Categorical Probability Proportion Difference (CPPD): A Feature Selection Method for Sentiment Classification

Basant Agarwal, *Namita Mittal*
Department of Computer Engineering,
Malaviya National Institute of Technology, Jaipur, India
thebasant@gmail.com, nmittal@mnit.ac.in

ABSTRACT

Sentiment analysis is to extract the opinion of the user from of the text documents. Sentiment classification using machine learning methods face problem of handling huge number of unique terms in a feature vector for the classification. Thus it is required to eliminate the irrelevant and noisy terms from the feature vector. Feature selection methods reduce the feature size by selecting prominent features for better classification. In this paper, a new feature selection method namely Probability Proportion Difference (PPD) is proposed which is based on the probability of belongingness of a term to a particular class. It is capable of removing irrelevant terms from the feature vector. Further, a Categorical Probability Proportion Difference (CPPD) feature selection method is proposed based on Probability Proportion Difference (PPD) and Categorical Proportion Difference (CPD). CPPD feature selection method is able to select the features which are relevant and capable of discriminating the class. The performance of the proposed feature selection methods is compared with the CPD method and Information Gain (IG) method which has been identified as one of the best feature selection method for sentiment classification. Experimentation of proposed feature selection methods was performed on two standard datasets viz. movie review dataset and product review (i.e. book) dataset. Experimental results show that proposed CPPD feature selection method outperforms other feature selection method for sentiment classification.

KEYWORDS : Feature Selection, Sentiment Classification, Categorical Probability Proportional Difference (CPPD), Probability Proportion Difference (PPD), CPD.

Proceedings of the 2nd Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2012), pages 17–26,
COLING 2012, Mumbai, December 2012.

1. Introduction

With the rapid growth of web technology, people now express their opinion, experience, attitude, feelings, and emotions on the web. So, it has increased the demand of processing, organizing, and analyzing the web content to know the opinion of the users (Pang B. and Lee L., 2008). An automatic sentiment text classification means to identify the sentiment orientation of the text documents i.e. positive or negative. It is important for users as well as companies to know the opinion of users, for example review for electronic products like laptop, car, movies etc. can be beneficial for users to take decision on which product to purchase and for companies to improve and market their products.
Various researchers have applied machine learning algorithms for sentiment analysis (Pang B. and Lee L., 2004; Tan S. and Zhang J., 2008; Pang B. and Lee L., 2008). One of the major problems in sentiment classification is to deal with huge number of features used for describing text documents, which produces hurdles to machine learning methods in determining the sentiment orientation of the document. Thus, it is required to select only prominent features which contribute majorly in the identification of sentiment of the document. The aim of feature selection methods is to produce the reduced feature set which is capable of determining sentiment orientation of the document by eliminating irrelevant and noisy features.
Various feature selection methods has been proposed for selecting predominating features for sentiment classification, for example Information Gain (IG), Mutual Information (MI), Chi square (CHI), Gain Ratio (GR), Document Frequency (DF) etc. (Tan S. and Zhang J., 2008; Pang B. and Lee L., 2008).

In the proposed approach, feature selection methods are used for improving the performance of the machine learning method. Initially, binary weighting scheme is used to represent the review documents, and then various feature selection methods are applied to reduce the feature set size. Further, machine learning methods are applied to the reduced and prominent feature set.

Our contribution:

1. Two new feature selection methods i.e. PPD and CPPD are proposed for sentiment classification.

2. Compared the performance of proposed feature selection methods on two different standard datasets of different domains.

The paper is organized as follows: A brief discussion of the related work is given in Section 2. Feature selection methods used for sentiment classification are discussed in Section 3. Dataset, Experimental setup and results are discussed in Section 4. Finally, conclusions and future work is described.

2. Related work

Machine learning methods have been widely applied for sentiment classification (Pang B. and Lee L., 2004; Tan S. and Zhang J., 2008; Pang B. and Lee L., 2008). Pang *et al.* 2002, applied machine learning methods viz. Support Vector Machine (SVM), Naïve Bayes (NB), and Maximum Entropy (ME) for sentiment classification on *unigram* and *bigram* features of movie review dataset. Authors found SVM to be performed best among classifiers. Authors also found that binary weighting scheme outperforms Term Frequency (TF) method for representing the text for sentiment classification. Later, a minimum cut method is proposed to eliminate objective

sentences from the text (Pang B. and Lee L., 2004), which showed improved performance. Authors (Tan S. and Zhang J., 2008), experimented on five machine learning algorithms i.e. K-nearest neighbour (KNN), Centroid classifier, Winnow classifier, NB and SVM with four feature selection methods those are MI, IG, CHI, and DF for sentiment classification on Chinese documents. Authors observed that IG performs best among all the feature selection methods and SVM gives best results among machine learning algorithms.

Various feature selection methods have been proposed by various researchers for reducing the feature vector for sentiment classification for improved performance of machine learning methods (Tan S. and Zhang J., 2008; Pang B. and Lee L., 2008). Entropy Weighted Genetic Algorithm (EWGA) is proposed by combining the IG and genetic algorithm, which improved the accuracy of sentiment classification (Abbasi *et al.* 2008). Sentiment features are highlighted by increasing their weights, further authors used multiple classifiers on various feature vectors to construct the aggregated classifier (Dai *et al*. 2011). O' keefe *et al.* 2009, compared three feature selection methods for sentiment classification, which are based on Categorical Proportional Difference (CPD) and Sentiment Orientation (SO) values. Wang *et al.* 2009, proposed Fisher's discriminant ratio based feature selection method text review sentiment classification.

3. Feature selection methods

Feature selection methods select prominent features from the high dimensional feature vector by eliminating noisy and irrelevant features. Optimal feature vector improves the performance of the machine learning method in terms of both accuracy and execution time.

3.1 Probability Proportion Difference (PPD)

Probability Proportion Difference (PPD) measures the degree of belongingness or probability that a term belongs to a particular class.

Algorithm 1: Probability Proportion Difference (PPD) Feature Selection Method

Input: Document corpus (D) with labels (C) positive or negative, k (number of Optimal features to be selected)
Output: OptimalFeatureSet
Step 1 ***Preprocessing***
$t \leftarrow$ ExtractUniqueTerms(D)
$F \leftarrow$ TotalUniqueTerms(D)
$W_p \leftarrow$ TotalTermsInPositiveClass(D,C)
$W_n \leftarrow$ TotalTermsInNegativeClass(D,C)
Step 2 ***Main Feature Selection loop***
for each $t \in \mathbf{F}$
N_{tp} =CountPositiveDocumentsInwhichTermAppears(D,t)
N_{tn} =CountNegativeDocumentsInwhichTermAppears(D,t)
end for
for each $t \in \mathbf{F}$

$$ppd = \frac{N_{tp}}{W_p+F} - \frac{N_{tn}}{W_n+F}$$

end for
OptimalFeatureSet $\leftarrow$ SelectTopTerm(k)

If a term has high probability of belongingness to dominantly one category/class (i.e. positive or negative) that indicates the term is important in identifying the category of unknown review. And if a term has almost equal probability of belongingness to both the categories, in that case the term is not useful in discriminating the class. PPD value of a term is calculated by computing the difference of probabilities that a term will belong to positive class or negative class. Thus, if a term has high PPD value, it indicates that the term is important for sentiment classification. Probability of belongingness of a term depends on the number of documents in which a term appears and number of unique terms appeared in that class. Algorithm for calculating PPD value of a term is given in Algorithm 1. Top k features can be selected on the basis of PPD value of the term.

3.2 Categorical Proportion Difference (CPD)

Categorical Proportional Different (CPD) value measures the degree to which a term contributes in discriminating the class (Simeon *et al.* 2008). O'Keefe *et al.* 2009, have used CPD value for feature selection method. CPD value of a term is computed by finding the ratio of the difference between the number of documents of a category in which it appears and the number of documents in which it appears of another category, to the total number of documents in which that term appears. CPD value for a feature can be calculated by using equation 1.

$$\text{cpd} = \frac{|\text{posD-negD}|}{\text{posD} + \text{negD}} \quad \ldots. (1)$$

Here, posD is the number of positive review document in which a term appears, and negD is the number of negative review documents in which that term appear. Range of CPD value is 0 to1. If any term appears dominantly in positive or negative class, then that feature is useful for the sentiment classification, and if a term is occurring in both the categories equally then that feature is not useful for classification. If CPD value of a feature is close to 1 it means that this feature is occurring dominantly in only one category of documents. For example if "Excellent" word is occurring in 150 positive review documents and in 2 negative review documents, then value of this feature will be (150-2)/(150+2)= 0.97 , its value is near to 1 indicates that this term is useful in identifying the class of unknown document. It indicates that if a new document is having "excellent" word, there is a high chance that this document belongs to positive category. Similarly if a word occurs in same number of positive and negative documents, then CPD value will be 0, which indicates that this term is not useful for classification.

3.3 Categorical Probability Proportion Difference (CPPD)

Categorical Probability Proportion Difference (CPPD) based feature selection methods combines the merits and eliminates the demerits of both CPD and PPD methods. Benefit of CPD method is that it measures the degree of class distinguishing property of a term, which is an important attribute of a prominent feature. It can eliminate terms, which are occurring in both the classes equally and are not important for classification. It can easily eliminate the terms with high document frequency but are not important like stop words. However, PPD value of term indicates the belongingness/relatedness of a term to the classes and difference measures the class discriminating ability. It can remove the terms with less document frequency, which is not important for sentiment classification like rare terms. PPD feature selection method also considers the documents length of positive and negative reviews, since generally positive orientation documents are more in length as compared to negative class documents. So, there is a

high probability that most of the feature selection method select more positive sentiment words, as compared to negative sentiment words that result in less recall. However, in the proposed CPPD method, length of documents is considered in computing the CPPD value.

Demerits of CPD feature selection method is that it can include rare term with less document frequencies but not important, which will be eliminated by PPD method. Similarly, PPD feature selection method may include term with high document frequency but not important, which will be removed by CPD method. So, by combining the merits and removing the demerits of CPD and PPD feature selection, a more reliable feature selection method is proposed for sentiment classification. CPPD feature selection method is described in algorithm2.

Algorithm 2: Categorical Probability Proportion Difference (CPPD) Feature Selection Method

Input: Document corpus (D) with labels (C) positive or negative
Output: ProminentFeatureSet
Step 1 ***Preprocessing***
$t \leftarrow$ ExtractUniqueTerms(D)
$F \leftarrow$ TotalUniqueTerms(D)
$W_p \leftarrow$ TotalTermsInPositiveClass(D,C)
$W_n \leftarrow$ TotalTermsInNegativeClass(D,C)
Step 2 ***Main Feature Selection loop***
for each $t \in$ **F**
N_{tp} =CountPositiveDocumentsInwhichTermAppears(D,t)
N_{tn} =CountNegativeDocumentsInwhichTermAppears(D,t)
end for
for each $t \in$ **F**
$cpd = \frac{N_{tp}-N_{tn}}{N_{tp}+N_{tn}}$
$ppd = \frac{N_{tp}}{W_p+F} - \frac{N_{tn}}{W_n+F}$
if (*cpd >T1 && ppd>T2*)
ProminentFeatureSet $\leftarrow$ SelectTerm(t)
end for

3.4 Information Gain (IG)

Information Gain has been identified as one of the best feature selection method for sentiment classification (Tan S. and Zhang J., 2008). Therefore, we compared proposed feature selection methods with IG. Information gain (IG) is a feature selection method, which computes importance of a feature with respect to class attribute. It is measured by the reduction in the uncertainty in classification when the value of the feature is known (Forman G. 2003). Top ranked features are selected for reducing the feature vector size in turn better classification results. IG of a term can be calculated by using equation 2 (Forman G. 2003).

$$\mathrm{IG(t)} = -\sum_{j=1}^{K} \mathrm{P}(C_j)\log \mathrm{P}(C_j) + \mathrm{P(w)} \sum_{j=1}^{K} \mathrm{P}(C_j|\mathrm{w})\log \mathrm{P}(C_j|\mathrm{w}) + \mathrm{P}(\overrightarrow{\mathrm{w}}) \sum_{j=1}^{K} \mathrm{P}(C_j|\overrightarrow{\mathrm{w}}) \log \mathrm{P}(C_j|\overrightarrow{\mathrm{w}}) \quad ..(2)$$

Here, $P(C_j)$ is the fraction of number of documents that belongs to class C_j out of total documents and $P(w)$ is fraction of documents in which term w occurs. $P(C_j|w)$ is computed as fraction of documents from class C_j that have term w and $P(C_j|\overrightarrow{w})$ is fraction of documents from class C_j that does not contain term w.

4. Experimental Setup and Result Analysis

4.1 Dataset and Experiments

One of the most popular publically available standard movie review dataset is used to test the proposed feature selection methods (Pang B., and Lee L., 2004). This standard dataset, known as Cornell Movie Review Dataset is consisting of 2000 reviews that contain 1000 positive and 1000 negative labeled reviews. In addition, product review dataset (book reviews) consisting amazon products reviews has also been used (Blitzer *et al.* 2007). This dataset contains 1000 positive and 1000 negative labeled book reviews.

Documents are initially pre-processed as follows:

(i) Negation handling, "NOT_" is added to every words occurring after the negation word (no, not, isn't, can't etc.) in the sentence. Since, a negation word inverts the sentiment of the sentence (Pang B. and Lee L., 2002).

(ii) Terms which are occurring in less than 2 documents are removed from the feature set.

The feature vector generated after pre-processing is further used for the classification. Binary weighting scheme is used for representing text since it has been proved the best method for sentiment classification (Pang B. and Lee L., 2002).

Among various machine learning algorithms Support Vector Machine (SVM) and Naïve Bayes (NB) classifiers are mostly used for sentiment classification (Pang B. and Lee L., 2002; O'Keefe *et al.* 2009; Abbasi *et al*. 2009; Pang B. and Lee L., 2008). So, in our experiments, SVM and NB are used for classifying review documents into positive or negative class. Evaluation of classification results is done by 10 folds cross validation (Kohavi R., 1995). Linear SVM and Naïve Bayes are used for all the experiments with default setting in weka machine learning tool (WEKA).

4.2 Performance measures

To evaluate the performance of sentiment classification with various feature selection methods, F-measure (given in equation 3) is used. It combines precision and recall, which are commonly used measure. Precision for a class C is the fraction of total number of documents that are correctly classified to the 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)).

$$F - Measure = \frac{2 * precision * recall}{(precision + recall)} \quad \ldots\ldots\ldots \quad (3)$$

4.3 Results and discussions

Some cases have been selected from movie review dataset and discussed. CPD and PPD values of some of cases have been shown in Table 1. CPD feature selection method has the drawback that less document frequent term can have very high CPD value, which is not important for classification. For example, if a term is having positive DF of 3 and negative DF of 0, then CPD value will be 1, which is maximum CPD value, even if the feature is not that important (refer case 1 of Table1). Similarly, if a term has positive DF of 1 and negative DF of 6, then CPD value comes out to be 0.714, which is quite high but the feature is not that important for classification (refer case 2 of Table1). This drawback is removed by using PPD feature selection method. Since, these types of terms have very low PPD value, so eliminated by PPD feature selection method. Also, in movie review dataset the term "poor" has low CPD value which is very important term for sentiment classification (refer case 3 of Table 1). This term will be eliminated by CPD method but would be selected by PPD method.

Similarly, cases 4, 5, 6, of Table1 for terms "Oscar", "perfect", and "bad" respectively are important for sentiment classification, which are eliminated by CPD method but included by PPD method. In contrary, few terms with high DF would have high PPD value, but not important. These terms are eliminated by CPD method. For example, In Table 1 case 7 shows PPD value high for term "because", it is eliminated by CPD method, but PPD value is high. It is due to the fact that PPD value depends on the DF and total terms in each class of the corpus. In this example, document length of positive reviews is larger as compared to length of negative reviews that is why the PPD value is high.

Cases	Positive DF	Negative DF	CPD	PPD
1	3	0	1	0.001
2	1	6	0.714	0.0016
3	57	122	0.36	0.025
4	137	62	0.375	0.024
5	201	94	0.362	0.03
6	260	515	0.329	0.099
7	461	461	0	0.011

TABLE 1. Case study of movie review dataset with different terms

Finally, by combining PPD and CPD method, a new feature selection method CPPD is proposed, which selects important features by considering the class distinguishing ability of a term and relevancy of a term based on probability with taking the size of negative and positive documents into consideration.

4.3.1 Comparison of feature selection methods

F- Measure for sentiment classification with various feature selection methods are shown in Table 2. *Unigram* feature set without any feature selection method is taken as baseline accuracy. It is observed from the experiments that all the feature selection methods improve the performance of both the classifiers (SVM and NB) as compared to baseline performance.

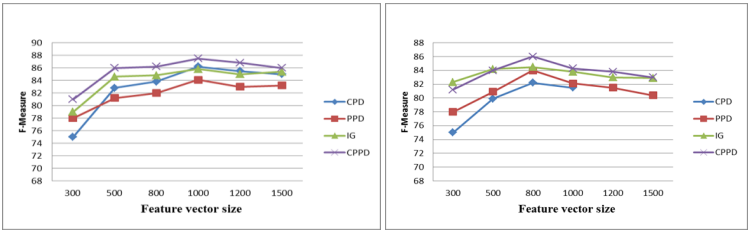
With CPPD feature selection method, F-measure of *unigram* feature set improves from 84.2 % to 87.5% (+3.9%) for SVM classifier and from 79.4% to 85.5 % (+7.6%) for NB classifier for movie review dataset. For book review dataset, F-measure significantly improves from 76.2% to 86% (+12.8%) for SVM classifier and from 74.5% to 80.1% (+7.5%) for NB classifier. With PPD feature selection method, F-measure improves for unigram features from 79.4% to 85.2% (+7.3%) for NB classifier and remains almost same for SVM classifier on movie review dataset.

	Movie reviews		Book reviews	
Features	SVM	NB	SVM	NB
Unigram	84.2	79.4	76.2	74.5
IG	85.8(+1.9%)	85.1(+7.1%)	84.5(+10.8%)	76.3(+2.4%)
CPD	86.2(+2.3%)	82.1(+3.4%)	82.2(+7.6%)	77.2(+3.6%)
PPD	84.1(-0.11%)	85.2(+7.3%)	84(+10.2%)	79(+6.0%)
CPPD	87.5(+3.9%)	85.5(+7.6%)	86(+12.8%)	80.1(+7.5%)

TABLE 2. F-Measure (%) for various feature selection method

4.3.2 Effect of different feature size on classification results:

F-Measure values for different feature size with various Feature Selection (FS) method for SVM classifier using movie review and book review dataset in shown in Figure 1.



(a) (b)

FIGURE 1**.** (a) F-Measure (%) for various FS methods with SVM on Movie review (b) F-Measure (%) for various FS methods with SVM on Book review dataset.

It is observed from Figure 1 that CPPD method outperforms other feature selection methods. As feature size increases F-measure increases upto a certain limit, after that it varies within a small range. Best F-measure is observed for 1000 and 800 features respectively for movie review and book review dataset, which are approximately 10-15% of total unigram features.

Conclusion

Prominent feature selection for sentiment classification is very important for better classification results. In this paper, two new feature selection methods are proposed PPD and CPPD. These are compared with other FS methods namely CPD and IG. Proposed CPPD feature selection method is computationally very efficient and filters irrelevant features. It selects relevant features to the class and which can contribute in discriminating classes. The proposed schemes are evaluated on two standard datasets. Experimental results show that proposed method improves the classification performance from the baseline results very efficiently. Proposed CPPD feature selection method performs better as compared to other feature selection methods. In future, we wish to evaluate the proposed scheme on various datasets of various domains and for non-English documents.

References

Abbasi A., Chen H.C., and Salem A. (2008). "Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums". *In ACM Transactions on Information Systems (TOIS),* 2008. 26(3).

Blitzer J., Dredze M., Pereira F., (2007). "Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification", *Proc. Assoc. Computational Linguistics. ACL Press, 2007,* pp 440-447.

Dai L., Chen H., and Li X., (2011). "Improving sentiment classification using feature highlighting and feature bagging", *In 11th IEEE International conference on Data Mining Workshops,* pp.61-66.

Forman G., (2003). "An extensive empirical study of feature selection metrics for text classification". *JMLR,* 3: pp 1289–1306.

Kohavi R., (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection", *IJCAI'95 Proceedings of the 14th international joint conference on Artificial intelligence - Vol 2,* pp1137-1143.

O'Keefe T., Koprinska I., (2009). "Feature Selection and Weighting Methods in Sentiment Analysis", In *Proceedings of the 14th Australasian Document Computing Symposium.*

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

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

Pang B., Lee L., (2004). "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.

Simeon M., Hilderman R., (2008). “Categorical proportional Difference: A feature selection method for text categorization”, *In Proceedings of the 17th Australasian Data Mining Conference*, pages 201-208.

Tan S., Zhang J., (2008). "An empirical study of sentiment analysis for chinese documents”, In *Expert Systems with Applications,* vol. 34, pp. 2622-2629.

Wang S., Li D., Wei Y., Li H.,(2009). “A Feature Selection Method based on Fisher's Discriminant Ratio for Text Sentiment Classification”, *In Proceeding WISM '09 Proceedings of the International Conference on Web Information Systems and Mining,* pp 88- 97.

WEKA.Open Source Machine Learning Software Weka. http://www.cs.waikato.ac.nz/ml/weka/.