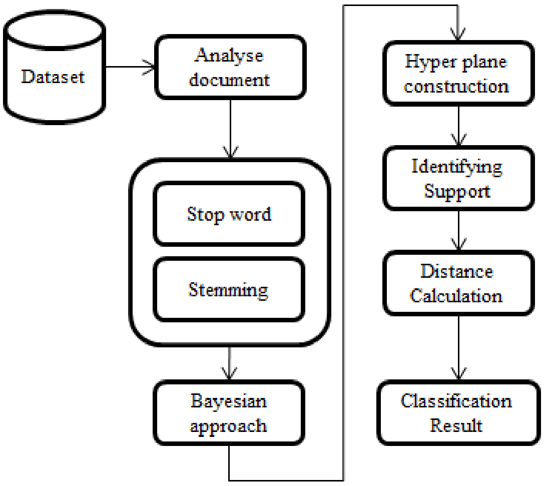


Fig 4. Algorithm flow

VI. MODULE DESCRIPTION

A. Analyze Dataset

Include all the documents and extract the content of the documents. The contents of the documents further performed the removal of stop words and stemming. This basic mining performance for the document content will be performed for all the words in the document. Here we use the Reuter dataset consists the classification for the title and its sentence. This phase runs with the training data set and pre-processing the original data set and identifying the short, stem, stops words. Finally the preprocessed dataset are considered for further process.

B. Training Sets

This phase we considering the input of pre-processed datasets and further more we classify the data and stored in the database. The database maintains the classified datasets. Further identifying the repetitive title and sentence values to identify the Bayesian vectorization module. The Bayesian value is identified for the repetitive values with the help of classifier method. Which helps in the identifying the classification values for the training sets. The KNN compiles the entire training data points again when there is a new input sample and it discards the immediate result. This involves the nearest neighbor method based graph is drawn. As per they show the positions of the title sentence. This generated nearest neighbor helps in easy text classification.

C. Hyper-Plane Construction

These hyper planes are said to the separation or classifier constraint for the text classification. Optimal separating hyper-plane plays important role in the identifying the support vector. As before performing the construction of optimal hyper-plane we insert all the training data points. There are many hyper planes generated with the help of support vector machine values. There are an infinite number of hyper-planes (the dashed lines) could be generated, but there is only one hyper-plane (the solid line) which could optimally separate the data points from different categories. These values will be identified with the help of size; price allocated for the transaction in GB or Mb, frequency of the dataset is taken. As these values iteratively find for all records under the field. Hence the dataset size remains same no further elimination is performed in this module. Finally we identify the minimum privacy preserving cost. That the minimum solution mentioned herein is somewhat psedominimum because an upper bound of joint privacy leakage is just an approximation of its exact value.

D. Support Vector

The support vector machine (SVM) has been reported as a discriminative classifier which is more accurate than most other classification models. The nearest data points to the optimal separating hyper-plane are called support vectors (SVs). There is a certain way to represent the SVs for a given set of training data points, and the maximal margin can be found by minimizing. Support vectors of each category are identified, and the remaining training data points are discarded. New unlabeled data point is mapped into the same vector space of support vectors which obtained from the training stage. The SVM-NN approach suffers from the high time consumption in the classification stage, due to the fact that the average Euclidean distances between the input data point and the support vectors for each of the categories are needed to be calculated in order to make the classification decision.

E. Distance Calculation

He Euclidean distance formula for this computation. The average distance of the SVs of a particular category and the new data point is calculated by using the formula as illustrated. The distance calculation defines the spaces in the hyper plane, for identify the points that falls inside the space region between the hyper plane and the Euclidean distance they are considered to the text classifications. Those count value of the points fall in between the region said to the text classifications.

II. EXPERIMENTS

For our experiments we used a variety of datasets, most of which are frequently used in the information retrieval research. The range of the number of classes is from 4 to 105 and the range of the number of documents is from 476 to 20,000, which seem varied enough to obtain good insights as to how GDA performs.

A. The characteristics of the dataset

REUTERS-21578

The Reuters-21578 R8 dataset which had been used in our experiments was acquired from Ana Cardoso-Cachopo's website, which is the same source where the WebKB dataset was acquired. This collection consists of 7670 documents which had been categorized into 8 categories. The documents in the collection had been divided into training set and testing set, which consist of 5483 documents and 2187 documents respectively, which had been categorized into 8 categories. The documents in the collection had been divided into training set and testing set, which consist of 5483 documents and 2187 documents respectively. The characteristics of the dataset are Text and the attribute characteristic is categorical type. Number of instance in the dataset is 21578 and the number of attribute is 5

List of categories of the Reuters-21578 R8 dataset.

1	**Acq**
2	**Crude**
3	**Earn**
4	**Grain**
5	**Interest**
6	**Money-FX**
7	**Ship**
8	**Trade**

Table: 1 Reuters Dataset

VII. RESULTS AND DISCUSSION

The performance analysis shows that the accuracy of the KNN classifier is good for lesser values of the parameter. But as the parameter value k increases the accuracy of classification and decreases gradually. In the proposed SVM-NN method the accuracy stays optimal for even huge values of the parameter c. The accuracy compared to the KNN method is higher in the SVM-NN. The classification can be calculated by using the metrics given below,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

T-True; F-False; P- Positive; N- Negative

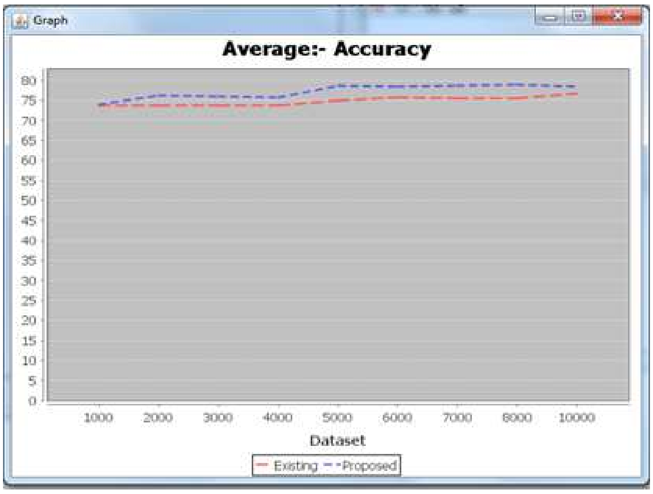


Fig: 5 Comparison of accuracy

VIII. Conclusion

A hybrid classification approach which incorporates the SVM to the training stage of the KNN classification approach is presented. Unlike the conventional KNN classification approach, the SVM-NN approaches have low impact on the implementation of the parameter. The classification accuracy of the SVM-NN approach is relatively consistent with the implementation of the different values of parameter, as compared to the conventional KNN approach. The classification accuracy is severely degraded if inappropriate values of parameter are reported to the classifier. Hence, the determination of the appropriate value for the parameter is not a critical requirement for the SVM-NN classification approach. Especially in the situation where the training samples are limited and insufficient for the preparation of the training set and the validation set. However, the SVM-NN approaches suffers from high time consumption in the classification stage ,due to the fact that the average Euclidean distances between the input data point and the support vectors for each of the categories are needed to be calculated in order to the classification decision. In the future, other alternative methods for calculating distance and similarity measurement, with lower computational cost, in order to propose a more effective and efficient classification approach.

IX. REFERENCES

[1] Blanzier, E. & Bryl, A. (2007b). Evaluation of the highest probability SVM neighbour classifier with variable relative error cost.IN Proceedings of the 4th conference on email and anti-spam, Aug. 2-3, Mountain View, California, USA., pp. 5-9J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Chan, J. N., Huang, H. K., Tians, S. F., Qu, Y.L.(2009). Feature selection for text classification with Naive Bayes Systems with Applications, 36(3),5423-5435.

[3] Han, E.H., Karypis, G. Kumar, V. (1999). Text categorization using weighted adjusted K-nearest neighbour classification. Technical Report , Department of Computer Science and Engineering, Army HPC Research Center, University of Minnesota, Minneapolis, USA.

[4] Isa, D., Lee, L. H., Kallimani, V. P., Rajkumar,R.(2008).text document preprocessing with the Bayes formula for classification using the support vector machine .IEEE Transactions on Knowledge and Data Engineering,20(9),1264-1272.

[5] Joachins, T. (1998). Text categoriation with support vector machines: learning with many relevant featuers. In Proceedings of the 10th European conference on machine learning (ECML-98), pp. 137-142.

[6] Lee , L. H.,Wan, C. H., Rajkumar, R., Isa, D.(2011a). An enhanced support vector machine classification framework by using Euclidean distance function for text document categorization, Applied Intelligence . DOI: 10.1007/s10489-011-0314-z.

[7] Lee, L. H., Rajkumar, R., & Isa, D. (2010b). Automatic folder allocation system using Bayesian-support vector machines hybrid classification approach. Applied Intelligence.DOI:10.1007/s10489-010-0261-0.

[8] Lee, L. H., Wan , C. H., Yong, T. F., & Kok , H. M.(2010c). A review of nearest neighbor support vector machines hybrid classification models .Journal of Applied Sciences,10(17),pp- 1841-1858

[9] Torkkola , K. "Discriminative features for text document classification" ,Pattern Analysis and Applications, Vol.6, pp. 301-308, 2003.

[10] McCallum, A. & Nigam, K.(1998).A comparison of vent models for Naive Bayes text classification . In AAAI-98Workshop on Learning for Text Categorization, pp.41-48.