Research Paper

SENTIMENT CLASSIFICATION USING SVM AND PSO

*Dr. Nirmala Devi K and Dr. Jayanthi P

Address for Correspondence

Department of CSE, Kongu Engineering College, Erode 638052, India

ABSTRACT

The growth of social network contributes huge quantity of user generated content like client reviews, comments and opinions. While this content will be helpful for decision making and analyzing this bulk of user generated content is difficult as well as time consuming. So there is a necessity to develop an intelligent system that automatically mine such vast content and classify them into positive and negative class. Sentiment analysis is useful in social media monitoring to automatically characterize the overall feeling or mood of consumers as reflected toward a specific brand or company and determine whether they are viewed positively or negatively on the web. This new kind of analysis has been widely addressed in customer relation management especially in the context of complaint management. For automating the task of classifying a single topic textual review, document-level sentiment classification is used. The document level classification approximately classifies the sentiment using SVM algorithm. The proposed SVM-PSO method obtains better result than the SVM on the benchmark dataset of Movies reviews dataset.

KEY WORDS: Sentiment Classification, SVM, PSO, Text mining, Social Media.

1. INTRODUCTION

Text Mining is used to extract previously unknown information from different written resources. A key element is used to link together the extracted information to form new facts or new hypotheses to be explored further by more conventional means of experimentation.

Sentiment analysis is the process used to determine the attitude or opinion or emotion expressed by a person about a particular topic. Sentiment analysis or opinion mining uses Natural Language Processing (NLP) and text analytics to identify and extract subjective information in source materials. The rise of social media such as blogs and social networks has fuelled interest in sentiment analysis [1][2]. In order to identify the new opportunities and to manage the reputations, business people usually view the reviews or ratings or recommendations and other forms of online opinion. This allows to not only find the words that are indicative of sentiment, but also to find the relationships between words so that both words that modify the sentiment and what the sentiment is about can be accurately identified. Scaling system is used to determine sentiment for the words having a positive, negative and neutral sentiment. It also analyses the subsequent concepts to understand the words and how they relate to the concept.

There are two different types of learning. One of the learning is supervised and the other is unsupervised learning. The supervised learning classifiers perform learning classifier during training and assign class labels to test data. The unsupervised learning performs learning without training data. There is a hybrid training called as semi-supervised learning, which uses both labelled and unlabeled training data. The sentiment learning uses machine learning or lexicon based learning. The lexicon is a sentiment bearing words and which is built with the dictionary based approach. The lexical resources WordNet and SentiWordNet are helpful for building such lexicons[3].

Text Classification (TC) [4][5] is one of the prime techniques to deal with the textual data. TC systems are used in a number of applications such as, filtering email messages, classifying customer reviews for large e-commerce sites, web page classification for an internet directory, evaluating exams paper answers and organizing document databases in semantic categories. Opinion mining is a hot research in the part of text mining that has been specific by different conditions such as sentiment analysis or sentiment orientation [6].

In traditional model the sentiment is considered as a binary classification [7], [8], [9], [10], [11]. There are different machine learning methods were used for sentiment classification such as Naïve Bayes, Support Vector Machines, Maximum Entropy, J48 etc. Among several SVM maintained to do the best. The proposed system uses SVM-PSO to perform sentiment classification.

The rest of this paper is organized as follows. The detailed proposed approach is discussed in Section 2.The experimental results and performance analysis are discussed in Section 3.The Section 4 contains the conclusion of the paper.

2. MATERIALS AND METHODS

Vapnik [6] proposed SVM are a group of supervised learning methods that performs classification by constructing an N-dimensional hyper plane that optimally separates the data into two categories. SVM [7], [12] has been shown to perform very good on a wide variety of classification problems that require large scale input space, such as handwritten character recognition, face detection, and most importantly in this case, text categorization.

SVM has been widely used in many text classification problems due to their major benefits as follows. It is robust in high, suitable for any function and moreover SVM gained great results in sentiment mining also. Even though it is novel and performs best in many applications, the practicality of SVM is impacted due to the problems of choosing suitable parameters of SVM (C,$\sigma$and ε) [13][14][15]. The PSO is based on the swarm intelligence optimization technique and is very simple to implement.

The proposed system aims to classify sentiment of movie review standard data set using PSO to tune parameters of SVM and accuracy of SVM PSO is better than SVM. The sentiment analysis is helpful to predict the stock market investment as well as movie box office prediction. The hotspot forums were also to be detected with the sentiments.

2.1 Data Set

The data set used in the proposed system is the Movie Reviews Dataset and Table 1 characterizes the attributes after preprocessing. An overview of steps and techniques commonly used in sentiment classification approaches, as shown in Figure 1.

Table 1: Movie Review Data set

Number of Positive Reviews	Number of Negative Reviews	Total Reviews	Number of distinct terms
1000	1000	2000	25162

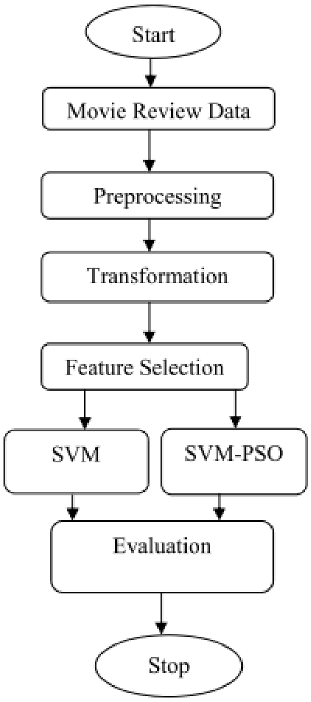

Figure 1: Proposed System

2.2 Preprocessing

The text pre-processing techniques are divided into two subcategories:

Tokenization: Textual data comprises block of characters called tokens. The documents are separated as tokens and used for further processing.

Removal of Stop Words: A stop-list is the name commonly given to a set or list of stop words. It is typically language specific, although it may contain words. A search engine or other natural language processing system may contain a variety of stop-lists, one per language, or it may contain a single stop-list that is multilingual. Some of the more frequently used stop words for English include "a", "of", "the", "I", "it", "you", and ”and” these are generally regarded as 'functional words' which do not carry meaning. When assessing the contents of natural language, the meaning can be conveyed more clearly by ignoring the functional words. Hence it is practical to remove those words which appear too often that support no information for the task. If the stop word removal is applied, all the stop words in the particular text file will not be loaded. If the stop word removal is not applied, the stop word removal algorithm will be disabled when the dataset is loaded.

2.3 Text Transformation

The score of each sentence in the source document is calculated by sum of weight of each term in the corresponding sentences. The weight of each term is calculated by multiplication of TF and IDF of that word based on adjective word extracted from Parts of speech tags. The TF and IDF are defined in Equation (1).

$$W(t,d) = TF(t,d) * IDF(t)$$
$$= TF(t,d) * \log\left[\frac{D}{DF(t)}\right] \quad (1)$$

2.4Feature Selection

Many statistical feature selection methods for document level classification can also be used for sentiment analysis. The simplest statistical approach for feature selection is to use the most frequently occurring words in the corpus as polarity indicators. The majority of the approaches for sentiment analysis involve a two-step process:

- Identify the parts of the document to contribute the positive or negative sentiments.
- Join these parts of the document in ways that increase the odds of the document falling into one of these two polar categories.

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.

2.5 SVM

The basic concept of SVM regression is to map non-linearly the original data ‘x’ into a high-dimensional feature space and to solve a linear regression problem in the feature space. The SVM regression function is represented by Equations (2).

$$K(x_i, x) = exp(-\|x_i - x\|^2/2\sigma^2) + \frac{C\varepsilon}{2\sigma^2} \quad (2)$$

The C and σ are in the range [10^{-3}, 10^{+3}] and ε is in the ranges [0, 1].

2.6 PSO

The working of PSO is governed by velocity vector and position vector with the Equations (3) and (4) given below. The process of PSO is also represented in Figure 2.

$$\mathrm{V_{ij}}^{r+1} = \mathrm{w\,V_{ij}}^{r} + \mathrm{C_1\,rand_1}\left(\mathrm{pbest_{ij}} - \mathrm{X_{ij}}^{r}\right) + \mathrm{C_2\,rand_2}\left(\mathrm{gbest_{ij}} - \mathrm{X_{ij}}^{(r)}\right) \quad (3)$$

$$\mathrm{X_{ij}}^{r+1} = \mathrm{X_{ij}}^{(r)} + \mathrm{V_{ij}}^{r+1} \quad (4)$$

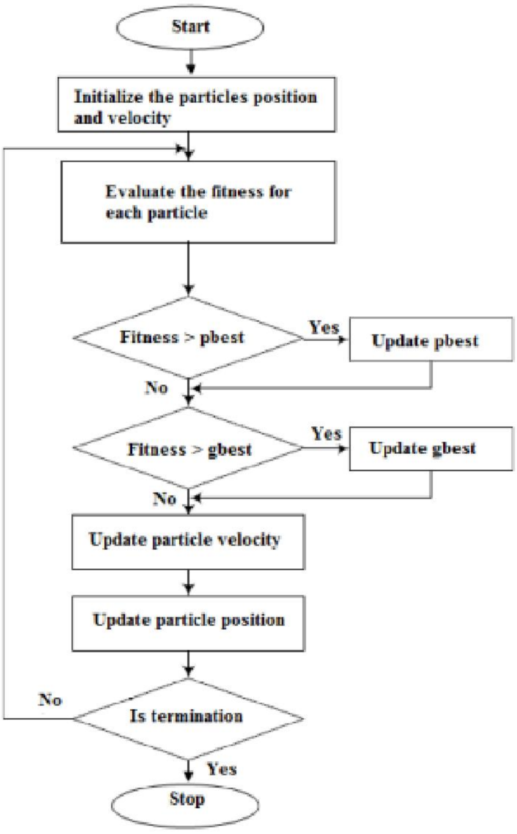

Figure 2: The Process of PSO

2.7SVM - PSO

In the proposed SVM-PSO, PSO is used to select the best features C,σ and ε in SVM. Then, feature subset selection and parameter values determination are performed. 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. Forecasting can be carried out using SVM-PSO algorithm.

3. RESULTS AND DISCUSSION

Table 2: Parameters of SVM and SVM-PSO

PSO Parameters	SVM Parameters

Particles = 30
Iterations=200
	C = [10^{-3}, 10^{+3}]
	With initial value [-10,10]

C1= 2.3, C2 = 1.8 rand1,rand2 =[0,1]	σ = [10^{-3}, 10^{+3}] With initial value [-10,10]
w from 0.9 to 0.4	ε = [0,1]
Fitness function = Accuracy	

The initial parameters for PSO and SVM are shown in the Table 2. The measures used in the proposed system evaluations are Precision, Recall and Accuracy are defined in Equations (5), (6) and (7). Figure 3 shows the comparison results between SVM and SVM-PSO.

$$\text{Precision (P)} = \frac{\text{TP}}{\text{TP+FP}} \qquad (5)$$

$$\text{Recall (R)} = \frac{\text{TN}}{\text{TP+FN}} \qquad (6)$$

$$\text{Accuracy} = \frac{\text{TP}*\text{TN}}{\text{TP+TN+FP+FN}} \qquad (7)$$

TP means, which are truly classified as the positive terms. True positives (TP) are examples that the classifier correctly labelled as belonging to the positive class. False positive (FP) are examples which were not labelled by the classifier as belonging to the positive class but should have been. True Negative (TN) is examples that the classifier correctly labelled as belonging to the negative class. True Negative means, which are truly classified as the Negative terms. At last there is False Negative (FN), which is an example which was not labelled by the classifier as belonging to the negative class but should have been.

The obtained results in Figures 3 reveal the better performance of SVM-PSO with respect to Precision, Recall and Accuracy while comparing with SVM. It is found that the proposed method for sentiment classification outperform the existing techniques.

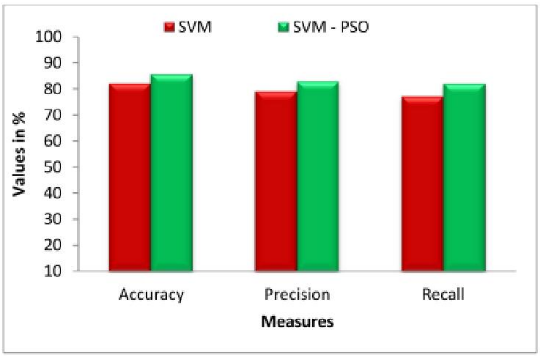

Figure 3: Results of SVM and SVM-PSO

4. CONCLUSION

The proposed system uses PSO-based methodto build novel sentiment classification. In SVM parameters C,σ and εre elected by PSO. The experiment results show that the approach is not only able to achieve the process of selecting important features but also to yield high accuracy for sentiment classification. This proposed system has shown that PSO affect the accuracy of SVM after the hybridization of SVM-PSO. In our future work, it will further improve the feature selection algorithm.

REFERENCES

1. Mark Dredze 2012, 'How Social Media Will Change Public Health', IEEE Intelligent Systems, vol.27, 2012, Volume 04, Pages 81-84.
2. Hong Liu & Xiaojun Li 2010, 'Internet Public Opinion Hotspot Detection Research Based on K-means Algorithm', ICSI 2010, Part II, LNCS 6146, pp. 594–602, Springer-Verlag Berlin Heidelberg.
3. J. Bollen, H. Mao, and X. Zeng , 'Twitter mood predicts the stock market', Journal of Computational Science, 2011, Volume 2, No. 1, Pages 1-8.
4. Y. Yang. Y, X. Liu, 'A re-examination of text categorization methods', Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), ACM, New York, NY, USA,1999, Pages 42–49.
5. Y. Wang, and X. Wang, 'New Approach to Feature selection in Text Classification', Proceedings of the 4th International Conference on Machine Learning and Cybernetics. IEEE 2005, Pages 145-189.
6. V. Vapnik, 'The Nature of Statistical Learning Theory', Springer- Verlag, 2000, Pages 863-884.
7. K.Nirmala Devi and V.Murali Bhaskaran ,'Online forums hotspot prediction base on Sentiment Analysis', Journal of Computer Science, 20012, Volume. 8, No.8, Pages 1219-1224.
8. Turney , 'Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews', In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia,2002, Pages 417-424.
9. B. Liu, 'Web Data Mining', Second Edition, Springer Berlin Heidelberg, 2011.
10. X.yu, Y.liu, and X.huang, 'Mining online reviews predicting sales performance', EEE Transactions on Knowledge and Data Engineering, 2012, Volume 24, No. 4, Pages 720-734.
11. Earle, PS, Bowden, DC & Guy, M 2011, 'Twitter Earthquake Detection: Earthquake Monitoring in a Social World', Annals Geophysics, 2011, Volume 54, No. 6.
12. B.Pang , L. Lee and S. Vaithyanathan, 'Thumbs up ? sentiment classification using machine learning techniques', Proceeding of the conference on empirical methods in natural language processing, 2002, Pages 79–86.
13. J. Kennedy and R. Eberhart ,Swarm intelligence, Morgan Kaufmann Publishers, Inc., San Francisco, CA, 2001.
14. M. Clerc and J.Kennedy , 'The particle swarm-explosion, stability, and convergence in a multidimensional complex space', IEEE Transactions on Evolutionary Computation, 2002, Volume 6, No.1, Pages 58-73.
15. K.Parsopoulos and M.N Vrahatis, 'On the computation of all global minimizers through particle swarm optimization', IEEE Transactions on Evolutionary Computation, 2004, Volume 8 No.3, Pages 211-224.