Sentiment Classification of Web Review Using Association Rules

Man Yuan[1,2], Yuanxin Ouyang[1,2,*], Zhang Xiong[1,2], and Hao Sheng[1,2]

[1] School of Computer Science and Technology, Beihang University, 100191 Beijing, P.R. China
[2] Research Institute of Beihang University in Shenzhen, 518000 Shenzhen, P.R. China
{ym,oyyx,xiongz,shenghao}@buaa.edu.cn

Abstract. Sentiment Classification of web reviews or comments is an important and challenging task in Web Mining and Data Mining. This paper presents a novel approach using association rules for sentiment classification of web reviews. A new restraint measure AD-Sup is used to extract discriminative frequent term sets and eliminate terms with no sentiment orientation which contain close frequency in both positive and negative reviews. An optimal classification rule set is then generated which abandons the redundant general rule with lower confidence than the specific one. In the class label prediction procedure, we proposed a new metric voting scheme to solve the problem when the covered rules are not adequately confident or not applicable. The final score of a test review depends on the overall contributions of four metrics. Extensive experiments on multiple domain datasets from web site demonstrate that 50% is the best min-conf to guarantee classification rules both abundant and persuasive and the voting strategy obtains improvements on other baselines of using confidence. Another comparison to popular machine learning algorithms such as SVM, Naïve Bayes and kNN also indicates that the proposed method outperforms these strong benchmarks.

Keywords: Association rule, sentiment classification, text categorization.

1 Introduction

Sentiment Classification, also referred as Polarity Classification or Binary Classification, aims to determine whether the semantic orientation of the given text is positive or negative. The rise of social media such as blogs and social networks as well as e-commerce suppliers such as Amazon and EBay has fueled interest in sentiment analysis. Automatic detection and analysis of consumer reviews or comments from the Web holds great promise for customer research, business intelligence, recommendation system and Smart City applications which need the human feedback analysis. As the number of Web reviews for any product (movies,

* Corresponding author.

A.A. Ozok and P. Zaphiris (Eds.): OCSC/HCII 2013, LNCS 8029, pp. 442–450, 2013.

e-commerce, social network content etc.) grows rapidly, it is hard for a potential consumer to make informed decision when reading hundreds of reviews on a single product, and if he/she reads part of the reviews, he/she may get a biased view point. For manufactures or online shopping sites, it also requires great effort to manage and keep track of the large scale review dataset. So, sentiment classification is becoming a challenging and interesting topic in text mining area.

Effective sentiment classification relies on multiple disciplines, such as machine learning, natural language processing, linguistic, statistic etc. One of the main methodologies for sentiment classification is to treat sentiment classification as a special case of Text Categorization [1], which has been a well studied field in the last few decades. Comparing with text categorization, the difference in sentiment classification is that the predefined labels only include "positive", "negative" and sometimes "neural", rather than the topics. Comparative studies demonstrated that these general techniques provide strong baseline accuracy for sentiment classification and outperforms other method based on lexicon analysis.

Pang [2] firstly tried to classify movie reviews into positive/negative by using several supervised machine learning methods: Naive Bayes, Maximum Entropy and SVM. In their following work [3], they added in subjectivity detection with minimum cuts algorithm to avoid the sentiment classifier from dealing with irrelevant "objective" sentences. As reported in their work, the classification performance of product reviews is worse than that of normal topical text categorization. One of the main difficulties is that people typically use both positive and negative words in the same review. In analyzing political speeches, [4] exploited the argument structure found in speaker reference links to help determine how a members of congress would vote given their congressional floor speeches. The method in [5] used bag-of-words, Part-Of-Speech information and sentence position as features for analyzing reviews, representing reviews as feature vectors. In [6], the problem of attributing a numerical score (one to five stars) to a review is presented using Naïve Bayes and SVM.

Among all the text categorization method, there is still a lack of investigation on sentiment classification based on association rules. Association rule based classifiers (associative classifier) originate from association rule mining task of data mining. Since association rules reflect strong associations between items and includes more underlying semantic and contextual meaning than individual word, it has been developed within the text mining domain in different aspects [7] [8] [9].

In this paper, we investigate the association rules in sentiment classification problem. The motivation is to the convert this general classification approach into a binary and special domain classification problem. The main contributions of this work involve: first, we introduce the AD-Sup metric to extract discriminative frequent term sets and eliminating terms with no sentiment orientation which contain close frequency in both positive and negative reviews; Second, optimal rule set is generated to abandon the redundant rules to construct classifier; Third, in rules matching phase, we propose a new metric named Maximum Term Weight and a multiple metric voting scheme to solve the problem when matched rules are not applicable or not confident enough.

2 Proposed Method

The proposed method includes four steps: (1) Data pre-processing and feature selection; (2) Frequent term set extraction and rule mining; (3) Mining optimal classification rules; (4) Predicting test review with multiple metric voting.

Before extracting frequent term set, feature selection is conducted for dimension reduction using Information Gain (IG) [10]. In this study, the magnitude of dimensions is reduced from 10^5 to 10^3. An important reason that makes feature selection essential is that the number of frequent term sets extracted from single terms grows exponentially when the input terms increase. So, the input number of single terms must be restricted to a reasonable scale.

2.1 Frequent Term Set Extraction

In text mining issues, each document d in D = $\{d_1,d_2,\ldots d_n\}$ is treated as a transaction and the set of terms T = $\{t_1,t_2\ldots t_m\}$ contained in D corresponds to the items set. A term set S in T is frequent if $Sup(S) \geq min\text{-}sup$. The *min-sup* constraint of term set is a key measure for frequent sets extraction because it determines the scale and quality of the selected frequent sets. When applied to text mining problem, the concept of support count corresponds to Document Frequency (DF). However, support count cannot be simply substituted by DF because DF only measures the occurrences and this is not sufficient to differentiate the discriminative effect of the frequent term sets. To solve this problem, our previous work proposed a new metric Average Deviation Support (AD-Sup), considering the distribution discrepancy of term sets in each class. Assume the documents set have n classes $\{class_1,\ldots class_i,\ldots class_n\}$ and let FS denote the term set and t is the term in FS, AD-Sup can be formulated as :

$$AD\text{-}Sup(FS) = \frac{\sqrt{\sum_{i=1}^{n}\{Sup(FS)_i - Ave(Sup(FS))\}^2}}{Ave(Sup(FS))} \tag{1}$$

The expression of AD-Sup (1) can be deemed as a modified support deviation, where $Sup(FS)_i$ means the local support of FS in class *i* and Ave(Sup(FS)) denotes the average value of Sup(FS) in all the classes. The frequent term extraction procedure is implemented using Apriori strategy [8]. After obtaining all the frequent term sets (FS), AD-Sup restraint is used to select the frequent features. The selected FS will involve more term sets that are not only frequent but distributed unevenly in different classes.

2.2 Optimal Rule Mining

Following the extraction of frequent term sets, an association rule is an implication denoted by $X \Rightarrow Y$, where both X and Y are subset from a frequent set. For classification problem, the consequent of the rules are class labels. However, it suffers

the following problem:(1) the confidence and support restraint is not always suitable for any mining pr paper problem;(2) the number of association rules are usually too large which makes the pruning quite challenging;(3) the vast amount of association rules involve much redundant information. To overcome these obstacles, many interesting metrics and pruning strategies have been proposed to find "optimal rules". There is no standard definition for "optimal rules". This paper utilizes a similar strategy close to [11]:

Definition (optimal association rule set): A rule set is optimal with respect to an interestingness metric if it contains all rules except those with no greater interestingness than one of its more general rules. Given two rules $P \Rightarrow C$ and $Q \Rightarrow C$ where $P \subset Q$, the latter is more specific than the former and the former is more general than the latter.

2.3 Predicting Sentiment Class by Association Rules

For general association rule based classification, the classifier is a collection of selected rules using different rule matching strategies. However, for sentiment classification, these strategies may fail to predict correctly in some cases. Given covered rule sets with positive and negative classification rules, following examples are hard to judge the sentiment: First, the covered positive rules have the highest confidence but the difference with that of the negative rules is quite small, while the number of negative rules is much more than that of the positive rules. In this case, the test review should be negative rules but it will be predicted to be positive. Second, the number of covered rules and the max-confidence of rules are both equal. Third, the situation can be more complicated, where the negative rules have a higher confidence with a little priority than the positive but besides the highest confidence rules, comparing other covered rules, the positive rules are more persuasive.

To overcome these obstacles, borrowing the ideas of democratic regime, we propose a new association rule based class predicting method by a voting scheme of the following metrics. The class label of test review will be determined by a combination of voting score.

$$Score(test\_review_i) = \sum_{0}^{m} Vote(metric_j) \tag{2}$$

The vote on a metric is 1,-1,or 0 depending on the different metric value on covered positive rules (PR) and negative rules (NR). Vote(metirc$_j$)=1 if metric$_j$(PR)>metirc$_j$(NR), and respectively, if metric$_j$(PR)<metirc$_j$(NR), Vote(metirc$_j$)=-1, if metric$_j$(PR)=metirc$_j$(NR), Vote(metirc$_j$)=0. The assigned class label depends on whether vote score of metrics is a positive number or negative number.

The metrics that are evaluated here include:

Definition 1 (Max-conf): the highest confidence value of covered rules.

Definition 2 (Cover-len): the number of rules in the covered rule set.

Definition 3 (Minor-conf): the average confidence of covered rules excluding the highest one.

All the above 3 metrics can be obtained by counting the recorded value generated in the rule mining procedure. However, sometimes there is no promising rule with a high confidence in the covered rules. When the Max-conf is not very high or the Max-conf in PR and NR are very close, it is hard to predict if it belongs to any category. To solve this problem, we propose a new metric named MTW (Max Term Weight):

Definition 4 (MTW): Maximum Term Weight, the average term weight of each rule clusters in covered rules. In this paper, we use the information gain (IG) of each term obtained in the frequent term set mining procedure for its weight. A rule cluster is a collection of rules which contains the same term. Given a covered rule set, the algorithm to get max term weight can be described as follows:

Algorithm 1: MTW metric generation
Input: single term set (TS) in descending order
For each term T_i in TS:
If covered rule set is not empty
For each rule in covered rule set:
If($Rule_j$ contains T_i)
Add $Rule_j$ to rule cluster(RC_i);
Set:weight(RC_i)=GetTermWeight(T_i);
Delete $Rule_j$ in covered rule set;
end If
end For
end If
end For

Return: $Average(\sum_{i}^{k} weight(RC_i))$

The motivation of MTW is to make use of discriminative measurement of single terms contained in the covered rules. Besides IG, similar measurement like TF*IDF and χ^2 are also applicable here.

3 Experimental Results

In this paper, Multi-Domain Sentiment Dataset [12] is used for experimental evaluation. This dataset contains product reviews taken from Amazon.com from many product types (domains). We selected four domains of this datasets: DVD, Book, Kitchen and Electronic. Each domain contains 1000 positive and 1000 negative reviews. All the above reviews are pre-processed by stemming and stop-word elimination. The evaluation is conducted through 3-folds cross validation.

3.1 Frequent Term Sets Extraction

The frequent term sets extraction starts from the selected single terms by IG. The first scan generates frequent term sets with two terms and these double term sets are used

as the input term sets for the candidate-generating algorithm. Then the iteration starts until no candidate is selected to be frequent term set. In this paper, input number of single terms is set to be 600 and *min-sup* is 2%.Table 1 reports more details of selected terms and extraction results. The top single terms are ranked by its IG value, followed by the frequent term sets by support count and refined term sets by AD-Sup. Note that the stemming changes the form of the words and makes the stemmed terms appear to be different from the real words. On Electronic dataset, all the top 5 frequent term sets by *min-sup* contain word "i". They are selected due to their high occurrences, but "i" is a common pronoun without discriminating value. We can observe similar results in all the four datasets. Contrarily, term sets selected by AD-Sup contain more sentiment oriented words and also have better consistency with the most important single terms. A *min-AD-Sup* threshold is then used to prune these un-discriminative term sets before mining classification rules.

Table 1. Top Single term and frequent term sets in Electronic and Kitchen dataset

Top5 terms	Frequent termSets	Refined Term Sets	Top5 terms	Frequent termSets	Refined Term Sets
great return	us/i work/i	terribl/i/ refund/i	easi return	num/my my/so	return/time/first/ worst/ever
excel	i/get	worst/i	great	num/so	wast/monei/do
price	i/all	return/i/bui	love	my/time	wast/monei/even
wast	i/when	wast/work	disappoint	my/get	wast/monei/what

We set *min-conf* as 50% in this experiment to extract classification rules. Table 2 lists the top classification rules of the two datasets.

Table 2. Top positive (Pos) and negative (Neg) rules for Electronics and Kitchen dataset

Pos Rules	Neg Rules	Pos Rules	Neg Rules
excel/price perfect/i/us	terribl/i refund/i	easi/so/love easi/num/love	wast/monei return/product
perfect/us	return/bui	easi/great/love	return/bui
price/good/well	worst/i/	love/dishwash	return/again
great/perfect	support/call	best/so/can	worst/ever

3.2 Sentiment Prediction Using Multiple Metric Votes

Figure 1 is the F_1 value comparison of classification result on the four dataset. To demonstrate the effectiveness of the metric voting method, we select two baseline algorithms: the Single Rule tries to match the maximum confidence rule of the covered rules set, and the Multi-Rules compares maximum confidence + average

confidence value. The results show that in all the four datasets, our strategy outperforms the other two baselines. The maximum improvement of 2.8% was obtained on DVD. On each dataset, all the algorithms were implemented with different *min-conf*. The results also prove that 50% is the best *min-conf* to generate classification rules. When the *min-conf* increases to 70% and 80%, the performance declined rapidly because the number of rules decreases to a very little scale and makes too many test documents covered by none of the rules.

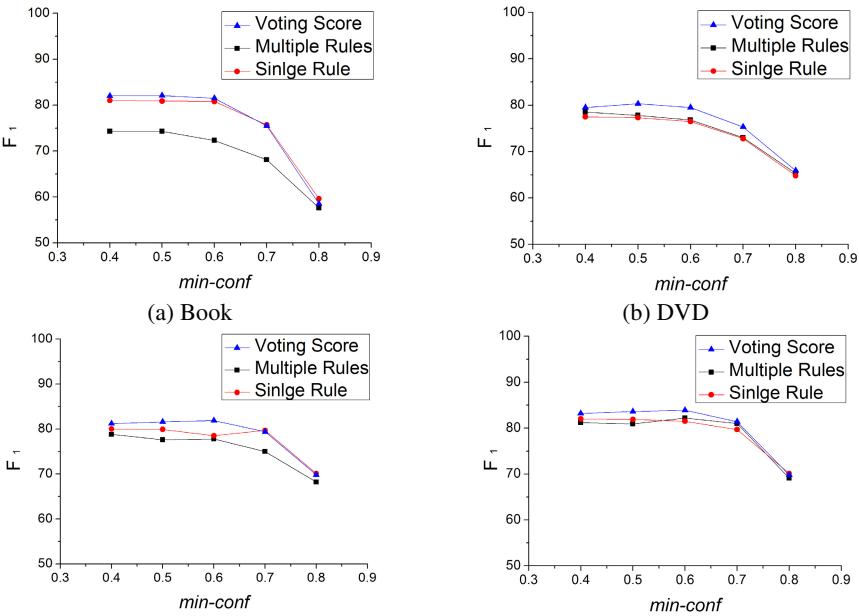

Fig. 1. Classification results of different strategies on four datasets

Table 3. Classification results vs. other classifiers : F_1 (%)

Dataset	SMO	LibSVM	NB	kNN	Voting Score
Book	77.1	81.8	77.7	66.6	82.1
DVD	78.7	76.0	77.1	66.4	80.3
Electronic	82.5	75.6	73.5	67.4	81.9
Kitchen	82.1	81.9	75.6	58.5	83.9

Table 3 summarizes the classicization results of the proposed method comparing with other popular machine learning classifiers. SVM, Naïve Bayes (NB), kNN are well studied text document classifiers with very good performance track in many previous researches. SVM was implemented with two algorithms, LibSVM and SMO.

In three of the four datasets, the F_1 value of multiple metric voting strategy surpassed the other four benchmark algorithms, except for Electronic where the result of our method is very close to SMO.

4 Conclusions

This paper has presented a novel approach using association rules for sentiment classification of web reviews. To extract discriminative frequent term sets, a new restraint measure AD-Sup was used which considers more on the term set distribution on different sentiment classes. The experiment results on multiple domain reviews from real web sites demonstrated that AD-Sup was an effective metric to eliminate terms with no sentiment orientation which contain close frequency in both positive and negative reviews. An optimal classification rule set was generated which abandons the redundant general rule with lower confidence than the specific one. In the class label prediction procedure, we proposed a new metric voting scheme to solve the problem when the covered rules are not adequately confident or not applicable. The final score of a test review depends on the overall contributions of four metrics. To demonstrate the effectiveness of the voting strategy, we compared the classification performance with different baselines of using confidence and the result shows 50% is the best minconf to guarantee classification rules both abundant and persuasive, and the voting method obtains improvements on all the four datasets. We also compared the proposed method with popular machine learning algorithms such as SVM, Naïve Bayes and kNN. The result also shows that our strategy is effective and outperforms the other strong benchmarks. Since this research focused on binary classification, our future work will concentrate on a further optimization and the extension to multiple label classification problem.

Acknowledgements. We are grateful to Shenzhen Key Laboratory of Data Vitalization (Smart City) for supporting this research. This work was also supported by National Natural Science Foundation of China (61103095), International S&T Cooperation Program of China (2010DFB13350) and National High Technology Research, Development Program ("863" Program) of China (2011AA010502) and Fundamental Research Funds for the Central Universities

References

1. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys 34(1), 1–47 (2002)
2. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: ACL 2002 Conference on Empirical Methods in Natural Language Processing, pp. 79–86 (2002)
3. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: The 42nd Annual Meeting of the Association for Computational Linguistics, pp. 271–278 (2004)

4. Thomas, M., Pang, B., Lee, L.: Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In: ACL 2005 Conference on Empirical Methods in Natural Language Processing, pp.327–335 (2006)
5. Baccianella, S., Esuli, A., Sebastiani, F.: Multi-Facet Rating of Product Reviews. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) ECIR 2009. LNCS, vol. 5478, pp. 461–472. Springer, Heidelberg (2009)
6. Prasad, P., Vasudeva, V.: Published Supervised Learning Approaches for Rating Customer Reviews. Journal of Intelligent Systems 19(1), 79–94 (2010)
7. Mahgoub, H.: Mining Association Rules from Unstructured Documents. In: The 3rd International Conference on Knowledge Mining, pp.167–172 (2006)
8. Thabtah, F.: A review of associative classification mining. The Knowledge Engineering Review 22, 37–65 (2007)
9. Han, J., Cheng, H., Xin, D., Yan, X.: Frequent pattern mining: current status and future directions. Data Mining and Knowledge Discovery 15, 55–86 (2007)
10. Forman, G.: An Extensive Empirical Study of Feature Selection Metrics for Text Classification. Journal of Machine Learning Research 3, 1289–1305 (2003)
11. Li, J.: On Optimal Rule Discovery. IEEE Transactions on Knowledge and Data Engineering 18(4), 460–471 (2006)
12. http://www.cs.jhu.edu/~mdredze/datasets/sentiment