International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:9, No:12, 2015

One-Class Support Vector Machine for Sentiment Analysis of Movie Review Documents

Chothmal, Basant Agarwal

***Abstract*—**Sentiment analysis means to classify a given review document into positive or negative polar document. Sentiment analysis research has been increased tremendously in recent times due to its large number of applications in the industry and academia. Sentiment analysis models can be used to determine the opinion of the user towards any entity or product. E-commerce companies can use sentiment analysis model to improve their products on the basis of users' opinion. In this paper, we propose a new One-class Support Vector Machine (One-class SVM) based sentiment analysis model for movie review documents. In the proposed approach, we initially extract features from one class of documents, and further test the given documents with the one-class SVM model if a given new test document lies in the model or it is an outlier. Experimental results show the effectiveness of the proposed sentiment analysis model.

***Keywords*—**Feature selection methods, Machine learning, NB, One-class SVM, Sentiment Analysis, Support Vector Machine.

I. Introduction

THE textual information available on the Web is of two types: facts and opinions statements. Facts are objective sentences about the entities, and do not show any sentiments. Opinions are subjective in nature and generally describe the people's sentiments towards entities and events. Sentiment analysis research has been increasing tremendously for last 10 years due to the wide range of business and social applications [1]. Opinion Mining or Sentiment Analysis is the study that analyses people's opinion and sentiment towards entities such as products, services etc. in the text. The automatic analysis of online contents to extract the opinion requires deep understating of natural text by the machine. Sentiment analysis research can be categorized among Document level, Sentence level and Aspect/Feature level sentiment analysis. Document level sentiment analysis classifies a review document as containing positive or negative polarity. It considers a document as a single unit. Sentence level sentiment analysis takes a sentence to extract the opinion or sentiment expressed in that sentence. Aspect based sentiment analysis deals with the methods that identify the aspects/entities in the text about which an opinion is expressed [2]. Further, the sentiments expressed about these entities are identified.

What other people think has always been very important in decision-making. Whenever people want to purchase a product (for e.g. mobile phone, camera, laptop, etc.), they ask their friends or their peers about the product if they have used that. Nowadays, due to the advent of Web recent trends, people express their opinion, feelings, and experiences about the products or services on the forums, blogs, social network and content-sharing services. The user can know the merits and demerits of the product from the experiences shared by people on the web, which can be useful for them in taking purchasing decisions [3]. E-commerce companies can improve their products or services on the basis of users' opinion and can also know the current trends of the market. Examples of sentiment analysis include identifying movie popularity from online reviews, which model of a camera is liked by most of the users and which music is liked by most of the people, etc. Opinion mining and sentiment analysis also have applications in political domain and brand analysis.

In this paper, we propose to use one-class SVM machine learning algorithm for the sentiment analysis. In this methods, only review document of one class either positive or negative class are required to develop the machine-learning model. We use positive review documents to develop machine learning model. It is due to the assumption that most of the reviews are positive reviews. Most of the people write the positive reviews.

The paper is organized as follows. In Section II, we give an overview of the related work done in the field of sentiment analysis. In Section III, we provide a description of the one-class SVM algorithm which we use for sentiment analysis. Section IV presents the proposed algorithm. Experimental results and discussion are presented in Section V. Finally, we conclude future work in Section VI.

II. Related Work

Sentiment analysis research has attracted large number of researchers around the globe [1], [17], [18]. Machine learning methods have been widely applied for sentiment analysis problem. Mainly Support Vector Machine (SVM), Naive Bayes (NB), Maximum Entropy, Artificial Neural Networks methods have been adopted by most of the researchers for sentiment analysis [4], [5], [19], [20]. Authors in [4] used different machine learning algorithms like NB, SVM, and Maximum Entropy (MaxEnt) for sentiment analysis of movie review dataset. Their experimental results show that SVM outperforms other machine learning method for sentiment analysis. Authors in [5] also explored that SVM performs better than other classifiers for sentiment analysis. Authors in [6] used SVM and NB classifiers for sentiment analysis with various feature weighting schemes and feature selection methods. Their experimental results showed that SVM classifier is better than NB classifier for sentiment analysis.

Prof. Chothmal (corresponding author) and Dr. Basant Agarwal are with the Department of Computer Science and Engineering, Swami Keshvan and Institute of Technology Management & Gramothan, Jaipur, India (phone: +919414305959; e-mail: hodcs@skit.ac.in, thebasant@gmail.com).

Authors in [7] experimented with three supervised learning algorithms namely NB, SVM and character based N-gram model for the domain of travel destination reviews. They used frequency of words to give weights to the features. Their experimental results showed that SVM outperforms other classifiers for sentiment analysis. Authors in [8] also investigated that discriminative classifiers such as SVM is more appropriate for sentiment analysis as compared to other generative models and winnow classifier.

Authors in [9] presented an empirical study on the comparison between SVM and Artificial Neural Network (ANN) classifiers for document-level sentiment analysis. They discuss the limitations of both the methods for sentiment analysis. Authors in [10] experimented SVM classifier with several weighting schemes for various different domains of datasets. Authors in [11] constructed various classifiers with different feature sets of unigrams and POS based features, and further, these classifiers are combined using several combination rules. It was investigated that combined classifier outperforms individual classifiers. Authors in [12] also proposed an integrated classifier consisting of SVM, Maximum Entropy and score calculation based classifiers, resulting into improved performance of movie review classification. Authors in [1] developed a three phase framework for choosing optimal combination of classifiers based on assembling multiple classifiers. Authors in [13] proposed various types of ensemble methods for various categories of features (i.e. POS based, word relation based) and classifiers (NB, SVM, Maximum Entropy) for sentiment analysis. Authors in [14] proposed hybrid classifier by combining rule-based classification, machine learning, and supervised learning method to improve the classification effectiveness.

III. One-Class Support Vector Machine

Support Vector Machines (SVMs) is a supervised learning method. SVM training algorithm builds a model that predicts whether a new example falls into one category or the other. Support Vector Machine (SVM) is a non-linear classifier which is often reported as producing superior classification results compared to other methods. The idea behind SVM classifier is that it map the non-linearly input data to high dimensional space, where the data can be linearly separated, thus provide better classification performance. A special property of SVM is that it simultaneously minimizes empirical classification errors and maximizes geometric margins.

One-class Support Vector Machine (SVM) was proposed by Scholkopf for estimating the support of a high-dimensional distribution [15]. Given a training dataset without any class information, the One-class SVM constructs a decision function that takes the value +1 in a small region capturing most of the data points, and -1 elsewhere. The strategy in this technique is to map the input vectors into a high dimension feature space corresponding to a kernel, and construct a linear decision function in this space to separate the dataset from the origin with maximum margin. With the freedom to utilize different types of kernel, the linear decision functions in the feature space are equivalent to a variety of non-linear decision functions in the input space. The One-class SVM introduces a parameter v (0, 1) to control a tradeoff between the fractions of data points in the region and the generalization ability of the decision function. One-class SVM is represented in Fig. 1.

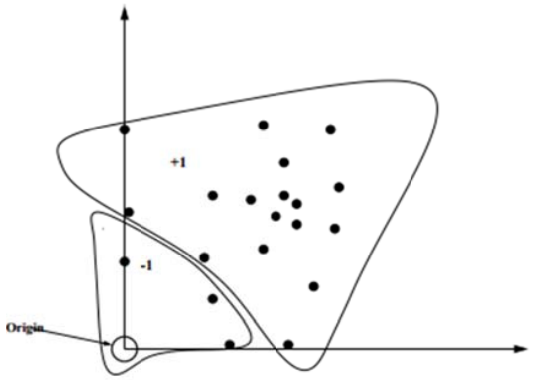

Fig. 1 One-class SVM

IV. Proposed Approach

The procedure of making a One-class Support Vector Machine is as following and is further explained in subsections.

1. Extract features from the review document corpus.
2. Transform data to the format of an SVM package.
3. Select parameters for the SVM in LibSVM (e.g. considering the RBF kernel, find the best parameter C and gamma).

A. Feature Extraction

Unigrams features are simply bag-of-words features which are extracted by eliminating extra spaces and noisy characters between any two words. For example sentence, “this is an awesome movie”. Here, words ‘this’, ‘is’, ‘an’, ‘awesome’, ‘movies’ are all unigram features. Bi-grams are the features, consisting of every two consecutive words in the text. For example, “this is not a good book”. Here, ‘this is’, ‘is not’, ‘not a’, ‘a good’, ‘good book’ are the bigram features. These features are capable of incorporating some contextual information. All the feature sets are represented using binary weighting schemes. Similar, trigram feature set is constructed by considering three consecutive words in the sentence. Trigram feature set contains contextual information. In addition to these, we construct composite feature sets by considering these features.

B. Transform the Data to LibSVM Format

LibSVM is used in our experiments to develop One-class SVM model. Features are represented in the LibSVM format to develop a one-class SVM model. Further, we applied min-max scaling on the data so that all the data is in the same range. The main advantage of scaling is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges.

C. Model Selection for One-Class SVM

There are two parameters need to be set before training the One-class SVM: v and gamma. The generalization performance of one-class SVM can be evaluated by two measures: the size of region and the generalization fraction of data points in the region. Small size indicates the probability that a data point of class "-1" falls into this region is small. Great generalization fraction of data points indicates that the probability that a new data point of this class (+1) falls into this region is great. The parameter decides the non-linear characteristics of the decision function, in other words, it decides the "shape" of the region. The parameter controls not only the fraction of data points in the region but also the generalization ability. Thus, the kernel and v all influence the size of region and the generalization fraction of data points in the region. We experiment with different combinations of v and to get the best results,

V. Dataset, Experimental Setup and Results

A. Dataset Used

To evaluate the performance of the proposed methods, one of the most popular publicly available movie review dataset is used [16]. This standard dataset, known as Cornell Movie Review Dataset consists of 2000 reviews containing 1000 positive and 1000 negative labelled reviews collected from Internet Movie Database (IMDb). Support Vector Machine (SVM), Naïve Bayes (NB) algorithms have been used extensively for sentiment analysis [4]-[6]. Therefore, we use these classifiers for classification of review documents in positive or negative class. In addition, we use One-class SVM classifier for sentiment analysis. To evaluate the performance of the proposed method, we use 90% documents positive documents to develop the one-class SVM model. Further, we use 10% positive and 10% negative documents for testing. Further, to evaluate the performance with other classifiers viz. SVM, NB, we divide the dataset into 90% training and 10% testing documents, such that both the sets are disjoint. LibSVM software tool is used to develop one-class SVM and linear SVM classifiers. WEKA software is used to develop NB classification model for sentiment analysis.

B. Evaluation Metrics

Precision, Recall, Accuracy and F- measure are used for evaluating performance of sentiment classification. Precision for a class C is the fraction of total number of documents that are correctly classified and total number of documents that classified to the class C (sum of True Positives (TP) and False Positives (FP)). Recall is the fraction of total number of correctly classified documents to the total number of documents that belongs to class C (sum of True Positives and False Negative (FN)). F –measure is the combination of both precision and recall, is given by

$$F - Measure = 2*(precision * recall)/(precision + recall) \quad (1)$$

F-measure is used to report the performance of classifiers for the sentiment classification.

C. Results and Discussions

Experimental results show that unigram features performs better than other features if we use them individually (results as shown in Table I) for movie review dataset. For example, unigram feature set produces the F-measure 84.2% as compared to 78.8% and 56.2% respectively for bi-grams, trigram features with SVM classifier on movie review datasets as shown in Table I. Further, composite features produce better results as compare to individual features. Composite feature set comprising of unigrams, bigrams, and trigrams produces the best results in comparison to all other features. For example, composite feature set of unigrams, bigrams, and trigrams gives the F-measure of 87.0 % with SVM classifier. Composite features gives better results as compare to their individual features due to the reason that by combing the features more information is used in the classification model.

TABLE I
F-Measure (in %) for the Various Feature Sets with Various Classifiers on Movie Review Dataset

Features	SVM	Naive Bayes	One-Class SVM
Unigrams	84.2	82.6	86.1
Bigrams	78.8	74.9	80.1
Trigrams	56.2	59.1	60.1
Unigrams + Bigrams	86.7	83.5	86.9
Unigrams + Trigrams	84.4	82.8	86.5
Unigrams + Bigrams + Trigrams	**87.0**	**84.9**	**87.9**

One-class SVM classifier performs best as compared to other classification algorithms for sentiment analysis (results as shown in Table I). For example, one-class SVM produces better F-measure of 86.1% as compared to 84.2% and 82.6% for SVM and NB classifiers respectively. Composite feature set of unigrams, bigrams and trigrams produces the best results with one-class SVM i.e. F-measure of 87.9% on movie review datasets. The possible reason that the one-class machine learning model performs better than other classifiers is that it makes use of both the positive and negative sample information in building the model.

VI. Conclusion

Sentiment analysis is to assign a given review document into positive or negative polarity. In this paper, we propose a new One-class Support Vector Machine (One-class SVM) based sentiment analysis model for movie review documents. Proposed method uses review documents of one of the class either positive or negative for developing the sentiment analysis model. We used positive review documents for developing one-class SVM model. Experimental results show the effectiveness of the proposed sentiment analysis model. We wish to compare the performance of these features on more datasets of different domain, and also study the effect of proposed method on non-English documents.

REFERENCES

[1] Liu B., "Sentiment Analysis and Opinion Mining", Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, 2012.
[2] Liu B., "Sentiment Analysis and Subjectivi," Handbook of Natural Language Processing", 2nd ed., N. Indurkhya and F.J. Damerau, eds., Chapman & Hall / CRC Press, 2010, pp. 627-666.
[3] Agarwal B., Mittal N., "Enhancing Performance of Sentiment Analysis by Semantic Clustering of Features", In IETE Journal of Research, Taylor and Francis, 2014, pp: 1-9.
[4] Agarwal B., Mittal N., "Prominent Feature Extraction for Sentiment Analysis", Springer Book Series: Socio-Affective Computing series, ISBN: 978-3-319-25343-5, DOI: 10.1007/978-3-319-25343-5, pages: 1-115.
[5] Pang B., Lee L., Vaithyanathan S., "Thumbs up? Sentiment classification using machine learning techniques", In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002, pp: 79-86.
[6] Tan S., Zhang J., "An empirical study of sentiment analysis for Chinese documents", In Expert Systems with Applications, Vol: 34, No: 4, 2008, pp: 2622-2629.
[7] O'keefe T., Koprinska I., "Feature Selection and Weighting Methods in Sentiment Analysis", In Proceedings of the 14th Australasian Document Computing Symposium, Sydney, Australia, 2009, pp: 67-74.
[8] Ye Q., Zhang Z., Law R., "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches", In Expert Systems with Applications, Vol: 36, No: 3, 2009, pp: 6527-6535.
[9] Cui H., Mittal V., Datar M., "Comparative experiments on sentiment classification for online product reviews", In Proceedings of the 21st national conference on Artificial Intelligence, 2006, pp: 1265-1270.
[10] Moraes R., Valiati JF, Neto WPG, "Document-level sentiment classification: An empirical comparison between SVM and ANN", In Expert Systems with Applications, Vol: 40, No: 2, 2013, pp: 621-633.
[11] Saleh MR, Martin-Valdivia MT, Montejo-Raez A., Urena-Lopez LA, "Experiments with SVM to classify opinions in different domains", In Expert Systems with Applications, Vol: 38, No: 12, 2011, pp: 14799-14804.
[12] Li S., Zong C., Wang X., "Sentiment Classification through Combining classifiers with multiple feature sets", In Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), 2007, pp: 135- 140.
[13] Tsutsumi K., Shimada K., Endo T., "Movie Review Classification Based on a Multiple Classifier", In Proceedings of the Annual meetings of the Pacific Asia Conference on Language, Information and Computation (PACLIC), 2007, pp: 481-488.
[14] Xia R., Zong C., Li S., "Ensemble of Feature Sets and Classification Algorithms for Sentiment Classification". In Journal of Information Sciences, Vol: 181, No: 6, 2011, pages: 1138-1152.
[15] Prabowo R., Thelwall M., "Sentiment analysis: A combined approach", In Journal of Informatics, Vol: 3, No: 2, 2009, pp:143-157.
[16] Agarwal B., Mittal N., "Optimal Feature Selection for Sentiment Analysis", In 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2013),Vol-7817, pages-13-24, Greece, Samos. 2013.
[17] Schaolkopf B., Platt J. C., Shawe-Taylor J. C., Smola A. J., Williamson, R.C, "Estimating the support of a high-dimensional distribution.", In Neural Comput.13, 7, 1443-1471.
[18] Pang B., Lee L., "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts", In Proceedings of the Association for Computational Linguistics (ACL), 2004, pp. 271-278.
[19] Agarwal B., Mittal N., Bansal P., Garg S., "Sentiment Analysis Using Common-Sense and Context Information", In Computational Intelligence and Neuroscience, Article ID 715730, 9 pages, 2015, DOI: http://dx.doi.org/10.1155/2015/715730.
[20] Agarwal B., Mittal N., "Prominent Feature Extraction for Review Analysis: An Empirical Study", In Journal of Experimental and theoretical Artificial Intelligence, Taylor Francis, 2014, DOI:10.1080/0952813X.2014.977830.