ISSN 2347 - 3983

Volume 8, No. 1 January 2020

International Journal of Emerging Trends in Engineering Research

Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter08812020.pdf

https://doi.org/10.30534/ijeter/2020/08812020

Ensemble of Classifiers in Text Categorization

M.Govindarajan
Associate Professor, Department of Computer Science and Engineering
Annamalai University, Annamalai Nagar – 608002
Tamil Nadu, India
govind_aucse@yahoo.com

ABSTRACT

Text Categorization has attracted the attention of the research community in the last decade. Algorithms like Support Vector Machines, Naïve Bayes, Genetic Algorithm have been used with good performance, confirmed by several comparative studies. Recently, ensemble-based classifiers have gained popularity in this domain. In this research work, efficient ensemble methods are addressed for developing accurate classifiers for usenet1 dataset. The proposed approach employs Naïve Bayes (NB), Support Vector Machine (SVM) and Genetic Algorithm (GA) as base classifiers along with different ensemble methods. The experimental results show that the ensemble classifiers were performing with accuracy greater than individual classifiers, and also hybrid model results are found to be better than the combined models for the usenet1 dataset. The proposed ensemble-based classifiers turn out to be good in terms of classification accuracy, which is considered to be an important criterion to develop ensemble classifiers for text categorization.

Key words: Accuracy, Genetic Algorithm, Naïve Bayes, Support Vector Machine, Text Categorization.

1. INTRODUCTION

Text categorization is the problem of automatically assigning one or more predefined categories to free text documents. While more and more textual information is available online, effective retrieval is difficult without good indexing and summarization of document content. Document categorization is one solution to this problem. A growing number of statistical classification methods and machine learning techniques have been applied to text categorization in recent years.

This problem is solved using supervised classification algorithms [11]. From the document set, a feature space is extracted based on a set of unique, uncommon and frequent terms which are evaluated for each document. Many comparative studies have been presented in the last years to understand which classifiers should be the most adequate to the Text categorization domain problems. In the last years, ensemble learning is also considered to improve the text categorization performance.

Ensemble Learning [14] [6] [1] is a technique where multiple models (such as expert system classifiers) are trained to solve the complex problem in the machine learning. It contains a collection of base learners that can work coherently with each other. The basic use of Ensemble Learning is to improve the performance of a classification model or prediction model and to minimize the likelihood of an ill-fated selection of a deprived classifier. Generally, in two steps Ensemble can be created; firstly, various base learners can be generated in sequential or parallel styles that are also used to influence the subsequent learner by the base learners. Consequently, the base learner is collectively used to calculate the majority voting and weighted aggregation for classification and regression, respectively.

In this paper, a classifier ensemble is designed using homogeneous and heterogeneous models for usenet1dataset and evaluated in terms of accuracy. The paper is conducted as follows**:** The related work is discussed in Section 2. The methodology is described in Section 3 and the performance evaluation measures are presented in Section 4. The experimental results and discussion are focused in Section 5. Section 6 summarizes and concludes the results.

2. RELATED WORK

A number of classification methods have been discussed in the literature for Text categorization. In [5], a study to compare Support Vector Machines (SVM), k Nearest Neighbors (kNN) and Naïve Bayes (NB) is presented to perform binary Text categorization. It concludes that all the algorithms should be considered as long as the optimal parameter settings could be used for each one. In [18], SVM, NB, logistic regression and LLSF (Linear Least Square Fit) are also compared. All but NB consistently achieve a top performance. Another algorithm usually considered for this task is the neural network (NNET) one [17].

Despite this straight forwarding knowledge achieved with single supervised learning techniques, the community attention changed its main focus in the last years: the researchers tend to use complex and advanced techniques to

solve these problems. Many hybrid techniques to build ensembles of classifiers for text categorization have been recently used [7]. It is commonly observed that the ensemble accuracy is superior when compared to its base classifiers
Reference [13] presented a new approach known as RSS-SVM, which is used to optimize kernel function and penalty parameters through the Ringed Seal Search algorithm. Experiments are conducted on three text datasets named: Reuter21578, 20 newsgroup and TDT2 with a different number of classes, which shows that, proposed RSS-SVM present significant results are having 79.22% accuracy, 70.79% recall, 58% precision and 54.71% f-measure among the previous GA-SVM and CS-SVM algorithms.

Reference [12] proposed three Bayesian counterparts, where it turns out that classical NB classifier with Bernoulli event model is equivalent to Bayesian counterpart. Finally, experimental results on 20 newsgroups and WebKB data sets show that the performance of Bayesian NB classifier with multinomial event model is similar to that of classical counterpart, but Bayesian NB classifier with Gaussian event model is obviously better than classical counterpart.

Reference [15] proposed a novel text classifier using DNN, in an effort to improve the computational performance of addressing big text data with hybrid outliers.

Reference [10] provided an empirical study of a feature selection method based on genetic algorithms for different text representation methods. This feature selection algorithm can accomplish two goals: in one hand is the search of a feature subset such that the performance of classifier is best; in other hands is find a feature subset with the smallest dimensionality which achieves higher accuracy in classification. To evaluate the performance of this approach, three from the best classifiers have been selected: Naive Bayes (NB), Nearest Neighbors (KNN) and Support Vector Machines (SVMs). The objective of the paper is to determine whether the genetic algorithms based feature selection will improve the performances in text classification with smaller size using F-measure. Experimentations were carried out on two benchmark document collections 20Newsgroups, and Reuters-21578.

Reference [16] proposed a new Hybrid Deep Belief Network (HDBN) which uses Deep Boltzmann Machine (DBM) on the lower layers together with Deep Belief Network (DBN) on the upper layers. The advantage of DBM is that it employs undirected connection when training weight parameters which can be used to sample the states of nodes on each layer more successfully and it is also an effective way to remove noise from the different document representation type; the DBN can enhance extract abstract of the document in depth, making the model learn sufficient semantic representation.

In this paper, a hybrid system is proposed using Naïve Bayes, Support Vector Machine and Genetic Algorithm and the effectiveness of the proposed bagged NB, bagged SVM, bagged GA and NB-SVM-GA hybrid system is evaluated by conducting several experiments on usenet1dataset.

3. METHODOLOGY

This research work proposes new hybrid methods for sentiment mining problems. In this paper, usenet1dataset from UCI machine learning repository is taken as input data for analyzing the various classification techniques using WEKA data mining tool [2]. The architecture based on combined and hybrid ensemble models is proposed by combining the base classifiers such as Naïve Bayes (NB), Support Vector Machine (SVM), Genetic Algorithm (GA) with bagging and arcing classifiers to enhance the classification performance. The various classification algorithms and ensemble models are analyzed to select the best classifier for usenet1 dataset.

3.1 Data Pre-processing

Data pre-processing, an important step in data mining, improves the performance and accuracy of the classifiers by improving the data quality.

3.2 Document Indexing

Indexing is an important process in Information Retrieval (IR) systems. It forms the core functionality of the IR process since it is the first step in IR and assists in efficient information retrieval. Indexing reduces the documents to the informative terms contained in them.

3.3 Dimensionality Reduction

To handle high dimensional real-world data adequately, its dimensionality needs to be reduced. Dimensionality reduction is the transformation of such data into a meaningful representation of reduced dimensions. It is an effective approach to downsizing data. It plays an important role in classification performance.

3.4 Base Classifiers

A. Naïve Bayes (NB)

First technology is the Naïve Bayes classifier algorithm [9] which is based on Bayes classification theory. The technique classifies text according to the particular feature of text. This value of particular feature is dependent on a probability of class variables.

Naïve Bayes theorem prepares the system efficiently follow the supervised learning strategy with respect to probability reasoning. The Naïve Bayes classifiers have worked, to solve many of the complex real world conditions. An important and effective benefit of the algorithm is requiring a small amount of the training data to evaluate parameters like means, variances for text classification. For predicting the future events Bayesian Reasoning is used to apply to make the decision and the inferential statistics which will deals with the probability of inference rule. Probability Rule, according to the Naïve Bayes theorem, which are as follows:

P(h/D)= {P(D/h) P(h)}

Where, P(D/h) - Probability of D under given h.

B. Support Vector Machine (SVM)

SVM was introduced by-Guyon, Boser and Vapnik, widely used for classification, pattern recognition and regression. SVM has the capability to classify the dimensions or the size of input space. SVM acquires major advantages because of high generalization performance with prior knowledge. The goal of SVM is find the best classification-function, even it aims to differentiate between the members of two classes in training the data. SVM needs to classify given patterns correctly which can maximize the efficiency of SVM Algorithm. SVM use the Vector Space Model (VSM) to separate samples into different classes, viz. done by the learning process of Support Vector Machine. The three types of learning process i.e. used in SVM are Supervised, Unsupervised and Semi-Supervised Learning [9].

C. Genetic Algorithm (GA)

Genetic Algorithm is an optimized technique which is derived from the Darwin's Principle. It gives an Adaptive Procedure for the survival of first Natural Genetics. GA maintains the number of potential solutions of candidate problem which can be termed as individuals, by the manipulation of these individuals with the help of genetic operators like Crossover, mutation, Selection [9].

3.5 Bagging Ensemble Classifier

In bagging [4], also called bootstrap ensemble technique, each classifier is trained to create the individual classifiers of an ensemble by redistributing the training set with random sampling. The algorithm of Bagging consists of training a pool of base models on the training sets sampled from the same distribution, and combining them by majority vote for classification tasks.

The Bagging Algorithm:

```
Bagging ({$(x_1, y_1), (x_2, y_2), ...., (x_N, y_N)\}, M$)
    For each m=1,2,….,M
     $T_m$=Sample_With_Replacement
    ({$(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}, N$)
     $h_m = L_b(T_m)$
    Return $h_{fin}(x) = \arg\max_{y \in Y} \sum_{m=1}^{M} I(hm(x) = y)$
    Sample_With_Replacement (T,N)
    $S = \{\}$
    For i = 1,2,….,N
     R= random_integer(1,N)
      Add T[r] to S.
     Return S.
```

3.6 Arcing Ensemble Classifier

The framework of arcing introduced by Breiman [3] is similar to the one employed in boosting. They both proceed in sequential steps. The major difference between arcing and boosting is that arcing improves its behavior based on the accumulation of its faults in history. It examines all previous base classifiers' faults for construction of a new base classifier while boosting only checks the previous one base classifier. Apart from this, arcing adopts un-weighted voting system whereas boosting uses weighted voting. In addition, unlike boosting, no checking procedure exists through the constructions of base classifiers.

Like Bagging, Arcing selects with replacement the samples from the original N training set and chooses a training set of size N for classifier K + 1. But the samples are not selected equally in the training set.

4. EVALUATION MEASURES

4.1 Cross Validation

This paper involves 10-fold cross-validation, the data are first partitioned into 10 equally (or nearly equally) sized segments or folds, trained and performed the validation [8].

4.2 Criteria for Evaluation

The efficiency of a classifier is best evaluated using accuracy as a metric. In this work, the performance of the classifier ensembles are analyzed and compared in terms of classification accuracy.

5. EXPERIMENTAL RESULTS

5.1 Dataset Description

The Usenet1Dataset is based on the 20 newsgroups collection. They simulate a stream of messages from different newsgroups that are sequentially presented to a user, who then labels them as interesting or junk, according to his/her personal interests.

5.2 Results and Discussion

In the current investigation, accuracy of usenet1 is the key criterion as it measures the degree of reliability. The usenet1 dataset is taken to evaluate the proposed Bagged and hybrid classifiers.

A. Performance of the Bagging Ensemble Classifier

Table 1: The performance of base and proposed bagged NB classifier for usenet1 data

Dataset	Classifiers	Accuracy
Usenet1 Data	Existing NB Classifier	66.40%
	Proposed Bagged NB Classifier	68.53%

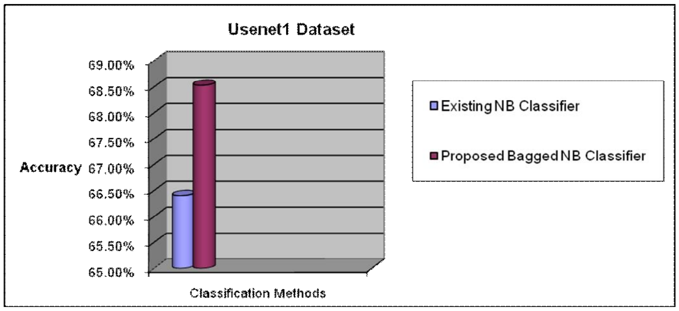

Figure 1: Classification Accuracy of Existing and Proposed Bagged NB Classifier using Usenet1Data

Table 2: The performance of base and proposed bagged SVM classifier for usenet1 data

Dataset	Classifiers	Accuracy
Usenet1 Data	Existing SVM Classifier	66.13%
	Proposed Bagged SVM Classifier	70.66%

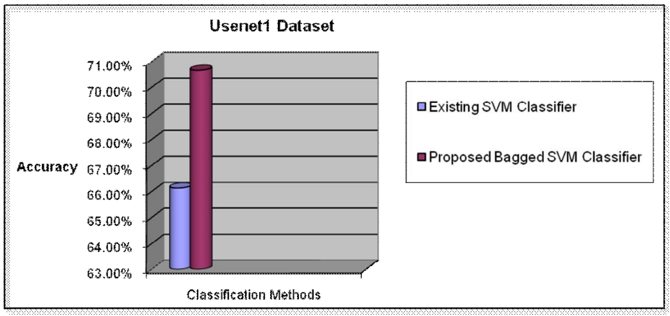

Figure 2: Classification Accuracy of Existing and Proposed Bagged SVM Classifier using Usenet1Data

Table 3: The performance of base and proposed bagged GA classifier for usenet1 data

Dataset	Classifiers	Accuracy
Usenet1 Data	Existing GA Classifier	59.46%
	Proposed Bagged GA Classifier	61.93%

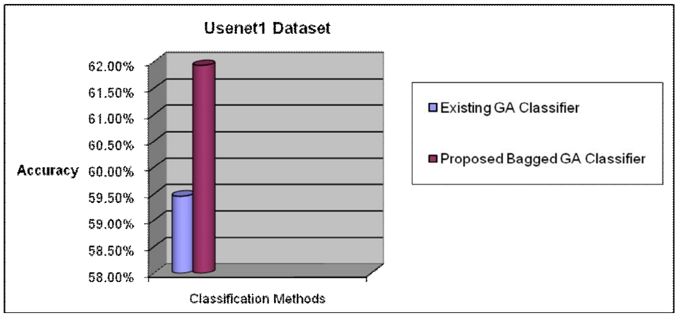

Figure 3: Classification Accuracy of Existing and Proposed Bagged GA Classifier using Usenet1Data

Table 1 to Table 3 shows the accuracy or the percentage of correctly classified instances of usenet1 dataset using the bagging ensemble technique. Clearly, from the experiment it can be observed that 'bagging' technique provides better accuracy values in comparison with individual approaches for usenet1dataset. Figure 1 to Figure 3 show the performance of the base models and proposed bagged models when applied to the usenet1dataset. The result of the proposed bagged model of a particular type, under the given metric, is reported. The accuracy of the combined models is better than the individual models for the usenet1dataset. In this work, the higher accuracy value has been achieved by proposed bagged model.

B. Performance of the Arcing Ensemble Classifier

Table 4: The performance of base and proposed hybrid classifier for usenet1 data

Dataset	Classifiers	Accuracy
Usenet1 Data	Naive Bayes	66.40%
	Support Vector Machine	66.13%
	Genetic Algorithm	59.46%
	Proposed Hybrid NB-SVM-GA	74.66%

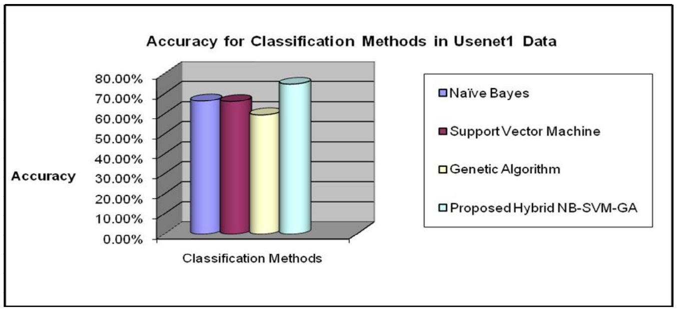

Figure 4: Classification Accuracy of Base and Proposed hybrid NB-SVM-GA Classifier using Usenet1Data

In Table 4, the accuracy or the percentage of correctly classified instances of usenet1 dataset using the arcing ensemble technique has been shown. Figure 4 shows the classification accuracy of usenet1dataset using proposed hybrid model. Clearly, it can be observed from the results that the proposed hybrid NB-SVM-GA is superior to individual approaches for usenet1 dataset in terms of classification accuracy and found to be statistically significant.

6. CONCLUSION

In this research work, ensemble based classifiers using bagging and arcing have been proposed for usenet1 dataset. As a result of experiments, some of the main findings of this work are as follows:

- ❖ NB performs better than SVM and GA in the important respects of accuracy.
- ❖ The proposed bagged methods exhibit significantly higher improvement of classification accuracy than the base classifiers.
- ❖ The hybrid NB-SVM-GA shows higher percentage of classification accuracy than the base classifiers.

- The result of $\chi 2$ statistic analysis shows that the proposed classifiers are significant at $p < 0.05$ than the existing classifiers.
- The heterogeneous model gives better results than homogeneous models for usenet1 data set in terms of accuracy.
- The usenet1 dataset could be detected with high accuracy for ensemble models.

Future ideas will be to develop more efficient ensemble models for large usenet1 datasets.

ACKNOWLEDGEMENT

Author gratefully acknowledges the authorities of Annamalai University for the facilities offered and encouragement to carry out this work.

REFERENCES

1. Allae Erraissi, Abdessamad Belangour, **Meta-modeling of Big Data management layer**, *International Journal of Emerging Trends in Engineering Research*, Vol.7, No.7, pp.36-43, 2019. https://doi.org/10.30534/ijeter/2019/01772019
2. Amita Dhankhar, Kamna Solanki, **A Comprehensive Review of Tools & Techniques for Big Data Analytics**, *International Journal of Emerging Trends in Engineering Research* , Vol.7, No.11, pp.556-562, 2019. https://doi.org/10.30534/ijeter/2019/257112019
3. Breiman. L, **Bias, Variance, and Arcing Classifiers**, *Technical Report 460*, Department of Statistics, University of California, Berkeley, CA, 1996.
4. Breiman, L, **Bagging predictors**, *Machine Learning*, vol.24, no.2, pp.123–140, 1996a. https://doi.org/10.1007/BF00058655
5. Colas, F., Brazdil, P, **Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks**, *In: Artificial Intelligence in Theory and Practice*, pp. 169–178, 2006.
6. J. Han, ***Data Mining: Concepts and Techniques***, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc., 2005.
7. Khan, A., Baharudin, B., Lee, L., Khan, K, **A Review of Machine Learning Algorithms for Text-Documents Classification**, *Journal of Advances in Information Technology*, 1, 2010. https://doi.org/10.4304/jait.1.1.4-20
8. Kohavi, R, **A study of cross-validation and bootstrap for accuracy estimation and model selection**, *Proceedings of International Joint Conference on Artificial Intelligence*, Montreal, Quebec, Canada, Vol.2, 1995, pp.1137–1143.
9. Meet Photographer, **Sentiment Analysis: Algorithmic and Opinion Mining Approach**, *International Research Journal of Engineering and Technology*, vol.6, no.3, pp. 49-53, 2019.
10. Noria Bidi,Zakaria Elberrichi, **Feature selection for text classification using genetic algorithms**, *8th International Conference on Modelling, Identification and Control (ICMIC)*, Algiers, Algeria, 2016.
11. Pankaj Hooda, Pooja Mittal, **An Exposition of Data Mining Techniques for Customer Churn in Telecom Sector**, *International Journal of Emerging Trends in Engineering Research*, Vol.7, No.11, pp.506-511, 2019. https://doi.org/10.30534/ijeter/2019/177112019
12. Shuo Xu, **Bayesian Naive Bayes classifiers to text classification**, *Journal of Information Science*, vol.44, no. 1, pp.48–59, 2018. https://doi.org/10.1177/0165551516677946
13. Wareesa Sharif, Iwan Tri Yadi Yanto, Noor Azah Samsudin, Mustafa Mat Deris, Abdullah Khan, Muhammad Faheem Mushtaq, Muhammad Ashraf, **An Optimised Support Vector Machine with Ringed Seal Search Algorithm For Efficient Text Classification**, *Journal of Engineering Science and Technology*, vol.14, no. 3, pp.1601 – 1613, 2019.
14. H. Witten and E. Frank, ***Data Mining: Practical Machine Learning Tools and Techniques***, Second Edition (Morgan Kaufmann Series in Data Management Systems). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005.
15. Wulamu Aziguli, Yuanyu Zhang,Yonghong Xie,Dezheng Zhang, Xiong Luo,Chunmiao Li, and Yao Zhang, **A Robust Text Classifier Based on Denoising Deep Neural Network in the Analysis of Big Data**, *Scientific Programming*, Volume 2017, pp.1-10. https://doi.org/10.1155/2017/3610378
16. Yan Yan, Xu-Cheng Yin, Sujian Li, Mingyuan Yang, and Hong-Wei Hao, **Learning Document Semantic Representation with Hybrid Deep Belief Network**, *Computational Intelligence and Neuroscience*, Volume 2015, pp. 1-9. https://doi.org/10.1155/2015/650527
17. Yang, Y., Liu, X, **A Re-Examination of Text Categorization Methods**, *In: 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, pp. 42–49. https://doi.org/10.1145/312624.312647
18. Zhang, T., Oles, F, **Text Categorization Based on Regularized Linear Classification Methods**, *Information Retrieval*, 4, pp. 5–31, 2001.