An Improved K-nearest-neighbor algorithm using Genetic Algorithm for Sentiment Classification

P.Kalaivani
Research Scholar
Department of CSE,
Sathyabama University, Chennai, India.
vaniraja2001@yahoo.com

K.L.Shunmuganathan
Professor & Head
R.M.K Engineering College
Chennai, India

***Abstract* — Sentiment classification is to find the polarity of product or user reviews. Supervised machine learning algorithms are used for opinion mining such as Navie Bayes, K-nearest neighbor and Support vector machine. KNN is simple algorithm but less efficient classification algorithm. In this paper we propose an improved KNN algorithm, genetic algorithm is developed which is a hybrid genetic algorithm that incorporates the information gain for feature selection and combined with KNN to improve its classification performance. Specifically, we compared other supervised machine learning approaches such as Navie Bayes and traditional kNN for Sentiment Classification of movie reviews and book reviews. The experimental results using genetic algorithm with improved indicate high performance levels with Fmeasure of over 87% on the movie reviews.**

Keywords—Opinion mining ,sentiment classification, machine learning algorithm, genetic algorithm, features selection.

I. INTRODUCTION

A basic task in sentiment classification/analysis is classifying the polarity of a given text at the document, sentence, or feature level whether the expressed opinion in a document, a sentence or an entity feature is positive, negative, or neutral. Sentiment analysis is one of the applications of natural language processing, and text analytics to identify and extract subjective information in source materials. It aims to determine the attitude of a writer with respect to some topic or the overall polarity of a document [1]. The attitude may be his/her judgment, affective state or the intended emotional communication.

We focus in particular on movie reviews for several reasons. There are several on-line movie review data bases. For instance, IMDb (Internet Movie Database) contains thousands of movie reviews. Cornell offer a large data base of movie reviews which they have labeled by sentiment. Movie reviews have a clear indication of sentiment such as "thumbs up" or "****" from which it is easy to gather labels [4].

The aim of the sentiment classification is to find the polarity of opinion of a writer with respect to product reviews, movie reviews or topics (Suge Wanga,Deyu Li, Lidong Zhao,Jiahao Zhang Wang , 2013). In this paper, we propose an improved KNN algorithm for sentiment classification.KNN is simple but less effective classification algorithm than SVM and NB. Genetic algorithm is combined with K-nearest neighbor algorithm to improve the performance and overcome the limitations of traditional KNN.

The proposed approach is based on supervised machine learning approach that uses information gain feature selection technique and TF-IDF weighting scheme along with optimized feature selection method. Our main objective is to design and develop a new classification algorithm able to improve the performance. We compared with other supervised machine learning algorithm such as NB, traditional KNN without optimized feature selection method.

The rest of the paper is organized as follows. Section 2 presents state-of-the-art, related to this study. Section 3 gives models and methodology. Section 4 presents evaluation models used in this study. In section 5 we discuss the empirical results and section 6 gives conclusion of the study and future research direction of this study.

II. RELATED WORK

The main objective of the feature selection is to reduce the number of features, the computational cost and improve the performance of classification. It has been proved that feature selection method is to remove the irrelevant and redundant feature and also increase the learning task, so it improves the efficiency of sentiment classification. In the movie review dataset, only few attributes gives useful information to the classifier. Several machine learning techniques were used for sentiment classification tasks in history. The following few works are selected to this study.

Earlier studies in sentiment classification several machine learning algorithms were analyzed on a movie review data set [2, 4] different feature selections techniques are used. The objective of the feature selection is to decrease the dimensionality of the feature space. Features are words or bigrams of works Pang & Lee [2, 3], achieved best result using SVM based in unigram. They utilized Naivebays (NB), Maximum entropy (ME) and Support vector machines (SVM). As per the results, on the movie review data set 82.9% accuracy was achieved, while the NB method gave lower accuracy. To improve the result of NB, Pang & Lee [4]

proposed first subjective sentences from the rest of the sentiment classification documents.

Ye, Q., Zhang, Z., and Law, R. [5], proposed sentiment classification of reviews and applied three supervised machine bearing algorithms of SVM, Navie Bayes, and character based N-gram model for sentiment classification of the reviews and they reported that all three approaches reached accuracies of at least 80% and also that SVM and N-gram approaches outperformed the Navie Bayes approache. Long-Sheng Chen and Hui-Ju Chiu [6], reports an evaluation machine learning algorithm combines proposed a Neural Network (NN) and semantic orientation indices to effectively classify sentiment.

Sentiment classification approach based on latent semantic analysis (LSA) to identify product features was adopted by Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, Gen-Chi Lu [7] and proposed a system called a movie-rating and review-summarization system in a mobile environment.

Pattoglou and Thekuall [8], proposed, TF-IDF weighting schemes combined with SVM classifier gave the best result. This solution achieved a significant improvement over the previous study.Rui Xia , Chengqing Zong and Shoushan Li [9], applied and applied Heuristic search-enhanced markov blanket model together with SVM for opinion mining from online text.

Kang, H., Yoo, S. J and Han, D. [10] used Navie Bayes, they reported that NB gave best result on restaurant reviews and obtained 83.6% accuracy on more than 6000 documents. In Annett and kondrak [11], approach based on lexical wordNet together with NB, SVM and decision tree for sentiment classification. They reported an performance accuracy of greater than 75%.

So, our focus in this work is to make an intensive study of the effectiveness of feature set reduction using optimized feature selection together with Information Gain (IG), for sentiment classification of movie reviews and book reviews.

III. METHODOLOGY

Sentiment analysis is conducted at any of the three levels: the document level, sentence level and the attribute level. In this study, we applied three supervised machine learning models for sentiment classification of reviews for the selected movie reviews. These models are Naive Bayes (NB), K-Nearest neighbourhood and an improved KNN algorithm together genetic algorithm.

In this work, support vector machine classification algorithm is applied to classify the documents and find a set of opinion as positive or negative. It has been shows that effective classification algorithm and used in sentiment analysis, among all classification algorithm, SVM outperformed NB and kNN. The SVM finds hyper plane using support vectors. This approach was developed by Vladimir vapnik, Bernhand Boser and Isabelle Guyan in 1992.

In the two category problem, the basic idea is to find a maximum margin hyper plane, represented b vector w, that not only separates the document vectors in one class from other class, but for which the separation is as large as possible.

A. Navie Bayess

The basic idea is to find the probabilities of categories given a text document by using the joint probabilities of words and categories. It is based on the assumption of word independence.

The starting point is the Bayes theorem for conditional probability, stating that, for a given data point x and class C:

Furthermore, by making the assumption that for a data point x = {x1, x2,…xj}, the probability of each of its attributes occurring in a given class is independent, when we estimate the probability of x as follows

$$P(C/x) = P(C).\prod P(x/C) \quad (1)$$

Training a Naïve Bayes classifier therefore requires calculating the conditional probabilities of each attributes occurring on the classes, which can be estimated from the training data set. To provide sentiment classification of online reviews about movie Information Gain was selected as feature selection technique in this study.

B. Nearest Neighbor Method

The Nearest neighbor methods are considered one of the simplest and most yet effective classes of classification algorithms in use. The k-Nearest Neighbor algorithm works by inspecting the k closest instances in the data set to a new occurrence that needs to be classified, and making a prediction based on what classes the majority of the k neighbours belong to. The closeness is given by a distance function between two points in the attribute space. An example of distance function typically used is the standard Euclidean distance between two points in an n-dimensional space, where n is the number of attributes in the data set

C. Datasets collection

We collected review dataset about book, DVD, electronics, kitchen from the web site and also we collected movie review dataset from web site[1]. The Cornell movie-review corpora[1] consist of movie review dataset contains 1000 positive reviews and 1000 negative reviews. In this study we used movie reviews and book reviews.

The User's opinions are the valuable sources of data which helps to improve the quality of service rendered. Blogs, review sites and micro blogs are some of the platforms where user expresses his/her opinions.

To prepare the review documents we used tokenization to split the review text of a document into a sequence of words. We converted all characters to lowercase in the example set. Porter stemming and filter stop words are used to reduce the wordlist. English stop word are used to remove tokens which equal stop words from the stop word list. We used feature measure schemes as TF-IDF to convert text representation vector. To conduct the research, Movie reviews and book reviews are considered here.

[1]www.cs.cornell.eu/people/pabo/movie-review-data.

D. Feature Selection

Feature selection is a method of selecting subgroup of the features, which is available in document used for describing the dataset. In this study we applied Information Gain (IG) feature selection method to reduce the original feature set by removing irrelevant for sentiment classification.

G. Information Gain

Information gain is one of the important feature selection method used in sentiment classification, it outperformed than other feature selection method [3, 4, 5]. It is based on the value or weight of information content in reviews, which select important feature with respect to class attributes.

H. KNN Classification based on Genetic Algorithm

Genetic algorithm (GA) is a search and optimization technique. An optimization is a process of finding best or optimal solution for a sentiment classification.

a) Initial Population - In GA, the initial population of n strings are randomly generated, a collection of such strings is called initial population. The information gain feature weights are used as the final strings in the initial population. In the string collection, 1 represents a selected feature and 0 represents a discarded one. Generate random population of n individuals. Each attribute is switched on with the probability n_i. In this study, the population size is set to 50, and the probability is set to 0.1.

b) Selection (or) Reproduction - The selection process is the first operator applied on the population and perform crossover with the probability cross n_c. The probability crossover is set to 0.6 and tournament selection is one of the common techniques used in the selection scheme.

c) Crossover - Crossover is the process of exchange information between two parents and to produce a new offspring. In the population, certain adjacent string pairs are randomly selected for crossover based on the crossover probability n_c. Different crossover types are used such as single point, uniform and shuffle. We used uniform crossover. If the mixing ratio is 0.5, then half of the genes in the offspring will come from parent 1 and half will come from parent 2.

d) Mutation - This operator randomly mutates individual feature characters in a solution string based on a fixed probability P_m. The mutation probability is set to 0.01.

IV. PERFORMANCE EVALUATION

In this study, we used finite set of dataset, evaluation involves splitting the available dataset into a training set and a testing set. Generally, we applied the NB, kNN algorithm to the dataset in the training set and evaluate the resulting model using the dataset in the test set.

The cross validation method involves partitioning the dataset randomly into n-folds. We one partition as a testing set and use the remaining partitions to form training set. As before, we apply an algorithm to the training dataset and evaluate the resulting model and the testing set, calculating percent correct. We repeat this process using each of the partitions as the testing dataset and using the remaining partitions to form a training set. In this work, four evaluation measures, Accuracy, Precision, Recall and F-measure are used to test the effectiveness of opinion mining. R and P denote Recall and precision of reviews. Once we selected an algorithm and an evaluation methodology, we need to select a performance metric.

For two class problems, a test case will be either positive or negative. The performance element, when given a test case will predict either correctly or incorrectly. This yields four quantities that we can compute by applying a model to a set of test cases, as shown in Table1.

TABLE I. QUANTITIES COMPUTED FROM A TEST SET FOR A TWO-CLASS PROBLEM

	True Positive Reviews	**True Negative Reviews**
Predict Positive	A	B
Predict Negative	C	D

For a set of test cases, let A be the number of time the model predictive positive when example's label is positive, let B be the number of time the model predictive negative when example's label is positive, let C be the number of time the model predictive positive when example's label is negative, let D be the number of time the model predictive negative when example's label is negative. Given these counts, we can define a variety of common performance metrics. Accuracy is the portion of the test examples that the model correctly predicted as explained below.

$$Accuracy = \frac{A + D}{A + B + C + D} \tag{2}$$

$$RP = \frac{A}{A + C} \tag{3}$$

$$RN = \frac{D}{B + D} \tag{4}$$

$$PP = \frac{A}{A + B} \tag{5}$$

$$PN = \frac{D}{C + D} \tag{6}$$

$$FP(F1measure - Pos) = \frac{2 * RP * PP}{RP + PP} \tag{7}$$

$$FN(F1measure - Neg) = \frac{2 * RN * PN}{RN + PN} \tag{8}$$

TABLE II. COMPARISION BETWEEN THE K VALUES FOR THREE CLASSIFICATION ALGORITHM IN MOVIE REVIEWS

Classifier	Performance	K Value								
		1	**5**	**10**	**20**	**30**	**35**	**40**	**45**	**50**
KNN	Accuracy	62.00	58.50	59.50	56.50	56.00	55.50	55.00	55.00	55.00
	Precision	63.28	57.19	57.43	55.37	55.37	55.16	54.79	54.79	54.79
	Recall	74.36	94.36	99.00	99.09	100	99.09	100	100	100
	Fmeasure	68.37	71.22	**72.69**	71.30	71.28	70.87	70.79	70.79	70.79
GIKNN	Accuracy	75.00	74.00	79.50	81.00	81.00	83.55	84.50	82.00	86.00
	Precision	70.76	76.38	80.81	85.31	79.05	83.70	85.05	80.66	87.30
	Recall	92.55	77.09	83.55	80.36	89.55	90.00	88.91	92.73	88.09
	Fmeasure	80.20	76.73	82.16	82.76	83.97	86.74	86.94	86.27	**87.69 (↑)**
NB	Accuracy	79.00								
	Precision	78.29								
	Recall	87.18								
	Fmeasure	**82.49**								

TABLE III. COMPARISION BETWEEN THE K VALUES FOR THREE CLASSIFICATION ALGORITHM IN BOOK REVIEWS.

Classifier	Performance	K Value								
		1	**5**	**10**	**20**	**30**	**35**	**40**	**45**	**50**
KNN	Accuracy	58.34	60.26	59.24	67.39	60.76	63.79	62.29	64.84	60.76
	Precision	54.63	59.12	58.91	74.70	80.00	89.93	91.25	91.11	92.00
	Recall	98.00	88.00	68.78	57.67	28.00	33.11	28.11	35.33	23.00
	Fmeasure	70.15	**70.73**	63.46	65.09	41.48	48.40	42.98	50.92	36.80
GIKNN	Accuracy	70.39	73.45	75.84	80.45	78.92	78.89	80.97	81.42	80.97
	Precision	71.26	69.78	73.01	84.28	79.44	82.64	84.97	80.97	85.20
	Recall	75.78	83.00	86.67	79.78	78.89	72.67	77.00	86.89	75.00
	Fmeasure	73.45	75.81	79.26	81.97	79.16	77.33	80.59	**83.83(↑)**	79.78
NB	Accuracy	75.95								
	Precision	78.89								
	Recall	73.78								
	Fmeasure	**76.25**								

V. RESULT AND DISCUSSION

To evaluate our model, we used Cornell movie review datasets frequently used in the sentiment classification. The Cornell movie review dataset contains 1000 positive and 1000 negative documents [5]. In this study, we used 2000 movie review documents, it is a challenging task because the reviewers use a lot of comparisons, and sometimes used an unclear language. The performance result on Cornell movie review dataset and book review dataset with different k values are shown in Fig 1 and 2.

A. Comparison of classifiers

Many classification algorithms are available for sentiment classification such as SVM, NB, kNN, Maximum Entropy, Decision tree etc. In this study, we used three classification algorithms GIKNN, KNN and NB. Among all these methods GIKNN is shown to perform better. The best Fmeasure (%) for each classification is shown in bold with statistically significant achieve over the baseline with an up arrow.

B. Performance of different classifier method

In this study we use F1measure to evaluate our proposed approach on movie review dataset and multi domain dataset.

Information gain feature selection is used to reduce feature vector space and TF-IDF feature weighting schemes were utilized and selected top k percent attributes with highest weights are selected for training the classifier where the k value is 0.5. All the experiments were validated using 10 fold cross validation. Table [2-3] shows the experimental results when using three different classifier together genetic algorithm. It is an optimized feature reduction technique. Information gain is selected as a feature selection method, because it outperforms than other feature selection method. We have a comparative study with the performance result of GIKNN and KNN with different k values. Fig. 1 Summarize the result obtained F1measure applying 10 fold cross validation for movie review dataset. Fig. 2 Summarize the result obtained F1measure applying 10 fold cross validation for book review dataset.

The performance results are compared with different classifier such as NB, KNN. The GIKNN classifier result is much better than KNN classifier with different k values. The GIKNN achieve best result when k is 50, in this case the GIKNN, F1measure value is 87.69 %.

The KNN achieve best result when k is 10. The GIKNN classifier significantly outperforms KNN classifier in most of the different k values.In table 2 and 3, results show that GIKNN classifier performance much better than KNN and NB.

C. *Results of individual classifier*

We give the results of three classifiers (KNN, NB and GIKNN) with information gain as a feature selection method. We observe the performance of different classifiers. The results are presented in table [2-3]. We focus on the result of movie review data set and book review dataset.

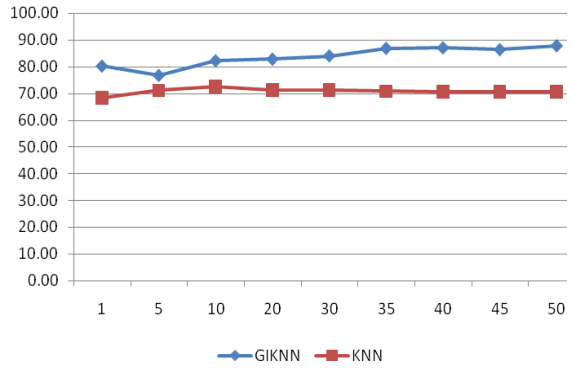

FIG 1 . THE CLASSIGICATION RESULT WITH DIFFERENT K VALUES IN MOVIE REVIEW

We conclude that GIKNN better on feature selection with IG and unigram feature weight scheme. We applied n-gram feature weighting schemes to generate the word vector space. In order to evaluate our system, we applied 10 fold cross validation for movie review corpus. If we compare our results with other similar work, we can find that our approach gave best results. For example, Pang & Lee applied different machine learning approaches. They found that SVM outperforms NB and ME. They obtained a maximum accuracy of 85.35% using trigram BO feature weighting scheme and 10 fold cross validation.

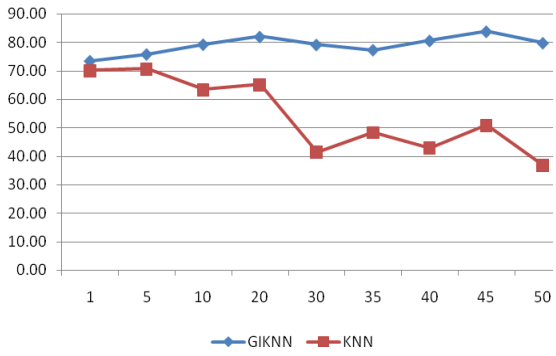

FIG 2 . THE CLASSIGICATION RESULT WITH DIFFERENT K VALUES IN BOOK REVIEW

In this paper we obtained performance result of 87.69% using GIKNNalgorithm with 10 fold cross validation.

VI. CONCLUSION

The main aim of the study is to improve the performance of opinion mining. We used movie review datasets and in multi domain dataset we used book review dataset**.** We proposed an improved KNN algorithm, genetic algorithm is developed which is a hybrid genetic algorithm that incorporates the information gain for feature selection and combined with KNN to improve its classification performance. The result shows that GIKNN approach outperformed the Navie Bayes and kNN approaches. A direction for future work is to study the performance of feature selection methods on different machine learning classifiers and to evaluate the model for cross domain sentiment analysis with other domain than movie reviews.

REFERENCES

[1] Suge Wanga,Deyu Li, Lidong Zhao,Jiahao Zhang, "Sample cutting method for imbalanced text sentiment classification based on BRC", Knowledge-Based Systems 37 (2013) 451–461.

[2] Pang, B., Lee, L. and Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10 (2002). Association for Computational Linguistics, 2002.

[3] Turney, P. D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings ofthe Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Philadelphia, Pennsylvania, 2002.

[4] Pang, B., & Lee, L. ,"A sentimental education: Sentiment analysis using subjectivity summarization based on Minimum Cuts". In *Proceedings of the 42nd Annual Meeting of the Association of Computational Linguistics*. 2004.

[5] Ye, Q., Zhang, Z., & Law, R. "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches ", Expert Systems with Applications, 36(3), 6527–6535., 2009.

[6] Long-Sheng Chen and Hui-Ju Chiu, "Developing a Neural Network based Index for Sentiment Classification", Proceedings of the International Multi Conference of Engineers and Computer Scientists, Hong Kong, pp 744-749, March2009

[7] Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, Gen-Chi Lu, And Emery Jou "Movie Rating and Review Summarization in Mobile Environment", IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, Vol. 42, No. 3, May 2012.

[8] Paltoglou, G., & Thelwall, M. ,"A study of information retrieval weighting schemes for sentiment analysis", In *Proceedings of the 48th Annual Meeting of the Association forComputational Linguistics*, pp. 1386-395.,2010

[9] Rui Xia , Chengqing Zong , Shoushan Li , "Ensemble of feature sets and classification algorithms for sentiment classification", Information Sciences 181, 1138–1152, 2011.

[10] Kang, H., Yoo, S. J., Han, D. Senti-lexicon and improved, "Naïve Bayes algorithms for sentiment analysis of restaurant reviews. Expert Systems with Applications", 2011.

[11] Annett, M. and Kondrak, G. "A comparison of sentiment analysis techniques: Polarizing movie blogs", Advances in Artificial Intelligence, 5032:25–35,2008.