International Journal of Engineering Science, Advanced Computing and Bio-Technology
Vol. 8, No. 2, April – June 2017, pp. 92 - 107

Mining Newsgroups Using Ensemble Classifiers in Social Network Analysis

M.Govindarajan

Department of Computer Science and Engineering

Annamalai University, Annamalai Nagar, Tamil Nadu, India

E-mail: *govind_aucse@yahoo.com*

Abstract: *Internet is the rapidly growing information gallery that contains rich textual information. This rapid growth makes it difficult for the users to locate relevant information quickly on the web. Document retrieval, categorization, routing and filtering systems are often based on text classification. Text Classification means allocating a document to one or more categories or classes. The ability to accurately perform a classification task depends on the representations of documents to be classified. In this research work, new ensemble classification methods are proposed for homogeneous ensemble classifiers using bagging and heterogeneous ensemble classifiers using arcing classifier and their performances are analyzed in terms of accuracy. A Classifier ensemble is designed using Naive Bayes (NB), Support Vector Machine (SVM) and Genetic Algorithm (GA) as base classifiers. The feasibility and the benefits of the proposed approaches are demonstrated by means of newsgroups dataset that is widely used in the field of sentiment classification. The main originality of the proposed approach is based on five main parts: preprocessing phase, document indexing phase, feature reduction phase, classification phase and combining phase to aggregate the best classification results. A wide range of comparative experiments are conducted for newsgroups dataset. The accuracy of base classifiers is compared with homogeneous and heterogeneous models for newsgroups dataset. The proposed ensemble methods provide significant improvement of accuracy compared to individual classifiers and also heterogeneous models exhibit better results than homogeneous models for newsgroups dataset.*

Key words: *Accuracy, Arcing, Bagging, Genetic Algorithm (GA), Naive Bayes (NB), Support Vector Machine, Text Classification.*

1. Introduction

According to the growth in the amount of text documents over the internet and news sources which make document classification is an important task in document processing. Document classification was widely used in many contexts like document indexing, document analysis, document filtering, automatic distribution or archiving of documents (F. Sebastiani, 2002 and N. Chen et al., 2007). This process difficult to be manual with huge number of documents so automatic classification is better than manual classification because it has more accuracy and time efficiency (N. VasfiSisi et al., 2013 and Nidhi et al., 2011). Natural language processing, data mining, and machine learning techniques work together to automatically classify documents.

There are many machine learning techniques which are used for document classification (Nidhi et al., 2011; Bhumika, S. S. Sehra et al., 2013; B. Baharudin et al., 2010; D. Kumar et al.,2013) such as Bayesian classifier, decision tree, K-nearest neighbor, support vector

Received: 19 September, 2016; Revised: 20 April, 2017; Accepted: 07 June, 2017

machines, neural networks, genetic algorithm, and genetic programming (GP), etc. GP is a supervised machine learning technique and a powerful evolutionary algorithm widely used to evolve computer programs automatically (D. Kumar et al., 2013). The principle components of the GP are a set of functions and terminals that are able to represent the solution of the problem. This paper proposes new ensemble classification methods to improve the classification accuracy. The main purpose of this paper is to apply homogeneous and heterogeneous ensemble classifiers for newsgroups dataset to improve classification accuracy.

The rest of this paper is organized as follows: Section 2 describes the related work. Section 3 presents proposed methodology and Section 4 explains the performance evaluation measures. Section 5 focuses on the experimental results and discussion. Finally, results are summarized and concluded in section 6.

2. Related work

A number of classification methods have been discussed in the literature for text classification. These include, naïve Bayes classifier, decision trees (D. Lewis et al.,1994), neural networks and support vector machines (H. Schutze, et al., 1995), rule learning algorithms (G. Salton et al., 1983), relevance feedback (Yan-Shi Dong et al., 2004).

Svetlana Kiritchenko et al., (2001) introduced a learning technique that decreased the effort needed in applying machine learning. Main problems in text classification are lack of labeled data and the cost required for labeling the unlabeled data.

M. Arun Kumar et. al., (2009) have enhanced TSVM to least squares TSVM (LSTSVM), which is an immensely simple algorithm for generating linear/nonlinear binary classifiers using two non-parallel hyper planes/ hyper surfaces. In LSTSVM, they have solved the two primal problems of TSVM using proximal SVM (PSVM) idea instead of two dual problems usually solved in TSVM. They have further investigated the application of linear LSTSVM to text categorization using three benchmark text categorization datasets: reuters-21578, ohsumed and 20 Newsgroups (20NG) and based on the Comparison of experimental results, against linear PSVM shows that linear LSTSVM has better generalization on all the three text corpuses considered.

Suresh Kumar et.al, (2015) firstly tested SVM classifier on a Labeled edition of unlabeled data and then Naive Bayes classifier was tested. As a result SVM performed very well in comparison with Naive Bayes. Experimental result also showed that the performance of co-training depends on learning method that it used.

The authors in (N. Priyadharshini et al., 2013) used an approach used to segment image document and classify the document regions as text, image, drawings and table. Document

image is divided into blocks using run length smearing rule and features are extracted from every blocks. Discipulus tool has been used to construct the genetic programming based classifier model.

The complexity of natural languages and the extremely high dimensionality of the feature space of documents have made this classification problem very difficult. Saad M. Darwish et al., (2015) proposed work mitigates this difficult by providing an algorithm to classify documents into more than two categories (multi-class classification) at the same time by combining multi-objective technique with the genetic programming of classifiers based on multi-tree representation of documents. This combination has the potential to attain lower errors because classification accuracy on each class is represented as a distinct objective.

Most of research in text categorization has been devoted to binary problems, where a document is classified as either relevant or not relevant with respect to predefined topic. However, there are many sources of textual data, such as Internet News, electronic mail and digital libraries, which are composed of different topics and which therefore pose a multi-class categorization problem.

The common approach for multi-class text categorization is to break the task into disjoint binary categorization problems, one for each class. To classify a new document, one needs to apply all the binary classifiers and combine their predictions into a single decision. The end result is a ranking of possible topics.

Xia et al. (2011) ensemble framework is applied to sentiment classification tasks with the aim of integrating different feature sets and different classification algorithms to produce a more accurate classification procedure.

Freund and Schapire (1995,1996) proposed an algorithm the basis of which is to adaptively resample and combine (hence the acronym--arcing) so that the weights in the resampling are increased for those cases most often misclassified and the combining is done by weighted voting. A hybrid model can improve the performance of basic classifier (Tsai 2009).

In this paper, a hybrid system is proposed using Naive 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 newsgroups dataset. The performance of the proposed bagged NB, bagged SVM, bagged GA and NB-SVM-GA hybrid classifiers are examined in comparison with

standalone NB, SVM and standalone GA classifier and also heterogeneous models exhibits better results than homogeneous models for newsgroups data set.

3. Proposed Methodology

Several researchers have investigated the combination of different classifiers to from an ensemble classifier (D. Tax et al, 2000). An important advantage for combining redundant and complementary classifiers is to increase robustness, accuracy, and better overall generalization. This research work aims to make an intensive study of the effectiveness of ensemble techniques for text sentiment classification tasks. In this work, first the base classifiers such as Naive Bayes (NB), Support Vector Machine (SVM), Genetic Algorithm (GA) are constructed to predict classification scores. All classification experiments were conducted using 10 × 10-fold cross-validation for evaluating accuracy. Secondly, well known homogeneous and heterogeneous ensemble techniques are performed with base classifiers to obtain a very good generalization performance. The feasibility and the benefits of the proposed approaches are demonstrated by means of newsgroups dataset that is widely used in the field of text sentiment classification. A wide range of comparative experiments are conducted and finally, some in-depth discussion is presented and conclusions are drawn about the effectiveness of ensemble technique for text sentiment classification.

This research work proposes new hybrid methods for sentiment mining problems. A new architecture based on coupling classification methods using bagging and arcing classifier adapted to sentiment mining problem is defined in order to get better results. The main originality of the proposed approach is based on five main parts: Preprocessing phase, Document Indexing phase, feature reduction phase, classification phase and combining phase to aggregate the best classification results.

A. Data Pre-processing:

Different pre-processing techniques were applied to remove the noise from out data set. It helped to reduce the dimension of our data set, and hence building more accurate classifier, in less time.

The main steps involved are i) document pre-processing, ii) feature extraction / selection, iii) model selection, iv) training and testing the classifier.

Data pre-processing reduces the size of the input text documents significantly. It involves activities like sentence boundary determination, natural language specific stop-word elimination and stemming. Stop-words are functional words which occur frequently in the language of the text (for example, „a , the , an , of etc. in English language), so that they are not useful for classification. Stemming is the action of reducing words to their root or base form. For English language, the Porter s stemmer is a popular algorithm,

which is a suffix stripping sequence of systematic steps for stemming an English word, reducing the vocabulary of the training text by approximately one-third of its original size. For example, using the Porter's stemmer, the English word "generalizations" would subsequently be stemmed as "generalizations → generalization → generalize → general → gener". In cases where the source documents are web pages, additional pre-processing is required to remove / modify HTML and other script tags.

Feature extraction / selection helps identify important words in a text document. This is done using methods like TF-IDF (term frequency-inverse document frequency), LSI (latent semantic indexing), multi-word etc. In the context of text classification, features or attributes usually mean significant words, multi-words or frequently occurring phrases indicative of the text category.

After feature selection, the text document is represented as a document vector, and an appropriate machine learning algorithm is used to train the text classifier. The trained classifier is tested using a test set of text documents. If the classification accuracy of the trained classifier is found to be acceptable for the test set, then this model is used to classify new instances of text documents.

B. Document Indexing

Creating a feature vector or other representation of a document is a process that is known in the IR community as *indexing*. There are a variety of ways to represent textual data in feature vector form, however most are based on word co-occurrence patterns. In these approaches, a vocabulary of words is defined for the representations, which are all possible words that might be important to classification. This is usually done by extracting all words occurring above a certain number of times (perhaps 3 times), and defining your feature space so that each dimension corresponds to one of these words.

When representing a given textual instance (perhaps a document or a sentence), the value of each dimension (also known as an attribute) is assigned based on whether the word corresponding to that dimension occurs in the given textual instance. If the document consists of only one word, then only that corresponding dimension will have a value, and every other dimension (i.e., every other attribute) will be zero. This is known as the ``bag of words'' approach. One important question is what values to use when the word is present. Perhaps the most common approach is to weight each present word using its frequency in the document and perhaps its frequency in the training corpus as a whole. The most common weighting function is the *tfidf* (term frequency-inverse document frequency) measure, but other approaches exist. In most sentiment classification work, a binary weighting function is used. Assigning 1 if the word is present, 0 otherwise, has been shown to be most effective.

C. Dimensionality Reduction:

Dimension Reduction techniques are proposed as a data pre-processing step. This process identifies a suitable low-dimensional representation of original data. Reducing the dimensionality improves the computational efficiency and accuracy of the data analysis.

Steps:

- ✓ Select the dataset.
- ✓ Perform discretization for pre-processing the data.
- ✓ Apply Best First Search algorithm to filter out redundant & super flows attributes.
- ✓ Using the redundant attributes apply classification algorithm and compare their performance.
- ✓ Identify the Best One.

1) Best first Search:

Best First Search (BFS) uses classifier evaluation model to estimate the merits of attributes. The attributes with high merit value is considered as potential attributes and used for classification Searches the space of attribute subsets by augmenting with a backtracking facility. Best first may start with the empty set of attributes and search forward, or start with the full set of attributes and search backward, or start at any point and search in both directions.

D. Existing Classification Methods:

1) Naive Bayes (NB)

The Naïve Bayes assumption of attribute independence works well for text categorization at the word feature level. When the number of attributes is large, the independence assumption allows for the parameters of each attribute to be learned separately, greatly simplifying the learning process.

There are two different event models. The multi-variate model uses a document event model, with the binary occurrence of words being attributes of the event. Here the model fails to account for multiple occurrences of words within the same document, which is a more simple model. However, if multiple word occurrences are meaningful, then a multinomial model should be used instead, where a multinomial distribution accounts for multiple word occurrences. Here, the words become the events.

2) Support Vector Machine (SVM)

The support vector machine (SVM) is a recently developed technique for multi dimensional function approximation. The objective of support vector machines is to determine a classifier or regression function which minimizes the empirical risk (that is

the training set error) and the confidence interval (which corresponds to the generalization or test set error).

Given a set of N linearly separable training examples $S = \left\{ x_i \in R^n | i = 1,2,...,N \right\}$, where each example belongs to one of the two classes, represented by $y_i \in \{\div 1,-1\}$, the SVM learning method seeks the optimal hyperplane w.x +b = 0, as the decision surface, which separates the positive and negative examples with the largest margins. The decision function for classifying linearly separable data is:

$$f(\mathrm{X}) = sign(W.X + b) \qquad (1)$$

Where w and b are found from the training set by solving a constrained quadratic optimization problem. The final decision function is

$$f(x) = sign\left(\sum_{i=1}^{N} a_i y_i (x_i..x) + b \right) \qquad (2)$$

The function depends on the training examples for which a_i s is non-zero. These examples are called support vectors. Often the number of support vectors is only a small fraction of the original data set. The basic SVM formulation can be extended to the non linear case by using the nonlinear kernels that maps the input space to a high dimensional feature space. In this high dimensional feature space, linear classification can be performed. The SVM classifier has become very popular due to its high performances in practical applications such as text classification and pattern recognition.

The support vector regression differs from SVM used in classification problem by introducing an alternative loss function that is modified to include a distance measure. Moreover, the parameters that control the regression quality are the cost of error C, the width of tube ε and the mapping function ϕ .

In this research work, the values for polynomial degree will be in the range of 0 to 5. In this work, best kernel to make the prediction is polynomial kernel with epsilon = 1.0E-12, parameter d=4 and parameter c=1.0.

3) Genetic Algorithm (GA):

The genetic algorithm is a model of machine learning which derives its behaviour from a metaphor of some of the mechanisms of evolution in nature. This done by the creation within a machine of a population of individuals represented by chromosomes, in essence a set of character strings.

The individuals represent candidate solutions to the optimization problem being solved. In genetic algorithms, the individuals are typically represented by n-bit binary vectors. The

resulting search space corresponds to an n–dimensional boolean space. It is assumed that the quality of each candidate solution can be evaluated using a fitness function.

Genetic algorithms use some form of fitness-dependent probabilistic selection of individuals from the current population to produce individuals for the next generation. The selected individuals are submitted to the action of genetic operators to obtain new individuals that constitute the next generation. Mutation and crossover are two of the most commonly used operators that are used with genetic algorithms that represent individuals as binary strings. Mutation operates on a single string and generally changes a bit at random while crossover operates on two parent strings to produce two offsprings. Other genetic representations require the use of appropriate genetic operators.

The process of fitness-dependent selection and application of genetic operators to generate successive generations of individuals is repeated many times until a satisfactory solution is found. In practice, the performance of genetic algorithm depends on a number of factors including: the choice of genetic representation and operators, the fitness function, the details of the fitness-dependent selection procedure, and the various user-determined parameters such as population size, probability of application of different genetic operators, etc. The basic operation of the genetic algorithm is outlined as follows:

Procedure:

```
begin
t <- 0
initialize P(t)
while (not termination condition)
t <- t + 1
select P(t) from p(t - 1)
crossover P(t)
mutate P(t)
evaluate P(t)
  end
end.
```

Our contribution relies on the association of all the techniques used in our method. First the small selection in grammatical categories and the use of bi-grams enhance the information contained in the vector representation, then the space reduction allows getting more efficient and accurate computations, and then the voting system enhance the results of each classifier. The overall process comes to be very competitive.

E. Proposed Bagged Ensemble Classifiers:

Given a set D, of d tuples, bagging (Breiman, L. 1996a) works as follows. For iteration i (i =1, 2,…..k), a training set, D_i, of d tuples is sampled with replacement from the original set of tuples, D. The bootstrap sample D_i, by sampling D with replacement, from the given training data set D repeatedly. Each example in the given training set D may appear repeated times or not at all in any particular replicate training data set D_i. A classifier model, M_i, is learned for each training set, D_i. To classify an unknown tuple, X, each classifier, M_i, returns its class prediction, which counts as one vote. The bagged (NB, SVM, GA), M^*, counts the votes and assigns the class with the most votes to X.

Algorithm: Bagged ensemble classifiers using bagging

Input:

- D, a set of d tuples.
- k = 3, the number of models in the ensemble.
- Base Classifiers (NB, SVM, GA)

Output: A Bagged (NB, SVM, GA), M^*

Method:

1. for i = 1 to k do // create k models
2. Create a bootstrap sample, D_i, by sampling D with replacement, from the given training data set D repeatedly. Each example in the given training set D may appear repeated times or not at all in any particular replicate training data set D_i
3. Use D_i to derive a model, M_i;
4. Classify each example d in training data D_i and initialized the weight, W_i for the model, M_i, based on the accuracies of percentage of correctly classified example in training data D_i.
5. endfor

To use the bagged ensemble models on a tuple, X:

1. if classification then
2. let each of the k models classify X and return the majority vote;
3. if prediction then
4. let each of the k models predict a value for X and return the average predicted value;

F. Heterogeneous Ensemble Classifiers using Arcing:

1) Proposed NB-SVM-GA Hybrid System

Given a set D, of d tuples, arcing (Breiman. L, 1996) works as follows; For iteration i (i =1, 2,…..k), a training set, D_i, of d tuples is sampled with replacement from the original set of tuples, D. some of the examples from the dataset D will occur more than once in the training dataset D_i. The examples that did not make it into the training dataset end up forming the test dataset. Then a classifier model, M_i, is learned for each training examples

d from training dataset D_i. A classifier model, M_i, is learned for each training set, D_i. To classify an unknown tuple, X, each classifier, M_i, returns its class prediction, which counts as one vote. The hybrid classifier (NB-SVM-GA), M^*, counts the votes and assigns the class with the most votes to X.

Algorithm: Hybrid NB-SVM-GA using Arcing Classifier

Input:

- D, a set of d tuples.
- k = 3, the number of models in the ensemble.
- Base Classifiers (NB, SVM, GA)

Output: Hybrid NB-SVM-GA model, M^*.

Procedure:

1. For i = 1 to k do // Create k models
2. Create a new training dataset, D_i, by sampling D with replacement. Same example from given dataset *D* may occur more than once in the training dataset D_i.
3. Use D_i to derive a model, M_i
4. Classify each example d in training data D_i and initialized the weight, W_i for the model, M_i, based on the accuracies of percentage of correctly classified example in training data D_i.
5. endfor

To use the hybrid model on a tuple, X:

1. if classification then
2. let each of the k models classify X and return the majority vote;
3. if prediction then
4. let each of the k models predict a value for X and return the average predicted value;

The basic idea in Arcing is like bagging, but some of the original tuples of D may not be included in Di, where as others may occur more than once.

4. Performance Evaluation Measures

A. Cross Validation Technique:

Cross-validation, sometimes called rotation estimation, is a technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. 10-fold cross validation is commonly used. In stratified K-fold cross-validation the folds are selected so that the mean response value is approximately equal in all the folds.

B. Criteria for Evaluation:

The primary metric for evaluating classifier performance is classification Accuracy - the percentage of test samples that are correctly classified. The accuracy of a classifier refers to

the ability of a given classifier to correctly predict the label of new or previously unseen data (i.e. tuples without class label information). Similarly, the accuracy of a predictor refers to how well a given predictor can guess the value of the predicted attribute for new or previously unseen data.

5. Experimental Results

A. Dataset Description:

The data set consists of Usenet articles collected from 20 different newsgroups. These were downloaded from https://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html

B. Results and Discussion:

In this section, new ensemble classification methods are proposed for homogeneous ensemble classifiers using bagging and heterogeneous ensemble classifiers using arcing classifier and their performances are analyzed in terms of accuracy.

A) Homogeneous Ensemble Classifiers using Bagging:

The newsgroups dataset is taken to evaluate the proposed Bagged NB, SVM and GA classifiers.

Newsgroup Dataset	Classifiers	Accuracy
misc.forsale	Existing NB Classifier	97.50 %
	Proposed Bagged NB Classifier	98.50 %

Table I. The Performance Of Base And Proposed Bagged Nb Classifier For Newsgroups Data

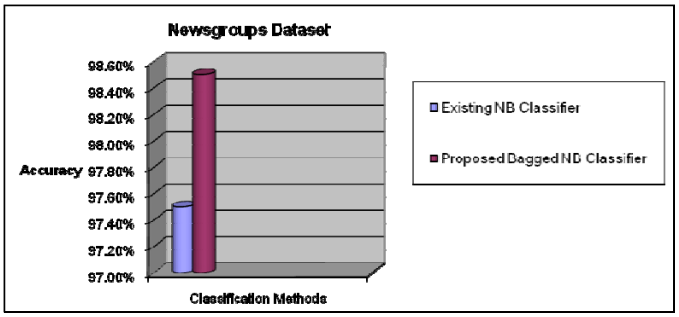

Figure 1.Classification Accuracy of Existing and Proposed Bagged NB Classifier using Newsgroups Data

Newsgroup Dataset	Classifiers	Accuracy
misc.forsale	Existing SVM Classifier	97.90 %
	Proposed Bagged SVM Classifier	98.20 %

Table II. The Performance Of Base And Proposed Bagged Svm Classifier For Newsgroups Data

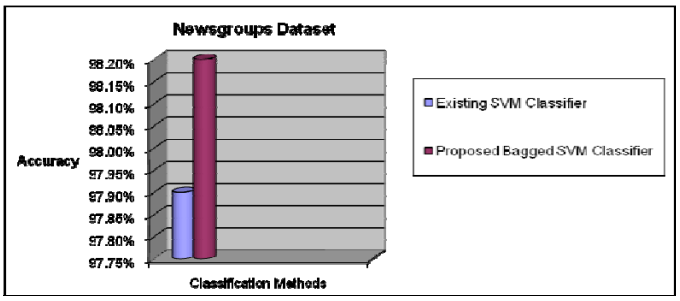

Figure 2. Classification Accuracy of Existing and Proposed Bagged SVM Classifier using Newsgroups Data

Newsgroup Dataset	**Classifiers**	**Accuracy**
misc.forsale	Existing GA Classifier	97.80 %
	Proposed Bagged GA Classifier	98.70%

Table III. The Performance Of Base And Proposed Bagged Ga Classifier For Newsgroups Data

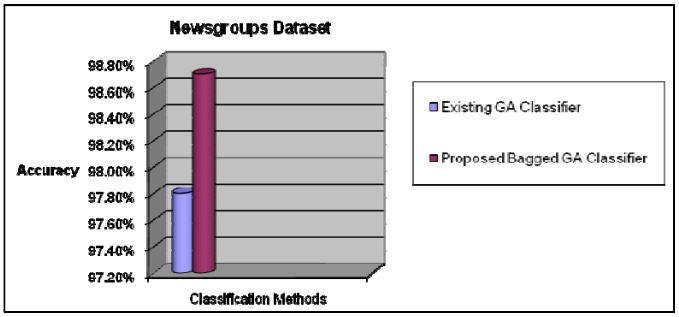

Figure 3.Classification Accuracy of Existing and Proposed Bagged GA Classifier using Newsgroups Data

In this research work, new ensemble classification method is proposed using bagging classifier in conjunction with NB, SVM, GA as the base learner and the performance is analyzed in terms of accuracy. Here, the base classifiers are constructed using NB, SVM, GA. 10-fold cross validation (Kohavi, R, 1995) technique is applied to the base classifiers and evaluated classification accuracy. Bagging is performed with NB, SVM, GA to obtain a very good classification performance. Table 1 to 3 shows classification performance for newsgroups dataset using existing and proposed bagged NB, SVM, GA. The analysis of results shows that the proposed bagged NB, SVM, GA are shown to be superior to individual approaches for newsgroups dataset in terms of classification accuracy. According to Figure. 1 to 3 proposed combined models show significantly larger improvement of classification accuracy than the base classifiers. This means that the combined methods are more accurate than the individual methods for the newsgroups dataset.

B) **Heterogeneous Ensemble Classifiers using Arcing**:

The newsgroups dataset is taken to evaluate the proposed hybrid NB-SVM-GA classifier.

Dataset	Classifiers	Accuracy
misc.forsale	Naive Bayes	97.50 %
	Support Vector Machine	97.90 %
	Genetic Algorithm	97.80 %
	Proposed Hybrid NB-SVM-GA	99.60 %

Table IV. The Performance Of Base And Proposed Hybrid Classifier For Newsgroups Data

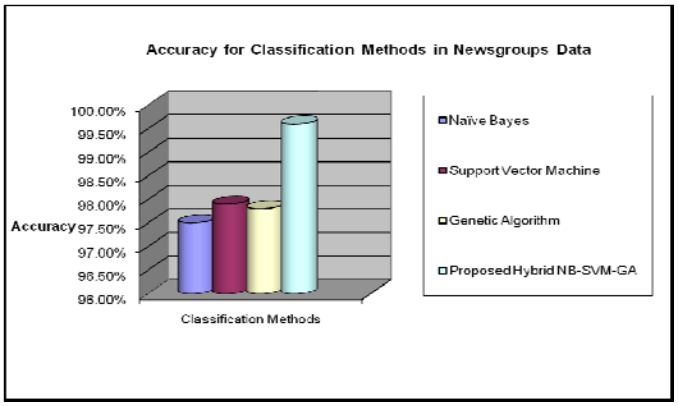

Figure 4.Classification Accuracy of Base and Proposed hybrid NB-SVM-GA Classifier using Newsgroups Data

In this research work, new hybrid classification method is proposed for heterogeneous ensemble classifiers using arcing classifier and their performances are analyzed in terms of accuracy. The data set described in section 5 is being used to test the performance of base classifiers and hybrid classifier. Classification accuracy was evaluated using 10-fold cross validation. In the proposed approach, first the base classifiers NB, SVM and GA are constructed individually to obtain a very good generalization performance. Secondly, the ensemble of NB, SVM and GA is designed. In the ensemble approach, the final output is decided as follows: base classifier's output is given a weight (0–1 scale) depending on the generalization performance as given in Table 4. According to figure 4, the proposed hybrid model show significantly larger improvement of classification accuracy than the base classifiers and the results are found to be statistically significant.

The experimental results show that proposed hybrid NB-SVM-GA is superior to individual approaches for newsgroups dataset in terms of classification accuracy.

6. Conclusion

In this research work, new combined classification methods are proposed for in homogeneous ensemble classifiers using bagging and the performance comparisons have been demonstrated using newsgroups dataset in terms of accuracy. Here, the proposed bagged NB, SVM and GA combines the complementary features of the base classifiers. Similarly, new hybrid NB-SVM-GA model is designed in heterogeneous ensemble classifiers involving NB, SVM and GA models as base classifiers and their performances are analyzed in terms of accuracy

The experiment results lead to the following observations.

- SVM exhibits better performance than GA and NB in the important respects of accuracy.
- The proposed bagged methods are shown to be 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 χ^2 statistic is determined for all the above approaches and their critical value is found to be less than 0.455. Hence corresponding probability is $p < 0.5$. This is smaller than the conventionally accepted significance level of 0.05 or 5%. Thus examining a χ^2 significance table, it is found that this value is significant with a degree of freedom of 1. In general, the result of χ^2 statistic analysis shows that the proposed classifiers are significant at $p < 0.05$ than the existing classifiers.
- The accuracy of base classifiers is compared with homogeneous and heterogeneous models for newsgroups dataset and heterogeneous models exhibit better results than homogeneous models for newsgroups data set.
- The newsgroups dataset could be detected with high accuracy for homogeneous and heterogeneous models.

Acknowledgment:

Author gratefully acknowledges the authorities of Annamalai University for the facilities offered and encouragement to carry out this work. This work is supported by DST-SERB Fast track Scheme for Young Scientists by the Department of science and technology, Government of India, New Delhi.

References:

[1] M. Arun Kumar and M. Gopal, "Least squares twin support vector machines for pattern classification", Expert Systems with Applications, 36(4), 2009, pp. 7535–7543.

[2] Baharudin, L. H. Lee , and K. Khan, "A review of machine learning algorithms for text-documents classification," Journal of Advances in Information Technology, 1(1), 2010, pp. 4-20.

[3] Bhumika, S. S. Sehra, and A. Nayyar, "A review paper on algorithms used for text classification," International Journal of Application or Innovation in Engineering & Management, 2(3), 2013, pp. 90-99.

[4] Breiman. L, "Bias, Variance, and Arcing Classifiers", Technical Report 460, Department of Statistics, University of California, Berkeley, CA, 1996.

[5] Breiman, "Bagging predictors", Machine Learning, 24(2), 1996a, pp.123–140.

[6] N. Chen and D. Blostein, "A survey of document image classification: Problem statement, classifier architecture and performance evaluation," International Journal of Document Analysis and Recognition (IJDAR), 10(1), 2007, pp. 1-16.

[7] Freund, Y. and Schapire, R. "A decision-theoretic generalization of on-line learning and an application to boosting", proceedings of the Second European Conference on Computational Learning Theory, 1995, pp.23-37.

[8] Freund, Y. and Schapire, R. "Experiments with a new boosting algorithm", Proceedings of the Thirteenth International Conference on Machine Learning, Bari, Italy,1996, pp.148-156.

[9] Kohavi, R, "A study of cross-validation and bootstrap for accuracy estimation and model selection", Proceedings of International Joint Conference on Artificial Intelligence, Vol.2, Montreal, Quebec, Canada, August 20-25, 1995, pp.1137–1143.

[10] D. Kumar and S. Beniwal, "Genetic algorithm and programming based classification: A survey," Journal of Theoretical and Applied Information Technology, 54(1), 2013, pp. 48-58.

[11] D. Lewis and M. Ringutte, "A comparison of Two Learning Algorithm for Text Categorization", Proceedings of the third Annual Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, 1994, pp. 81–93.

[12] Nidhi and V. Gupta, "Recent trends in text classification techniques," International Journal of Computer Applications, 35(6), 2011, pp. 45-51.

[13] N. Priyadharshini and V. MS, "Genetic programming for document segmentation and region classification using discipulus," International Journal of Advanced Research in Artificial Intelligence, 2(2), 2013, pp. 15-22.

[14] Saad M. Darwish, Adel A. EL-Zoghabi, and Doaa B. Ebaid, "A Novel System for Document Classification Using Genetic Programming", Journal of Advances in Information Technology, 6(4), 2015,pp. 194-200.

[15] G. Salton and M. McGill, "Introduction to Modern Information Retrieval", McGraw-Hill, 1983.

[16] F. Sebastiani, "Machine learning in automated text categorization," ACM Computing Surveys (CSUR), 34(1), 2002, pp. 1-47.

[17] H. Schutze, D. Hull, and J. Pedersen, "A Comparison of Classifiers and Document Representations for the Routing Problem", In SIGIR'95, Washington D.C., 1995, pp. 229–237.

[18] Suresh Kumar and Shivani Goel, "Enhancing Text Classification by Stochastic Optimization method and Support Vector Machine", International Journal of Computer Science and Information Technologies, 6(4),2015,pp.3742-3745.

[19] Svetlana Kiritchenko, Stan Matwin, "Email Classification with Co-training", Proceedings of the 2001 conference of the Centre for Advanced Studies on Collaborative Research, 2001, CASCON '01, pp. 8.

[20] Tsai, C. F., Lu, Y.F, "Customer Churn Prediction by Hybrid Neural Network", Expert Systems with Application,39, 2009, pp.12547-12553.

[21] D. Tax, M. Breukelen, R. Duin, and J. Kittler, "Combining multiple classifiers by averaging or by mutliplying?", Pattern Recognition, Vol 33, 2000, pp. 1475-1485.

[22] N. VasfiSisi and M. R. F. Derakhshi, "Text classification with machine learning algorithms," Journal of Basic and Applied Scientific Research, 3(1), 2013, pp. 31-35.

[23] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification", Information Sciences, 181(6), 2011, pp. 1138–1152.

[24] Yan-Shi Dong, Ke-Song Han, "A Comparison of Several Ensemble Methods for Text Categorization", Proceedings of the 2004 IEEE International Conference on Services Computing, Shanghai, China, Sept. 15-18, 2004, pp. 419-422.

Authors' Profile:

M.Govindarajan received the B.E and M.E and Ph.D Degree in Computer Science and Engineering from Annamalai University, Tamil Nadu, India in 2001 and 2005 and 2010 respectively. He did his post-doctoral research in the Department of Computing, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, Surrey, United Kingdom in 2011 and CSIR Centre for Mathematical Modelling and Computer Simulation, Bangalore in 2013. He is currently an Assistant Professor at the Department of Computer Science and Engineering, Annamalai University, Tamil Nadu, India. He has presented and published more than 100 papers at Conferences and Journals and also received best paper awards. He has delivered invited talks at various national and international conferences. His current Research Interests include Data Mining and its applications, Web Mining, Text Mining, and Sentiment Mining. He was the recipient of the Achievement Award for the field and to the Conference Bio-Engineering, Computer Science, Knowledge Mining (2006), Prague, Czech Republic, Career Award for Young Teachers (2006), All India Council for Technical Education, New Delhi, India and Young Scientist International Travel Award (2012), Department of Science and Technology, Government of India New Delhi. He is Young Scientists awardee under Fast Track Scheme (2013), Department of Science and Technology, Government of India, New Delhi and also granted Young Scientist Fellowship (2013), Tamil Nadu State Council for Science and Technology, Government of Tamil Nadu, Chennai. He also received the Senior Scientist International Travel Award (2016), Department of Science and Technology, Government of India. He has completed two major projects as principle investigator and has produced three Ph.Ds and also applied patent in the area of data mining. He has visited countries like Czech Republic, Austria, Thailand, United Kingdom (twice), Malaysia, U.S.A (twice), and Singapore. He is an active Member of various professional bodies and Editorial Board Member of various conferences and journals.