Sentiment Classification through Combining Classifiers with Multiple Feature Sets

Shoushan Li and Chengqing Zong
National Laboratory of Pattern Recognition,
Institute of Automation
Chinese Academy of Sciences, Beijing
100080, China
{sshanli,cqzong}@nlpr.ia.ac.cn

Xia Wang
Nokia Research Center, Beijing,
100013, China
xia.S.wang@Nokia.com

Abstract

Sentiment classification aims at assigning a document to a predefined category according to the polarity of its subjective information (e.g. 'thumbs up' or 'thumbs down'). In this paper, we present a classifier combination approach to this task. First, different classifiers are generated through training the review data with different features: unigram and some POS features. Then, classifier selection method is used to select a part of the classifiers for the next-step combination. Finally, these selected classifiers are combined using several combining rules. The experimental results show that all the combination approaches with different combining rules outperform individual classifiers and the sum rule achieves the best performance with an improvement of 2.56% over the best individual classifier.

1 Introduction

In the recent years, subjective analysis has drawn a growing interest in the research field of computational linguistics. Subjective analysis is concerned with the seeking information about the opinions, feelings and attitudes expressed in a text, rather than just the facts. A key task in this area is sentiment classification which aims at classifying a document according to the polarity of its subjective information ('thumbs up' or 'thumbs down'). This problem benefits many potential applications, such as automation classification of movie reviews (Pang et al., 2002) or product reviews (Cui et al. , 2006), question answering (Kim and Hovy, 2005), and automation summarization (Ku et al., 2006).

Pang et al. (2002) are the first to apply some machine learning methods, instead of traditional rule-based methods, to sentiment classification. In machine learning methods, one document is usually represented as bag-of-features. Although these methods have been proved to be effective for sentiment classification, there appears to remain considerable room for improvement (Pang and Lee, 2004). A straightforward way to improve the learning performance is to use new types of features, such as fixed-length n-grams (e.g., word bi-gram or tri-gram) or part-of-speech (POS) features besides the traditional bag-of-words features. Riloff et al. (2006) creates a subsumption hierarchy that defines the representational scope of different types of features. Based on different feature sets, classifiers lead to different classification performances. The traditional way is to choose the classifier with the best-performance feature set for the final decision. Note that the misclassified samples by these different classifiers may be not overlap. Thus, an alternative (may be better) way is to combine some of these classifiers instead of choose only one classifier.

In this paper, we explore and identify the benefits of classifier combination for sentiment classification through combining different classifiers based on different features. Classifier combination is an effective and broadly useful method for improving system performance. It is designed to combine multiple classifiers as to take advantage of the strengths of individual classifiers and have been successfully applied to various fields of com-

putational linguistics including text categorization (Switzerland, 1996), named entity recognition (Florian et al., 2003), word sense disambiguation (Klein et al., 2002), etc. To our best knowledge, there have been very few attempts to apply classifier combination to sentiment classification. Kennedy and Inkpen (2006) combine two systems based on two different classification algorithms: the term-counting and machine learning method (SVM method). They find that combining the two systems slightly improves the results, getting an improvement of 1.3% over using only the SVM method. Different from their work, we attempt to combine different classifiers based on different feature sets which are known as a requirement for a good-performance combination (Kittler et al., 1998).

The remainder of this paper is organized as follows. Section 2 briefly introduces the classifier combination methods. Section 3 describes the implementation of classifier combination on sentiment classification. Experimental results are presented and analyzed in Section 4. Finally, Section 5 draws some conclusions and outlines the future work.

2 Classifier Combination

Classifier combination has been studied intensively in the last decade, and has been shown to be successful in improving performance on diverse applications. Generally, the construction of a multiple classifier system (MCS) consists of three main steps, i.e. training a number of component classifiers, selecting some 'good' classifiers for further combination, and using combining rules to integrate the results from these component classifiers for the final decision. Figure 1 shows a traditional framework of a multiple classifier system.

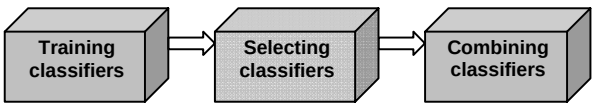

Figure 1: A traditional framework of a multiple classifier system

In the first step of training multiple classifiers, three popular mechanisms are usually used to generate the component classifiers. The first mechanism is based on different training sets. In this mechanism, several training sets are generated from the original training set, and then multiple classifiers are trained on these training sets respectively. One representative approach using this mechanism is Boosting (Shapire, 1999). In this approach, a series of classifiers are generated whose training sets are determined by the performance of the corresponding previous ones. More specifically, the instances in the training set that are wrongly classified by the previous classifier play more important roles in the training set of the next classifier. The second mechanism is based on different feature sets, in which each classifier is trained on one feature set. These different feature sets may come from different physical features or from different representations of one physical feature. A notable application of MCS using this mechanism is in biometric authentication systems, in which many classifiers are developed by utilizing different personal biometric features, such as face, voice, and fingerprint (Snelick et al., 2005). The last mechanism is based on different learning algorithms. One successful method of this mechanism is called stacking, in which multiple classifiers are generated by using several learning algorithms on a single dataset (Ting and Witten, 1999).

As for the second step of classifier combination, Zhou et al. (2002) shows that the combination of some classifiers is better reduces the generalization error of the ensemble than the combination of all the classifiers both theoretically and experimentally. Therefore, we prefer selecting only some classifiers over using all classifiers. The selection methods are mainly based on performance and classification diversity (Aksela, 2003).

Regarding the third step of combining the component predictions, the combining methods (or rules) are used to combine the output from all the component classifiers. The combining rules are usually categorized into two classes, i.e., fixed and trained rules. Fixed rules, such as the majority voting rule, combine the classification results in some fixed mode independent of the application task, while trained rules combine the results in a trained way, such as in the weighted sum rule (Fumera and Roli, 2005), and Dempster-Shafer method (Sugie and Kobayashi, 2002). Kittler et al. (1998) develop a common theoretical framework on classifier combination based on different feature sets, where many fixed combining rules, such as the product

rule, sum rule, min rule, max rule and vote rule can be derived.

3 Classifier Combination on Sentiment Classification

Sentiment classification aims to assign a document with a category from a predefined category set. The predefined category set usually consists of some sentiment classes, e.g. 'thumbs up' or 'thumbs down', which is the key difference from the topic-based text classification. To implement the machine learning algorithms, standard bag-of-features framework are used. Let $\{ f_1, f_2, ..., f_m \}$ be a predefined set of m features that can appear in a document. Several types of feature sets can be considered (Riloff et al., 2006). As a result, several classifiers can be generated through using different feature sets. We would like to combine these classifiers instead of to choose only one best-performance classifier.

3.1 SVM Classifiers using Multiple Feature Sets

In Pang et al. (2002), some types of features are considered including unigrams, bigrams, and adjectives. Among these features, unigrams are reported as the most effective features. Riloff et al. (2006) create a subsumption hierarchy that defines the representational scope of different types of features. Many kinds of features, such as N-grams, POS features, can be generated from the hierarchy. Although the subsumption hierarchy offers a good way to find effective features for classification with improved performance than unigrams, its implementation is much more complex. Moreover, the improvement sometimes is quite small (from 74.8% to 74.9% on the MPQA data) (Riloff et al., 2006). Therefore, we prefer to use unigrams and POS features. Here, POS features mean the words that are selected from the training data according to their POS tags. Thus there are many different POS feature sets, such as adjectives, nouns and verbs. For example, the word 'good' belongs to the adjectives while the word 'goodness' belongs to the nouns.

Documents are represented by the features. Suppose $n_i(d)$ is the number of times f_i occurs in the document d. Then, each document d is usually represented by the document vector $\vec{d} = (n_1(d), n_2(d), ..., n_m(d))$. Pang and Lee (2002) report that better performance is achieved by accounting only for feature presence than feature frequency. Thus we set $n_i(d)$ to 1 if and only feature f_i appears in the document d, otherwise set $n_i(d)$ to 0.

Many classification methods are available to apply to sentiment classification, such as Naïve Bayes, Maximum Entropy, and Support Vector Machines (SVM). Among these methods, SVM method outperforms other methods (Pang et al., 2002). We use only SVM in our experiments. SVM is a machine learning algorithm for a linear binary classifier, which is learned to maximize the margin of confidence of the classification on the training data set (Sugie and Kobayashi, 2002). To solve the non-linear classification problem, different kernel functions are employed to transform the problem into a linear problem by mapping the original space to a multidimensional space (Vapnik, 1995). We use LIBSVM[1] for the training and testing, with almost all parameters set to their default values. One exception is that we replace the polynomial kernel function with linear kernel function. This is because the linear kernel function performs much better than the polymomial kernel function when train the data using the unigrams.

3.2 Classifier Selection using N-best Method

After obtaining multiple SVM classifiers, classifier selection is a crucial step which aims to select some good-performance classifiers. We use one heuristic method, called N-best, to select classifiers because of its cheap computation. N-best method firstly sorts classifiers according to their performance and then to check how many best performing classifiers form the best ensemble (Ruta and Gabrys, 2005). Then, it works to examine single best classifier, a pair of best classifiers, best three classifiers and so on. Finally, the committee of classifiers with the best performances is chosen.

3.3 Combination of the Selected Classifiers using Different Combining Rules

Many combining rules can be applied to combine different classifiers. The theoretical framework for

[1] LIBSVM is an integrated software for support vector classification which is available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/

combining classifiers based on different feature sets developed in Kittler et al. (1998) can be applied for our combination. We briefly introduce the framework as following.

Suppose R individual classifiers $c_k (k = 1,....,R)$ are selected through the classifier-selection step. Each classifier assigns one input sample (represented as x_k) to a label L_k $(L_k = w_1,...,w_m)$. Assume the classifier c_k gives every output a measurement which is represented as a posterior probability vector, $P_k = [p(w_1 \mid x_k),\ ...,\ p(w_m \mid x_k)]^t$, (where $p(w_i \mid x_k)$ denotes the probability that the classifier considers that x was labeled with w_i).

Majority voting rule:

$$assign \quad Z \to w_j$$

$$j = \arg\max_i \sum_{i=1}^{R} \Delta_i$$

Where

$$\Delta_i = \begin{cases} 1 & L_k = w_i \\ 0 & L_k \neq w_i \end{cases}$$

Max rule:

$$assign \quad Z \to w_j$$

$$j = \arg\max_i \{\max_{k=1}^{R} p(w_i \mid x_k)\}$$

Min rule:

$$assign \quad Z \to w_j$$

$$j = \arg\max_i \{\min_{k=1}^{R} p(w_i \mid x_k)\}$$

Product rule:

$$assign \quad Z \to w_j$$

$$j = \arg\max_i p(w_i) \prod_{k=1}^{R} p(w_i \mid x_k)$$

Sum rule:

$$assign \quad Z \to w_j$$

$$j = \arg\max_i \sum_{k=1}^{R} p(w_i \mid x_k)$$

4 Experiments

We use the data set of classified movie reviews presented by Pang and Lee (2002)[2]. This data set contains 1,400 movie reviews: 700 positive and 700 negative. Classification effectiveness is evaluated in terms of the standard precision (P) which is defined as:

$$P = \frac{\alpha}{\gamma}.$$

Where α represents the number of documents correctly classified by system, and γ is the number of all the documents.

4.1 Individual Classifier Results

Six different types of features are used in our experiment. They are unigrams, adjectives[3], adjective+adverbs (words belongs to the adjective feature set or the adverb feature set), nouns, verb+adjectives, and verb+adverbs.

Table 1 presents the results of the six individual classifiers using SVM method. They are obtained by ten-fold cross-validation on the review data set. For each fold, we used 90% positive and 90% negative reviews for the training and 10% positive and 10% negative reviews for the testing.

From Table 1, we can see that the unigrams achieve the best precision of 80.44%. This is not so surprising, considering that the unigram feature set contains much more features than the other POS feature sets. Among the POS feature sets, the adjective+adverbs perform best, slightly better than the adjectives. An interesting phenomenon observed from the result is that the adjective+adverbs achieve the performance of 76.14% using only 1,350 features. This implies that feature selection may be promising in sentiment classification.

Features	Number of features	Precision (%)
Unigrams	11,226	**80.44**
Adjectives	1,064	76.00
Adjective+Adverbs	1,350	76.14
Nouns	2,190	65.35
Verb+Adjectives	2,177	75.78
Verb+Adverb	1,808	73.29

Table 1: The Performances of the Six Individual Classifiers on the Test Data.

[2] Version 1.0, which is available at: http://www.cs.cornell.edu/people/pabo/movie-review-data/

[3] We use Oliver Mason's QTag for the POS tagging which is available at: http://www.english.bham.ac.uk/staff/oliver/software/tagger/index.htm

4.2 Classifier Selection Results

After six different classifiers are obtained, we use N-best classifier selection method to select the best committee of the classifiers for the further combination. Table 2 presents the result using N-best classifier selection method. Note that we use the sum rule to combine N classifiers during the classifier selection process. From the result, we can see that the combination classifier performs best when N equals 3. Therefore, we choose three individual classifiers for the combination which use the feature sets: unigrams, adjective+adverbs and adjectives.

N-best	Precision (%)
N=1	81.23
N=2	83.48
N=3	84.29
N=4	82.46
N=5	79.87
N=6	77.14

Table 2: The Performances of all N-best-classifier Combination Classifiers on the Training Data.

4.3 Classifier Combination Results

Table 3 presents the results of the combination classifier which combine the selected three individual classifiers. They are performed by ten-fold cross-validation on the review data set. We test five combining rules: the sum, product, max, min, and vote rules.

Combining rules	Precision (%)
Sum	**83.00**
Product	82.71
Max	82.36
Min	82.36
Vote	81.43

Table 3: The Results of Classifier Combination Using Different Combining Rules on the Test Set.

Comparing the experimental results in Table 1 and Table 3, we can find that the combination of multiple classifiers is better than that achieved by a single classifier whatever the combining rule is. The best combination classifier, using the sum rule, outperforms the best single classifier, using the unigram feature, with an improvement of 2.54%.

From Table 3, we can find that the sum rule gets the best performance among all the combining rules. This finding is in good agreement with that reported in (Kittler et al., 1998). The sum rule leads a much better result than the vote rule, which implies that utilizing probability information in the outputs benefits the combination.

5 Conclusions and Future work

We combine multiple SVM classifiers on sentiment classification with several feature sets. We also study some combining rules for the combination of these classifiers. Based on our experiment results, we have the following observation.

- ✧ Comparison among different POS feature sets: the adjective+adverb feature set outperforms the other feature sets.
- ✧ Comparison between the best individual classifier and the combination classifier: it is clear that the combination classifiers show the significant improvement over the single classifiers.
- ✧ Comparison among different combining rules: the sum rule is preferable over the other combining rules.

Future work extending in our research includes several aspects. One aspect is to add more effective individual classifiers using some complex features, e.g., N-grams. The second aspect is to apply some trained combining rules, such as Dempster-Shafer method, to enhance the performance.

References

A. Kennedy, and D. Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2), pages 110-125.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques, In: *Proceedings of EMNLP-02, the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, US: Association for Computational Linguistics, pages 79-86.

B. Pang, and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: *Proceedings of ACL-04, 42nd Meeting of the Association for Computational*

Linguistics, Barcelona, ES: Association for Computational Linguistics, pages 271-278.

D. Klein, K. Toutanova, H.T. Ilhan, S.D. Kamvar, and C.D. Manning. 2002. Combining heterogeneous classifiers for word-sense disambiguation, In: *Proceedings of the. SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation*, pages. 74-80.

D. Ruta, and B. Gabrys. 2005. Classifier selection for majority voting. *Information Fusion*, vol. 6: 63-81.

E. Riloff, S. Patwardhan, and J. Wiebe. 2006. Feature Subsumption for opinion analysis. *In: Proceedings of EMNLP-06, the Conference on Empirical Methods in Natural Language Processing*, Sydney, AUS: Association for Computational Linguistics, pages 440-448.

G. Fumera, and F. Roli. 2005. A theoretical and experimental analysis of linear combiners for multiple classifier systems", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27: 942 - 956.

H. Cui, V. Mittal, and M. Datar. 2006. Comparative experiments on sentiment classification for online product reviews. In: *Proceedings of AAAI-06, the 21st National Conference on Artificial Intelligence*, Boston, US: AAAI Press.

J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. 1998. On combining classifiers. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20: 226-239.

K. M. Ting, and I.H. Witten. 1999. Issues in stacked generalization. *Journal of Artificial Intelligence Research*, vol. 10: 271-289.

L. W. Ku, Y.T. Liang, and H.H. Chen. 2006. Opinion extraction, summarization and tracking in news and Blog corpora," In: *Proceedings ofAAAI-CAAW-06, the Spring Symposia on Computational Approaches to Analyzing Weblogs*.

M. Aksela. 2003. Comparison of classifier selection methods for improving committee performance. In: *Proceeding of the 4th International Workshop, Multiple Classifier Systems (MCS)*, Lecture Notes in Computer Science, vol. 2709, pages 84-93.

R. Florian, A. Ittycheriah, H. Jing, and T. Zhang. 2003. Named Entity Recognition through Classifier Combination. In: *Proceedings of CoNLL-2003*, pages 168-171.

R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. Jain. 2005. Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27(3): 450-455.

R. E. Shapire. 1999. A brief introduction to boosting. In: *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*. pages 1-5.

S. M. Kim, and E. Hovy. 2005. Identifying opinion holders for question answering in opinion texts. In: *Proceedings of AAAI-05 Workshop on Question Answering in Restricted Domains*.

V. Vapnik. 1995. *The Nature of Statistical Learning Theory*. NewYork：SprierVerhg.

Y. Sugie, and T. Kobayashi. 2002. Media-integrated biometric person recognition based on the Dempster-Shafer theory. In: *Proceeding of the 16th International Conference on Pattern Recognition (ICPR)*, vol. 4: 381-384.

Z. Switzerland. 1996. Combining classifiers in text categorization. In: *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval*, pages 289-297.

Z. H. Zhou, J. Wu, and W. Tang. 2002. Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, vol.137(1-2): 239-263.