International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-3 Number-4 Issue-13 December-2013

Sentiment Analysis of Movie Reviews using Hybrid Method of Naive Bayes and Genetic Algorithm

M.Govindarajan

Abstract

The area of sentiment mining (also called sentiment extraction, opinion mining, opinion extraction, sentiment analysis, etc.) has seen a large increase in academic interest in the last few years. Researchers in the areas of natural language processing, data mining, machine learning, and others have tested a variety of methods of automating the sentiment analysis process. In this research work, new hybrid classification method is proposed based on coupling classification methods using arcing classifier and their performances are analyzed in terms of accuracy. A Classifier ensemble was designed using Naive Bayes (NB), Genetic Algorithm (GA). In the proposed work, a comparative study of the effectiveness of ensemble technique is made for sentiment classification. The ensemble framework is applied to sentiment classification tasks, with the aim of efficiently integrating different feature sets and classification algorithms to synthesize a more accurate classification procedure. The feasibility and the benefits of the proposed approaches are demonstrated by means of movie review that is widely used in the field of 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 sentiment classification.

Keywords

Accuracy, Arcing classifier, Sentiment Mining, Genetic Algorithm (GA), Naïve Bayes (NB).

1. Introduction

Recently, many web sites have emerged that offer reviews of items like books, cars, snow tires, vacation destinations, etc. They describe the items in some detail and evaluate them as good/bad, preferred/not preferred.

M.Govindarajan, Assistant Professor, Department of Computer Science and Engineering, Annamalai University.

So, there is motivation to categorize these reviews in an automated way by a property other than topic, namely, by what is called their 'sentiment' or 'polarity'. That is, whether they recommend or do not recommend a particular item. One speaks of a review as having positive or negative polarity. Now, such automated categorization by sentiment, if it worked effectively, would have many applications. First, it would help users quickly to classify and organize on-line reviews of goods and services, political commentaries, etc. Secondly, categorization by sentiment would also help businesses to handle 'form free' customer feed-back. They could use it to classify and tabulate such feedback automatically and could thereby determine, for instance, the percentage of happy clientele without having actually to read any customer input. Not only businesses but governments and non-profit organizations might benefit from such an application. Thirdly, categorization by sentiment could also be used to filter email and other messages. A mail program might use it to eliminate so-called 'flames'. Finally, perhaps a word processor might employ it to warn an author that he is using bombastic or other undesirable language. In this light, there is suitable motivation to look at the possibility of automated categorization by sentiment. Sentiment analysis is an emerging research area in text mining and computational linguistics, and has attracted considerable research attention in the past few years. Sentiment analysis is a kind of text classification that classifies text based on the sentimental orientation of opinions they contain. It is also known as opinion mining, opinion extraction and affects analysis in the literature. 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

Sentiment analysis of movie reviews is considered to be very challenging since movie reviewers often present lengthy plot summaries and also use complex literary devices such as rhetoric and sarcasm.

Previously used techniques for sentiment classification can be classified into three categories. These include machine learning algorithms, link analysis methods, and score based approaches. The effectiveness of machine learning techniques when applied to sentiment classification tasks is evaluated in the pioneering research by Pang et al, 2002. Many studies have used machine learning algorithms with support vector machines (SVM) and Naïve Bayes (NB) being the most commonly used. SVM has been used extensively for movie reviews (Pang et al, 2002; Pang and Lee, 2004; Whitelaw et al., 2005) while Naïve Bayes has been applied to reviews and web discourse (Pang et al, 2002; Pang and Lee, 2004; Efron, 2004). In comparisons, SVM has outperformed other classifiers such as NB (Pang et al., 2002). Hesham Arafat et al., (2014) results show that mRMR is better compared to IG for sentiment classification, Hybrid feature selection method based on the RST and Information Gain (IG) is better compared to the previous methods. Proposed methods are evaluated on four standard datasets viz. Movie review, Product (book, DVD, and electronics) reviewed datasets, and Experimental results show that hybrid feature selection method outperforms than feature selection methods for sentimental classification. Sumathi T et al., (2013) have compared three methods RIDOR, Naïve Bayes and FURIA. Further it can extent to improve the performance of the system using feature reduction technique. Also 400 reviews were randomly selected from IMDb dataset and feature extracted using stop word, stemming and IDF. The performance of FURIA classifier is better than Naïve Bayes by 8.21 % and by 21.71 compared to RIDOR. Jotheeswaran et al., (2012) proposed feature set extraction from movie reviews. Inverse document frequency is computed and feature set reduced using Principal Component Analysis. Pre processing's effectiveness is evaluated using Naive Bayes and Linear Vector Quantization. Kabinsinghaetal., (2012) investigated movie ratings. Data mining was applied to movie classification. Movies are rated into PG, PG-13 and R in the prototype. The 240 prototype movies from IMDb (http://imdb.com) were used. The other work that used sophisticated feature selection was by Abbasi et al. (2008). They found that using either information gain (IG) or genetic algorithms (GA) resulted in an improvement in accuracy. They also combined the two in a new algorithm called the Entropy Weighted Genetic Algorithm (EWGA), which achieved the highest level of accuracy in sentiment analysis to date of 91.7%. The drawback of this new method is that while it can efficiently classify items, it is very computationally expensive to conduct the initial feature selection, since both GA and IG are expensive to run. Genetic algorithms are search heuristics that are similar to the process of biological evolution and natural selection and survival of the fittest. Genetic Algorithms (GAs) are probabilistic search methods. GAs are applied for natural selection and natural genetics in artificial intelligence to find the globally optimal solution from the set of feasible solutions (S Chandrakala et al, 2012). The experiments with GA's start with a large set of possible extractable syntactic, semantic and discourse level feature set. The fitness function calculates the accuracy of the subjectivity classifier based on the feature set identified by natural selection through the process of crossover and mutation after each generation. The ensemble technique, which combines the outputs of several base classification models to form an integrated output, has become an effective classification method for many domains (T. Ho, 1994; J. Kittler,, 1998). In topical text classification, several researchers have achieved improvements in classification accuracy via the ensemble technique. In the early work (L. Larkey et al, 1996), a combination of different classification algorithms (k-NN, Relevance feedback and Bayesian classifier) produces better results than any single type of classifier. Freund and Schapire (1995,1996) proposed an algorithm the basis of which is to adaptively resample and **c**ombine (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. In this research work, proposes a new hybrid method for sentiment mining problem. A new architecture based on coupling classification methods (NB and GA) using arcing classifier adapted to sentiment mining problem is defined in order to get better results.

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 sentiment classification tasks. In this work, first the base classifiers such as Naïve Bayes (NB), Genetic

Algorithm (GA) are constructed to predict classification scores. The reason for that choice is that they are representative classification methods and very heterogeneous techniques in terms of their philosophies and strengths. All classification experiments were conducted using 10 × 10-fold cross-validation for evaluating accuracy. Secondly, well known 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 movie review that is widely used in the field of 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 sentiment classification. This research work proposes new hybrid method for sentiment mining problem. A new architecture based on coupling classification methods using 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: Pre-processing 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

Two classification methods are adapted for each training set. The most competitive classification methods are used for a given corpus. The results are evaluated using the cross validation method on movie review based on the classification accuracy.

1) Naïve 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) 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 Hybrid NB-GA Method

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-GA), M^*, counts the votes and assigns the class with the most votes to X.

Algorithm: Hybrid NB-GA using Arcing Classifier

Input:

- D, a set of d tuples.
- *k* = 2, the number of models in the ensemble.
- Base Classifiers (Naïve Bayes, Genetic Algorithm)

Output: Hybrid NB-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 basic data set consist of 2000 movie reviews, 1000 labelled positive and 1000 labelled negative (so they have a uniform class distribution). These were downloaded from Bo Pang's web page:

http://www.cs.cornell.edu/people/pabo/movie-eview-data/.

B. Results and Discussion

Table 1: The performance of base and hybrid classifier for movie review data

Dataset	Classifiers	Accuracy
Movie-Review Data	Naïve Bayes (NB)	91.15 %
	Genetic Algorithm (GA)	91.25 %
	Proposed Hybrid NB-GA Method	93.80 %

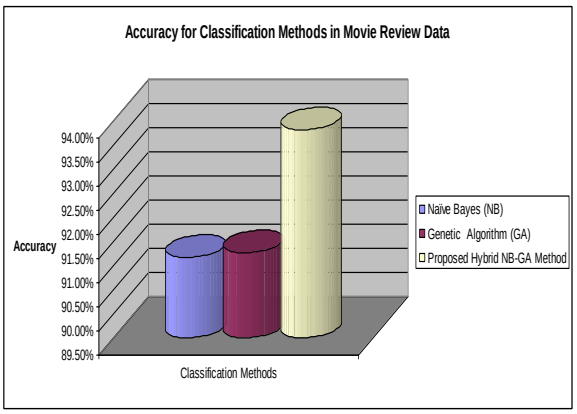

Figure 1: Classification Accuracy for Movie Review Data

The data set described in section 2 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 and Genetic Algorithm are constructed individually to obtain a very good generalization performance. Secondly, the ensemble of NB, 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 1. According to Table 1, the proposed hybrid NB-GA model shows significantly larger improvement of classification accuracy than the base classifiers and the results are found to be statistically significant. The proposed hybrid NB-GA method is shown to be superior to individual approaches for movie review data in terms of Classification accuracy.

6. Conclusions

In this research, a new hybrid technique is investigated and evaluated their performance based on the movie review data and then classifying the reduced data by NB and GA. Next a hybrid NB-GA model and NB, GA models as base classifiers are designed. Finally, a hybrid system is proposed to make optimum use of the best performances delivered by the individual base classifiers and the hybrid approach. The hybrid NB-GA shows higher percentage of classification accuracy than the base classifiers and enhances the testing time due to data dimensions reduction. The experiment results lead to the following observations.

- ❖ GA exhibits better performance than NB in the important respects of accuracy.
- ❖ Comparison between the individual classifier and the hybrid classifier: it is clear that the hybrid classifier show the significant improvement over the single classifiers.

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] A Abbasi, HC Chen and A Salem, (2008), "Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums". ACM Transactions On Information Systems, Volume 26, issue 3, No 12.

[2] L. Brieman. (1996), "Bias, Variance, and Arcing Classifiers", Technical Report 460, Department of Statistics, University of California at Berkeley, CA 94720.

[3] S Chandrakala and C Sindhu, (2012), "Opinion Mining and sentiment classification a survey", ICTACT journal on soft computing.

[4] Efron, M. (2004), "Cultural orientations: Classifying subjective documents by cocitation analysis", In Proceedings of the AAAI Fall Symposium Series on Style and Meaning in Language, Art, Music, and Design, pp. 41-48.

[5] Freund, Y. and Schapire, R. (1995), "A decision-theoretic generalization of on-line learning and an application to boosting", In proceedings of the

Second European Conference on Computational Learning Theory, pp. 23-37.

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

[7] Hesham Arafat, Rasheed M. Elawady, Sherif Barakat,Nora M.Elrashidy., (2014), "Different Feature Selection for Sentiment Classification", International Journal of Information Science and Intelligent System, vol 3issue 1, pp. 137-150.

[8] T. Ho, J. Hull, S. Srihari, (1994), "Decision combination in multiple classifier systems", IEEE Transactions on Pattern Analysis and Machine Intelligence, 16, pp. 66–75.

[9] Jotheeswaran, J., Loganathan, R., and MadhuSudhanan, B, (2012), "Feature Reduction using Principal Component Analysis for Opinion Mining", International Journal of Computer Science and Telecommunications ,Vol.3, No.5,pp.118-121.

[10] S.Kabinsingha,S., Chindasorn,C., and Chantrapornchai.A, (2012), "Movie Rating Approach and Application Based on Data Mining", International Journal of Engineering and Innovative Technology ,Vol. 2, No.1, pp.77-83.

[11] J. Kittler, (1998), "Combining classifiers: a theoretical framework", Pattern Analysis and Applications, 1, pp.18–27.

[12] L. Larkey, W. Croft, (1996), "Combining classifiers in text categorization", in: Proceeding of ACM SIGIR Conference, ACM, New York, NY, USA, pp. 289–297.

[13] B. Pang, L. Lee, S. Vaithyanathan, (2002), "Thumbs up? Sentiment classification using machine learning techniques", in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86.

[14] Pang, B., and Lee, L. (2004), "A sentimental education: Sentimental analysis using subjectivity summarization based on minimum cuts", In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, pp. 271-278.

[15] Sumathi T, Karthik S, Marikannan M, (2013), "Performance Analysis of Classification Methods for Opinion Mining", International Journal of Innovations in Engineering and Technology (IJIET) Vol. 2 Issue 4, pp 171-177.

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

[17] Whitelaw, C., Garg, N., and Argamon, S. (2005), "Using appraisal groups for sentiment analysis", In Proceedings of the 14th ACM Conference on Information and Knowledge Management, pp. 625-631.

Dr. 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 pursuing Doctor of Science at Utkal University, orissa, India. 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 75 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 has visited countries like Czech Republic, Austria, Thailand, United Kingdom, Malaysia, U.S.A, and Singapore. He is an active Member of various professional bodies and Editorial Board Member of various conferences and journals.