Sentiment Classification using Rough Set based Hybrid Feature Selection

Basant Agarwal
Department of Computer Engineering
Malaviya National Institute Technology
Jaipur, India
thebasant@gmail.com

Namita Mittal
Department of Computer Engineering
Malaviya National Institute Technology
Jaipur, India
nmittal@mnit.ac.in

Abstract

Sentiment analysis means to extract opinion of users from review documents. Sentiment classification using Machine Learning (ML) methods faces the problem of high dimensionality of feature vector. Therefore, a feature selection method is required to eliminate the irrelevant and noisy features from the feature vector for efficient working of ML algorithms. Rough Set Theory based feature selection method finds the optimal feature subset by eliminating the redundant features. In this paper, Rough Set Theory (RST) based feature selection method is applied for sentiment classification. A Hybrid feature selection method based on RST and Information Gain (IG) is proposed for sentiment classification. Proposed methods are evaluated on four standard datasets viz. Movie review, product (book, DVD and electronics) review dataset. Experimental results show that Hybrid feature selection method outperforms than other feature selection methods for sentiment classification.

1 Introduction

Sentiment analysis is to extract the users' opinion by analysing the text documents (Pang et al. 2008). Nowadays people are using web for writing their opinion on blogs, social networking websites, discussion forums etc. Hence, it is very much needed to analyse these web contents. Thus, it increases the demand of sentiment analysis research. Sentiment analysis has been very important for the users as well as for business with the drastic increase of online content. For users, it is important to know past experiences about some product or services for taking decision in purchasing products. Companies can use sentiment analysis in improving their products based on the users' feedback written about their products on blogs. E-commerce based companies know the online trends about the products. Example of sentiment analysis is - knowing which model of a camera is liked by most of the users.

Sentiment classification can be considered as a text classification problem. Bag-of-Words (BOW) representation is commonly used for sentiment classification using machine learning approaches. The words present in all the documents create the feature vector. Generally, this feature vector is huge in dimension that is used by machine learning methods for classification. This high dimensional feature vector deteriorates the performance of machine learning algorithm. Rough set theory has been used for reducing the feature vector size for text classification (Jensen et al. 2001; Jensen et al. 2009; Wakaki et al. 2004). However, it has not been investigated for sentiment analysis yet.

Contribution of this paper:-

1. Rough Set theory based feature selection method is applied for sentiment classification.
2. Hybrid Feature selection method is proposed based on Rough Set and Information Gain which performs better than other feature selection methods.
3. Proposed methods are experimented with four different standard datasets.

The paper is organized as follows: A brief discussion of the earlier research work is given in Section 2. Section 3 describes the feature selections method used for sentiment classification. Dataset, Experimental setup and results are discussed in Section 4. Finally, Section 5 describes conclusions.

Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 115–119,
Atlanta, Georgia, 14 June 2013.

2 Related Work

Machine Learning methods have been widely applied for sentiment analysis (Pang et al. 2008; Pang et al. 2002; Tan et al. 2008). Pang and Lee (2004) experimented with various features like unigrams, bi-grams and adjectives for sentiment classification of movie reviews using different machine learning algorithms namely Naïve Bayes (NB), Support Vector Machines (SVM), and Maximum-Entropy (ME). Feature selection methods improve the performance of sentiment classification by eliminating the noisy and irrelevant features from feature vector. Tan et al. (2008) investigated with various feature selection methods with different machine learning algorithm for sentiment classification. Their experimental results show that IG performs better as compared to other feature selection methods and SVM is best machine learning algorithms. Categorical Probability Proportion Difference (CPPD) feature selection method is proposed which computes the importance of a feature based on its class discriminating ability for sentiment classification (Agarwal et al. 2012). Various features are extracted from the text for sentiment classification. Further, Minimum Redundancy Maximum Relevancy (mRMR) and IG feature selection methods are used to select prominent features for better sentiment classification by machine learning algorithms (Agarwal et al. 2013).

Rough set based dimensionality reduction method is applied for data reduction to characterize bookmarks and it is compared with conventional entropy based reduction method (Jensen et al. 2009). Dimension reduction method based on fuzzy-rough sets and Ant Colony Optimization (ACO) method is proposed (Jensen et al. 2006), which is applied to the web categorisation problem. Experimental result show significant reduction in the data redundancy. Rough set theory is applied to select relevant features for web-page classification. Their experimental results show that the rough set based feature selection method with SVM gives better accuracy (Wakaki et al. 2004). Applicability of RS theory for various existing text classification techniques are discussed in detail with e-mail categorization as an example application (Chouchoulas et al. 2001).

3 Methodology Used

3.1 Rough Set Attribute Reduction (RSAR)

Rough Sets Theory (RST) (Jensen et al. 2007) is a mathematical tool to make attribute reduction by eliminating redundant condition attributes (features). The rough set is the approximation of a vague concept (set) by a pair of precise concepts, called lower and upper approximations. Rough Set Attribute Reduction (RSAR) (Jensen et al. 2007) is a filter based method by which redundant features are eliminated by keeping the amount of knowledge intact in the System. Basic intuition behind RSAR is that objects belonging to the same category (same attributes) are not distinguishable (Jensen et al. 2009).

RSAR algorithm finds the vague attributes which do not have important role in the classification. Therefore, it is needed to remove redundant features without changing the knowledge embedded in the information system. An important issue in data analysis is to discover dependencies between the attributes. QUICKREDUCT method (Jensen et al. 2007; Jensen et al. 2009) calculate a minimal reduct without exhaustively generating all possible subsets, it is used in our experiments for obtaining optimal feature subset. Main advantage of RSAR is that it does not require any additional parameter to operate like threshold is required in case of IG.

3.2 Information Gain (IG)

Information gain (IG) is one of the important feature selection techniques for sentiment classification. IG is used to select important features with respect to class attribute. It is measured by the reduction in the uncertainty in identifying the class attribute when the value of the feature is known. The top ranked (important) features are selected for reducing the feature vector size in turn better classification results.

3.3. Proposed Hybrid Approach to Feature Selection

The usefulness of an attribute is determined by both its relevancy and redundancy. An attribute is relevant if it is predictive to the class attribute, otherwise it is irrelevant. An attribute is consid-

ered to be redundant if it is correlated with other attributes. Hence, The Aim is to find the attributes that are highly correlated with the class attribute, but not with other attributes for a good attribute subset (Jensen et al. 2007).

Information Gain based feature selection methods determine the importance of a feature in the documents. But, it has disadvantage that threshold value is required initially which is not known generally. This method does not consider the redundancy among the attributes. In addition, it will return large number of features when massive amount of documents are to be considered. RSAR can reduce most of the irrelevant and noisy features. It reduces the redundancy among the features. It has advantage that it considers the dependency of combination of features on decision attribute in contrast to other conventional feature selection methods (Jensen et al. 2007). However, it has some disadvantages. Firstly, to get an optimal reduct is a NP-hard problem, some heuristic algorithms are used to get approximate reduction (Jensen et al. 2004; Jensen et al. 2009). Secondly, it is very time consuming. Therefore, an integrated method is developed which can reduce most of the redundant features and get the minimal feature set with reduced time complexity for sentiment classification.

Proposed Algorithm works in two steps. Firstly, Information Gain (IG) of each feature is computed and all the features are taken which has information gain value to be greater than 0. So that initially irrelevant and noisy features are removed from the feature vector, by this a lot computational efforts are reduced. Main assumption and motivation behind this step is that IG would eliminate the features which are likely to be noisy and irrelevant features. Further, Reduced feature set is sent to the RSAR feature selection method to get optimal feature subset. So, by combining both the methods a feature selection is proposed which is more efficient in terms of computational and time complexity.

4 Dataset Used and Experimental Setup

For the evaluation of the proposed method, one of the most popular publically available movie review dataset (Pang et al. 2004) is used. This standard dataset contains 2000 reviews comprising 1000 positive and 1000 negative reviews. Product review dataset consisting amazon products reviews is also used provided by Blitzer et al. (2007). We used product reviews of books, DVD and electronics for experiments. Each domain has 1000 positive and 1000 negative labelled reviews. Documents are initially pre-processed as follows:

(i) Negation handling is performed as Pang et al. (2002), "NOT_" is added to every words occurring after the negation word (no, not, isn't, can't, never, couldn't, didn't, wouldn't, don't) and first punctuation mark in the sentence.

(ii) Words occurring in less than 3 documents are removed from the feature set.

Binary weighting scheme has been identified as a better weighting scheme as compared to frequency based schemes for sentiment classification (Pang et al. 2002); therefore we also used binary weighting method for representing text. In addition, there is no need of using separate discretisation method in case of binary weighting scheme as required by RSAR feature selection algorithm.

Noisy and irrelevant features are eliminated from the feature vector generated after pre-processing using various feature selection methods discussed before. Further, prominent feature vector is used by machine learning algorithms. Support Vector Machine (SVM) and Naïve Bayes (NB) classifiers are the mostly used for sentiment classification (Pang et al. 2002; Tan et al. 2008). Therefore, we report the classification results of SVM and NB classifier for classifying review documents into positive or negative sentiment polarity. For the evaluation of proposed methods 10 fold cross validation method is used. F-measure value is reported as a performance measure of various classifiers (Agarwal et al. 2013)

4.1 Experimental results and discussions

Initially, unigram features are extracted from the review documents. Feature set without using any feature selection method is taken as a baseline. Further, various feature selection algorithms are used for selecting optimal feature subset. IG is used for comparison with the proposed feature selection method as it has been considered as one of the best feature selection method for sentiment classification (Pang et al. 2008; Tan et al. 2008). Feature subsets obtained after applying RSAR, IG

and proposed hybrid feature selection algorithm are called Rough features, IG features and Hybrid IG-Rough features respectively. Feature vector lengths for various features used for sentiment classification of different datasets are shown in Table 1. In the experiments, Firstly, RSAR algorithm is applied to get the best optimal feature subset. Further, according to the feature subset size obtained from RSAR method, threshold is set for IG based to get the feature vector, which is further used for classification. Experiments are conducted in this way so that results of Rough features and IG features can be compared.

	Movie	**Book**	**DVD**	**Electronics**
Unigram Features	9045	5391	5955	4270
Rough Features	263	310	350	371
IG Features	263	310	350	371
Hybrid IG-Rough Features	339	410	403	405

Table 1. Feature Length for Various Features Used With Four Datasets

Experimental results show that both feature selection methods (RSAR and IG) are able to improve the performance from baseline (as shown in Table 2). For example from Table 2, F-measure is increased from 84.2% to 85.9% (+2.1) and 85.6% (+1.6) for Rough features and IG features respectively with SVM classifier when movie review dataset is considered. Similarly, when electronics dataset is used, SVM classifier increased the performance from 76.5% to 82.9% (+8.3) and 81.1% (+6.01) for Rough and IG features. It is due to the fact that RSAR algorithm removes the redundancy and selects the prominent feature subset, and IG selects the top ranked features by its importance to the class attribute.

When hybrid features selection approach is used for movie review dataset, F- measure is increased from 84.2% to 87.7 (+4.15) for SVM classifier as given in Table 1. Hybrid IG-Rough features gives better classification results as compare to other features with very small feature vector length. It is due to the fact that IG in its first phase eliminates the irrelevant and noisy features and in second phase RSAR algorithm decreases the redundancy among features and extracts the optimal feature subset. By combining both the methods, a more robust feature selection method is developed for sentiment classification which is more efficient in selecting optimal feature set for massive dataset. Because when dataset size would be very large, RSAR algorithm will take much time and IG algorithm would be having problem of large feature size and pre-setting the threshold value.

		Unigram Features	**rough Features**	**IG Features**	**Hybrid IG-Rough Features**
Movie	SVM	84.2	85.9 (+2.1)	85.6 (+1.6)	87.7 (+4.15)
	NB	77.1	78.7 (+2.1)	78.6 (+2.0)	80.9 (+4.9)
Book	SVM	76.2	78.0 (+2.3)	77.0 (+1.0)	80.2 (+5.2)
	NB	74.4	74.9 (+0.1)	76.3 (+2.5)	79.1 (+6.3)
DVD	SVM	77.3	80.4 (+4.0)	79.1 (+2.3)	83.2 (+7.6)
	NB	74.2	76.5 (+3.1)	75.1 (+1.2)	78.1 (+5.2)
Electronics	SVM	76.5	82.9 (+8.3)	81.1 (+6.0)	83.5 (+9.1)
	NB	74.9	75.5 (+0.1)	75.2 (+.04)	78.1 (+4.2)

Table 2 F-measure (in %) for various features with four datasets

5 Conclusion

Rough set based dimension reduction method is applied for sentiment analysis. It is capable of reducing the redundancy among the attributes. Rough set based methods computes the best feature subset based on minimized redundancy in contrast to information gain which computes the importance of the attribute based on the entropy. Hybrid feature selection method is proposed which is based on RSAR and IG. Experimental results show that Hybrid feature selection method with very less number of features produces better results as compared to other feature selection methods. All the methods are experimented using four standard datasets. In future, more methods can be explored for making rough set based feature selection method computationally more efficient by incorporating evolutionary approaches in selecting feature subsets.

References

Alexios Chouchoulas, Qiang Shen, "Rough set-aided key- word reduction for text categorization", *Applied Artificial Intelligence,* Vol. 15, No. 9, pp. 843-873. 2001.

Basant Agarwal, Namita Mittal, "Categorical Probability Proportion Difference (CPPD): A Feature Selection Method for Sentiment Classification", *In Proceedings of the 2nd Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), COLING 2012*, pp 17–26, 2012.

Basant Agarwal, Namita Mittal, "Optimal Feature Selection Methods for Sentiment Analysis", *In 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2013)*, Vol-7817,pp:13-24, 2013.

Bo Pang, Lillian Lee. "Opinion mining and sentiment analysis*", Foundations and Trends in Information Retrieval*, Vol. 2(1-2):pp. 1–135, 2008.

Bo Pang, Lillian Lee, Shivakumar Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques", *In the Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 79–86. 2002.

Bo Pang, Lillian Lee, "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts", *In the Proceedings of the Association for Computational Linguistics (ACL)*, 2004, pp. 271–278. 2004.

John Blitzer, Mark Dredze, Fernando Pereira, "Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification", *In Proc. Assoc. Computational Linguistics. ACL Press*, pp 440-447, 2007.

Richard Jensen, Qiang Shen, "Fuzzy-Rough Sets Assisted Attribute Selection*", In IEEE Transactions on Fuzzy Systems*, Vol. 15, No. 1, February 2007.

Richard Jensen, Qiang Shen, "A Rough Set-Aided System for Sorting WWW Bookmarks". *In N. Zhong et al. (Eds.), Web Intelligence: Research and Development.* pp. 95-105, 2001.

Richard Jensen, Qiang Shen, "New Approaches to Fuzzy-Rough Feature Selection", *In the IEEE Transactions on Fuzzy Systems*, vol. 17, no. 4, pp. 824-838, 2009.

Richard Jensen, Qiang Shen, "Webpage Classification with ACO-enhanced Fuzzy-Rough Feature Selection", *In the Proceedings of the Fifth International Conference on Rough Sets and Current Trends in Computing (RSCTC 2006), LNAI* 4259, pp. 147-156. 2006

Richard Jensen, Qiang Shen "Fuzzy-Rough Attribute Reduction with Application to Web Categorization". *In the Transaction on Fuzzy Sets and Systems 141(3),* pp. 469-485. 2004.

Songbo Tan , Jin Zhang "An empirical study of sentiment analysis for chinese documents*", In Expert Systems with Applications* , pp:2622–2629 (2008).

Toshiko Wakaki, Hiroyuki Itakura, Masaki Tamura, "Rough Set-Aided Feature Selection for Automatic Web-Page Classification". *In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, Pages 70-76, 2004