ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Efficient Sentiment Analysis Using Hybrid PSO-GA Approach

Archana Sonagi[1], Deipali Gore[2]

PG Student, Dept. of Comp. Engineering, PESs MCOE, Savitribai Phule Pune University, Pune, Maharashtra, India

Asst. Prof., Dept. of Comp. Engineering, PESs MCOE, Savitribai Phule Pune University, Pune, Maharashtra, India

ABSTRACT: The utilization of web 2.0 as a platform has made the web a treasure trove of sentiments. People around the world express and share their opinions and reactions on web regarding day-to-day activities and global issues as well thus making a huge amount of intelligent data available to all. With this vast wealth of data arises the need for automatic opinion classification. Sentiment classification using machine learning algorithms faces the problem of high dimensionality of feature space. Therefore, a feature selection method is used to eliminate the uncounted features from the word vector space for efficiency improvement to machine learning approaches. In this paper, we intend to apply GA and swarm optimization (i.e., PSO) technique to optimize the feature selection. We exemplify our proposed method on the online movie reviews using sentiment analysis approach. SVM is proposed for sentiment classification. From experimental results it can be ascertained that combined approach i.e., PSO-GA gives better classification accuracy compared to PSO-based method. PSO provides the advantages in providing the solution to discrete problems.

KEYWORDS: Feature Selection, Genetic algorithm, Particle Swarm Optimization, Sentiment Classification, Support Vector Machine.

I. INTRODUCTION

With the advent of technology web today has become a treasure trove of sentiments. Millions of people around the world share and discuss their opinions and knowledge regarding various topics under the sun every day. Thus, this has led to a vast amount of intelligent and useful data to be available on the web. To make this data to be of use to all a need to filter and classify this data, and to extract the various emotions and opinions from the data available, becomes necessary. Sentiment analysis, also called opinion mining, is the field of study that analyses people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [1]. Sentiment Analysis is the process of determining the feeling or opinion of a piece of text. Automatic sentiment classification is an active research topic in the field of information retrieval and data mining, since the results are still subject to improvements. A large amount of resources such as social networking sites, blogs, review sites, online shopping sites, have a huge amount of reviews available. This data sources has become a gold mine for organizations to monitor their brand and reputation by extracting and analysing the sentiment of tweets posted by the public about them, their markets and competitors. So this gives a motivation to categorize the reviews in an automated way by a property other than topic, namely by what is called their sentiment or polarity. Reading all those reviews is time-consuming, but, if only few reviews were read, the evaluation would be biased. Sentiment analysis thus aims to solve this problem by automatically classifying user reviews into positive or negative opinions [2].

Without prior knowledge, it is difficult to determine which features are useful. As a result a large number of features are usually introduced to the dataset which include relevant, irrelevant and redundant features. Irrelevant and redundant features are not useful for classification and they even reduce the classification performance due to a large search space known as curse of dimensionality [3]. Feature selection can address this problem by selecting only relevant features for classification. Feature Selection can reduce the number of features, lessen the training time, simplify the learned classifiers and or improve the performance of classification. Sentiment classification using machine learning algorithms faces the problem of high dimensionality of feature space. Therefore, a feature selection method is used to eliminate the uncounted features from the word vector space for efficiency improvement to machine learning approaches.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Among too many methodologies, which are proposed for sentiment classification, Particle Swarm Optimization (PSO) based method have attracted a lot of attention. PSO has advantages in providing the solution to solve complex optimization problems mainly in discrete optimization problem. PSO is encouraging algorithm in data classification and clustering. PSO originated from the idea of swarm intelligence and Evolutionary Computation technique. It is inspired by the ability of flock of birds to find rich sources of food, and avoid predators by implementing an information sharing mechanism. It helps a great deal in selection of optimal feature subset.

High dimensionality of feature space is yet another major problem in sentiment classification. Many more features of the original feature set are irrelevant to sentiment classification, which increase the noisy data and lay impact on the overall performance of the classifier. So we need to select the subset from the original feature set to reduce the dimensionality of feature space which improves the efficiency and performance of the classifier. There were several approaches applied to solve the problem of feature selection in sentiment classification. Few of the most commonly used feature selection techniques include Chi-Square, Information Gain, Term Frequency, Term Presence, TF-IDF, PSO, ACO algorithms etc. Among all the approaches which are proposed for feature selection, genetic algorithm (GA) which is population-based method and Particle Swarm Optimization (PSO) based method have gained more attention. In our proposed approach we are considering combined PSO-GA method for feature selection which selects the most representative features for sentiment classification. The main contribution of this paper is implementation of a hybrid approach of particle swarm optimization and genetic algorithm to select optimal feature subset to increase the performance and efficiency of SVM classifier.

II. RELATED WORK

Feature selection has long been a fertile field of research and a vast literature exists on the various techniques of feature selection. PSO and GA have been used to promote the optimal subset for feature selection and also to tune SVM's parameters to increase the performance of classification in different practical fields like Facial and clothing information for gender classification, diagnosis of digital mammogram, VLSI Floor planning using smart decision making, Palm print Recognition[4-6].

EC techniques have been applied to address feature selection problems. Among many methods which are proposed for FS, heuristic methods such as genetic algorithm, ant colony optimization and particle swarm optimization have been of interest for researchers. These methods try to gather better solutions by using knowledge from previous steps. GA's are optimization methods based on the natural selection. They applied operations found in natural genetics to guide search in the search space. Because of their advantages, GA has been widely used as a tool for FS in data mining. Particle swarm optimization which is introduced by Kennedy and Eberhart is based on a social-psychological model of social influence and social learning. Particles in a swarm follow a very simple behaviour: emulate the success of neighbouring individuals. The emerged group behaviour discovers optimal regions of a high dimensional search space.

Bing et. al [2] presents the first study on multi-objective particle swarm optimization (PSO) for feature selection. The task is to generate a Pareto front of non dominated solutions (feature subsets). Xue et. al [7] proposed a multi-objective filter feature selection algorithm based on binary particle swarm optimization (PSO) and probabilistic rough set theory in which the results showed that the proposed algorithm can automatically evolve a set of non-dominated feature subsets. Chen et. al [8] proposed a regression based particle swarm optimization for feature selection problem. The proposed algorithm can increase population diversity and avoid local optimal trapping by improving the jump ability of flying particles. Nazir et. al. [9] proposed an efficient gender classification technique for real-world face images (Labeled faces in the Wild). In this work, they extracted facial local features using local binary pattern (LBP) and then, fused these features with clothing features, which enhanced the classification accuracy rate remarkably. In the further steps, particle swarm optimization (PSO) and genetic algorithms (GA) were combined to select the most important features' set which more clearly represent the gender and thus, the data size dimension gets reduced. In this paper Sheikhalishahi et. al. [10] proposes a novel hybrid genetic algorithm (GA)-particle swarm optimization (PSO) approach for reliability redundancy allocation problem (RRAP) in series, series parallel, and complex (bridge) systems. The proposed approach maximizes overall system reliability while minimizing system cost, system weight and volume, simultaneously, under nonlinear constraints. To meet these objectives, an adaptive hybrid GAPSO approach is developed to identify the optimal solutions and improve computation efficiency for these NP-hard problems. An illustrative example is applied to show the capability and effectiveness of the proposed approach. According to the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

results, in all three cases, reliability values are improved. Moreover, computational time and variance are decreased compared to the similar studies. In this paper Li at. al.[11] proposed an SVM classification system based on particle swarm optimization (PSO) is proposed to improve the generalization performance of the SVM classifier. Authors have optimized the SVM classifier design by searching for the best value of the parameters that tune its discriminant function. The experiments are conducted on the basis of benchmark dataset. Results confirm the superiority of the PSO-SVM approach.

B. Particle Swarm Optimization

Particle Swarm Optimization is an optimization technique proposed by Kennedy and Eberhart in 1995. PSO is inspired by the ability of flocks of birds to find rich sources of food, and avoid predators by implementing an information sharing mechanism. In fact they used the mechanism of birds flocking to solve optimization problems. It means that a group of particles search the solution space for the best solution. Each particle has a position, velocity, and a memory to save its best position from the beginning of the process. PSO is a global optimization technique which is known for its speed of convergence, easy to implement and only few parameters to adjust but it has the drawback that it may quickly cause a particle to stagnate and also prematurely converge on suboptimal solution.

PSO is based on the principle that each candidate solution can be represented as a particle in the swarm. Each particle has a position in the search space. These particles move in the search space to search for the optimal solutions. Therefore, each particle has a velocity, which is represented as Vi. During the movement, each particle updates its position and velocity according to its own experience and that of its neighbors. The best previous position of the particle is recorded as the personal best pbest, and the best position obtained by the population thus far is called gbest. Based on pbest and gbest, PSO searches for the optimal solutions by updating the velocity and the position of each particle according to eq. given below. The algorithm stops when a predefined criterion is met, which could be a good fitness value or a predefined maximum number of iterations.

$$V_{ij}{}^{r+1} = V_{ij}{}^{r+1} + C_1\ rand_1\ \left(pbest_{ij} - X_{ij}{}^{r}\right) + C_2\ rand_2\ \left(gbest_{ij} - X_{ij}{}^{r}\right)$$

$$X_{ij}{}^{r+1} = X_{ij}{}^{r} + V_{ij}{}^{r+1}$$

C. Genetic Algorithm

GA, proposed by Holland (1992), is a stochastic algorithm that mimics natural evolution. The most distinct aspect of GA is that it maintains a set of chromosomes (i.e. solutions) in a population. As in the case of biological evolution, it has a mechanism of selecting fitter chromosomes at each generation. To simulate the process of evolution, the selected chromosomes undergo several genetic operations, e.g. crossover and mutation. The evaluation stops when the termination criteria is met.

III. PROPOSED ALGORITHM

A. Document Preprocessing:

The collection of reviews set provided as an input to the system must be pre-processed as it contains data that are redundant, unambiguous, noisy and irrelevant for sentiment classification. The pre-processing includes data cleaning, tokenization, stop word removal, stemming, n-gram generation and term weighting. In this literature we find that the commonly used classification algorithms are quite slow in terms of processing time as well as the classification rate. In our proposed system we are applying a combination of PSO and GA for feature selection which will give more precise subset of features for sentiment classification. To determine the review document category, we adopt the algorithm of SVM classification suggested in [2]. Our choice is inspired by the flexibility of the metaheuristics to find optimal feature subsets that are common to NP hard. In general, a sentiment classification system includes numerous essential like feature extraction and feature selection. Once the review documents are pre-processed, feature extraction is applied to convert the input text document into a feature vector. For dimensionality reduction feature selection is applied to the feature vector. The overall process of proposed system is depicted in Fig.1. and explained in the following steps.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

1) Tokenization: It is the process of breaking a stream of text into meaningful words or phrases called tokens. The words in the review here are split based on the special delimiting characters such as a space and a line break. The list of tokens becomes input for further process of parsing.
2) Stop Word Removal: Stop words are frequently occurring words like pronouns, prepositions and conjunctions that do not provide any additional improvement in performance but increase the computational complexity by increasing the size of the dictionary. These words from the text documents are having a very low discriminative value. It includes creating a list of stop words and then scanning the tokens to remove the stop words that have occurred. These common stop words are collected from [8].
3) Stemming: It is the process of finding the root morphemes of the token words. For example, the words "purification", "purity", "purify" and "purifying" have a stemmed root as "pure". Stemming words helps to achieve efficient reduction in the dimensionality of the feature space. To stem the tokens the Porter stemming algorithm is used, which is a natural language processing (NLP) task by removing the suffix, to narrow down the size of the feature space.
4) N-grams: An n-gram is a contiguous sequence of n items from a given sequence of text. Generally information about any sentiment is conveyed by use of adjectives along with other parts of speech. By using unigrams, bigrams and trigrams we were able to capture information about adjectives and adverbs. Using bigrams and trigrams require a huge amount of training set available.
5) Term Frequency: It is a measure of how frequently a term showed up in a document. Term Frequency studies such information which may be beneficial in selection of essential features.
6) Term Weighting: TF-IDF is a term weighting approach which is one of the widely used methods to evaluate the importance of a term in the corpus or it identifies how relevant a term is to the classification. It can be calculated with the help of formula (2) as follows.

$$W(i,j) = tf_{ij} . \frac{N}{df_{ij}}$$

B. Feature Extraction

It is the process of converting the text feature into feature vector. For the representation of text we are going to use the count vectorizer which represents the documents in the form of matrix.

C. Feature Selection

Feature selection ranks terms by considering their presence and absence in each class. A high score is assigned to term that occur frequently in a class. Feature selection is used for dimensionality reduction of original feature set to get the more relevant feature space for classification. In our proposed system we used the combination of PSO and GA proposed in [4]. Each particle represents a solution, which denotes the selected subset of features and parameter values. The selected features, parameter values and training dataset are used for building SVM classifier models.

D. Classification Method

In the proposed system we adopt the algorithm of Support Vector Machine (SVM) suggested in [2] for the classification of reviews. SVM takes data (features) as input and predicts based on training data set to which class these features belong. The goal of SVM is to find the optimal hyperplane such that the error rate is minimized for an unseen test sample. Once we get the precise set of features, which is outcome of the proposed PSO-GA approach for feature selection, on this feature set we will apply the SVM based approach for sentiment classification. There are several techniques available for sentiment classification. The proposed approach will give better results for classification of review documents into its correct category than other approaches thereby improving the efficiency and performance of the classification.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

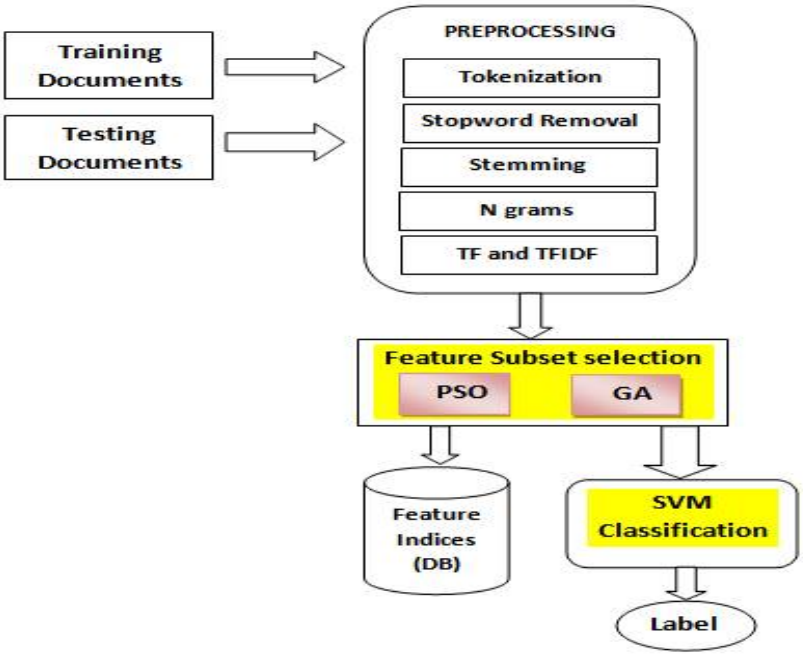

Figure 1: Proposed Approach

F. Evaluation of Classification

The performance of the classifier is evaluated according to the accuracy results. In order to compare the predicted categories assigned by the classifier with the actual categories of the test documents, first of all the number of True Positives, False Negatives and False Positives are determined, then precision, recall and accuracy is computed using these values.

IV. SIMULATION RESULTS

A. Dataset Used:

To evaluate the effectiveness of the sentiment classification algorithms, several standard datasets are available. These collections are useful for research in Information Retrieval, Natural Language processing and other corpus-based research. For evaluating the algorithmic approach that we introduce in this paper, we have selected the training dataset and test dataset as reviews for movie domain. We are filtering the reviews for removal of noise and then providing it as an input to the system. Pre processing steps are applied on the review documents and then stored in the database. And then further steps from proposed system are applied on this collection for sentiment collection to classify the document depending upon the emotion of the opinion. The dataset used in our experiments is twitter movie reviews from `trainingandtestdataset.zip`

B. Performance Measure:

To evaluate the performance of the classifier the important tool used is the confusion matrix. Confusion matrix is used to describe the performance of the classification model (or classifier) on a set of test data for which the values are previously known.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

	Document belonging to the category	Document not belonging to the category
Category assigned to the document by the classifier	TP	FP
Document category rejected by the Classifier	FN	TN

Table 1 : Confusion Matrix

In information retrieval systems precision, recall and accuracy are the most used measurements to evaluate the performance. According to the Table 2, precision and recall are defined as follows,

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + FN}{N}$$

C. Result:

For analysing the performance of PSO and PSO-GA algorithms for feature selection we have considered following confusion matrix for test documents. Analysing the precision, recall and accuracy shown in Table V, we see that on average, PSO-GA algorithms obtained a higher accuracy value than PSO alone.

Method Used	Precision (%)	Recall (%)	Accuracy (%)
PSO	81.20	75.17	76.20
PSO – GA	84.70	72.30	80.87

Table 2: Evaluation Table

The results show that as the percentage of selected features exceeds 4% in accuracy measures, the PSO-GA algorithm outperforms PSO algorithm.

V. CONCLUSION AND FUTURE WORK

In this paper a hybrid combination of PSO and GA, feature selection algorithm for sentiment classification is presented. In the proposed algorithm, the classifier performance and the length of the selected feature subset are adopted as heuristic information. Therefore, it can select the optimal feature subset without the prior knowledge of features. Proposed algorithm has the ability to converge quickly; it has a strong search capability in the problem space and can efficiently find minimal feature subset. Experimental results demonstrate competitive performance. Proposed PSO-GA algorithm is compared with PSO algorithm alone for text sentiment classification. In order to evaluate the performance of proposed algorithm, experiments were carried out on the dataset in the literature, i.e. movie reviews.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

The accuracy of SVM before using the combination of feature selection methods and after using the combination PSO-GA for feature selection has shown an increase. The accuracy increase obtained is of approximately 4%. The computational results indicate that the proposed algorithm outperforms PSO, since it has achieved better performance with lower number of features.

REFERENCES

1. Bing Liu, 2012, Sentiment analysis and opinion mining, Morgan and Claypool publishers.
2. Abd. Samad Hasan Basaria,*, Burairah Hussina, I. Gede Pramudya Anantaa, Junta Zeniarja, "Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization", Elsevier 2013.
3. Bing Xue, Mengjie Zhang and Will N. Browne, "Particle Swarm Optimization for Feature Selection in Classification: A Multi-Objective Approach", IEEE Transactions on Cybernetics 2012.
4. Nazir, A. Majid-Mirza1, S. Ali-Khan ,"PSO-GA Based Optimized Feature Selection Using Facial and Clothing Information for Gender Classification", Journal of Applied Research and Technology, 2014.
5. M. Sheikhalishahi, V. Ebrahimipour, H. Shiri,H. Zaman and M. Jeihoonian ,"A hybrid GAPSO approach for reliability optimization in redundancy allocation problem", Springer 2013.
6. M. Nazir, A. Majid-Mirza1, S. Ali-Khan, "PSO-GA Based Optimized Feature Selection Using Facial and Clothing Information for Gender Classification", Journal of Applied Research and Technology, 2014.
7. Liam Cervantes, Bing Xue, Lin Shang, and Mengjie Zhang, "A Multi-objective Feature Selection Approach Based on Binary PSO and Rough Set Theory", Springer 2013.
8. Kun-Huang Chen , Li-Fei Chen , Chao-Ton Su , "A new particle swarm feature method for sentiment classification", Springer 2013.
9. Jun Li and Bo Li, "Parameters Selection for Support Vector Machine Based on Particle Swarm Optimization", Springer 2014.
10. Kun-Huang Chen , Li-Fei Chen , Chao-Ton Su, "A new particle swarm feature method for sentiment classification", Springer, 2013.
11. Jyoti Ahuja, Saroj Dahiya Ratnoo, "Feature Selection using Multi-objective Genetic Algorithm: A Hybrid Approach", INFOCOMP 2015.
12. Iman Behravan, Oveis Dehghantanha, Seyed Hamid Zahiri, and Nasser Mehrshad," An Optimal SVM with Feature Selection Using Multi objective PSO", Journal of Optimization, 2015.
13. Mehdi Hosseinzadeh , Aghdam and Setareh Heidari., "Feature Selection Using Particle Swarm Optimization in Text Categorization", JAISCR, 2015.

BIOGRAPHY

Archana Sonagi received the B.E. degree in Information Technology Engineering from MGM's JNEC, BAMU University in 2012. She is now pursuing M.E. degree in Computer Engineering from P.E.S.'s Modern College of Engineering, Savitribai Phule Pune University, Pune and her area of interest is Data Mining and Information Retrieval.

Deipali Gore received the M.E degree in Computer Engineering from D.Y. Patil, Akurdi College under Savitribai Phule Pune University in 2006. Her current research area includes Information Retrieval and Web Mining. She is currently working as an Assistant Professor in P.E.S.'s Modern College of Engineering, Savitribai Phule Pune University, Pune.