Document-Level Sentiment Classification Based on Behavior-Knowledge Space Method

Zhifei Zhang, Duoqian Miao, Zhihua Wei, and Lei Wang

Department of Computer Science and Technology,
Tongji University, Shanghai 201804, China
Key Laboratory of Embedded System and Service Computing,
Ministry of Education, Tongji University, Shanghai 201804, China
zhifei.zhang@gmail.com

Abstract. There are mainly two kinds of methods for document-level sentiment classification, unsupervised learning and supervised learning. When ensemble learning is introduced, existing methods only combine unsupervised learning algorithms or supervised learning algorithms. To overcome each other's flaws, a novel sentiment classification method based on behavior-knowledge space is proposed, in which two unsupervised and two supervised learning algorithms are utilized. The experiment results not only explain the effectiveness by diversity measure, but also show that the proposed method is significantly superior to each individual classifier. In addition, our method is better than the other two common ensemble methods.

Keywords: sentiment classification, ensemble learning, behavior-knowledge space.

1 Introduction

More and more people express their attitudes and opinions about products, persons or events using the Internet. These user-generated texts are unstructured or semi-structured, which contain subjective information, such as attitudes, opinions, and emotions. Sentiment classification aims to determine the polarity of a given text at document, sentence or word level, i.e., whether the expressed opinion is positive, negative or neutral. Sentiment classification is a research hotspot of web mining, which has been widely used in many domains [1], e.g., commerce product recommendation, unhealthy information filtering, public opinion monitoring, and stock trend prediction.

Depending on the granularity of text, sentiment classification can be divided into three levels: word, sentence and document. Word-level sentiment classification is the basis of the other two tasks, whose methods are dictionary-based or corpus-based [2]. Dictionary-based methods utilize dictionaries, lexicons or their extensions to obtain the polarity of a word, e.g., WordNet [3], HowNet [4], while corpus-based methods fully utilize corpus to obtain the polarity of a word, e.g., adjectives' clustering [5], point-wise mutual information [6]. Sentence-level

S. Zhou, S. Zhang, and G. Karypis (Eds.): ADMA 2012, LNAI 7713, pp. 330–339, 2012.

sentiment classification is to decide the polarity of a sentence and also extracts sentiment-related elements, such as opinion holders and opinion aspects. As the key of sentiment classification, sentence-level sentiment classification can support fine-grained sentiment analysis. One method is to use sentiment dictionaries and domain dictionaries, extract subjective elements and calculate the polarity with weighted sum [7]. Another method is to construct classifiers based on machine learning, e.g., CRFs-based sentiment classification with redundant features [8], subjectivity summarization based on minimum cuts [9]. Document-level sentiment classification is to give the overall polarity of a document, by unsupervised learning [10][11] or supervised learning [12][13].

Ensemble learning is to combine outputs of multiple classifiers and can gain better results. It has been successfully applied in text classification [14], but less in sentiment classification, which can be regarded as the special case of text classification. Xia et al. [15] focus on ensemble of feature sets and classifiers, in which all classifiers belong to supervised learning. Wan [16] combines unsupervised learning methods using bilingual knowledge. In term of the relation among basic classifiers, ensemble learning can be classified into two groups, homogeneous ensemble learning and heterogeneous ensemble learning [17]. The common heterogeneous ensemble strategies are Voting, Bayes' Rule and Behavior-Knowledge Space [18]. Voting is simple but designs equality of basic classifiers. Bayes' Rule is limited to the assumption of independence among basic classifiers. Behavior-Knowledge Space (BKS, for short) solves their defects. In this paper, we focus on ensemble of unsupervised learning and supervised learning algorithms based on Behavior-Knowledge Space method for document-level sentiment classification, and verify the effectiveness of the proposed method on real sentiment corpus. Our contributions are summarized as follows.

- Propose an ensemble of unsupervised learning methods and supervised learning methods for sentiment classification.
- Explore the effectiveness of Behavior-Knowledge Space ensemble method for document-level sentiment classification.

The remainder of this paper is organized as follows. Section 2 introduces ensemble learning briefly and BKS method in detail. Our method is described in Section 3. The experiment results are illustrated in Section 4. Section 5 concludes this paper.

2 Ensemble Learning

2.1 Main Idea

A topic within machine learning where there has been a lot of recent activity is ensemble learning, which is to improve classification accuracy by learning ensembles of classifiers [19].

Given a set of training examples, a classifier is obtained by a learning algorithm. Given a new example, each classifier predicts the corresponding output.

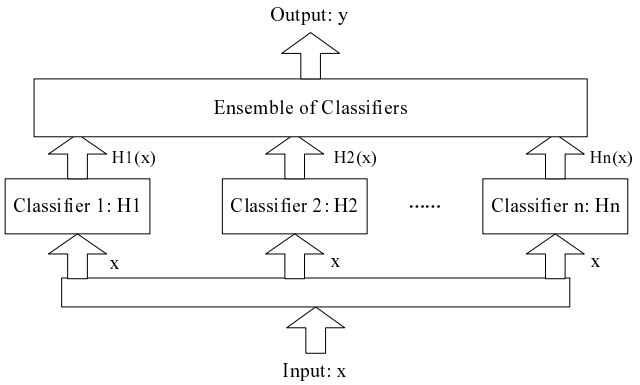

Fig. 1. Main Idea of Ensemble Learning

An ensemble of classifiers is multiple classifiers whose individual outputs are combined in some way to classify a new example. The main idea of ensemble learning is described in Fig.1.

An ensemble is more accurate than each basic classifier only if the component classifiers disagree with one another and the accurate rate of each basic classifier is not below 0.5 [19]. Diversity is the basis of the ensemble learning, which can be measured by Q-static. Given two classifiers H_i and H_j, Table 1 is the contingency table between each pair of classifiers.

Table 1. Contingency Table between Classifiers H_i and H_j

	H_j True	H_j False
H_i True	A	B
H_i False	C	D

A is the probability that both classifiers truly label input data. B is the probability that H_i truly labels but H_j falsely labels input data. C is the probability that H_i falsely labels but H_j truly labels input data. D is the probability that both classifiers falsely label input data. By this definition, $A+B+C+D=1$ is hold. Then the Q-static measure of diversity for these two classifiers is calculated by Eq.1 [20].

$$Q_{ij} = \left| \frac{AD - BC}{AD + BC} \right| \tag{1}$$

We can see that $0 \leq Q_{ij} \leq 1$, when Q_{ij} is approaching 0, the diversity of two classifiers is bigger, otherwise the diversity is smaller.

2.2 Behavior-Knowledge Space Method

Behavior-Knowledge Space method [18] is proposed by Huang and Suen in 1993, which is for combination of multiple classifiers. Denote the number of basic classifiers K, the number of decision classes M, the class set $\Lambda = \{1, 2, ..., M\}$. If a classifier rejects an input, the output class is set to be $M+1$. The decision of each classifier H_i is marked by $e(i)$, $e(i) \in \Lambda \cup \{M+1\}$. A behavior-knowledge space BKS is a K-dimensional space. The intersection of the decisions of classifiers is one unit of the BKS, denoted by $BKS(e(1), e(2), ..., e(K))$. For example, $K = 2$, $M = 3$, the BKS is constructed, as presented in Table 2.

Table 2. An Example of 2-Dimensional BKS

$e(1)$ \ $e(2)$	1	2	3	4
1	(1,1)	(1,2)	(1,3)	(1,4)
2	(2,1)	(2,2)	(2,3)	(2,4)
3	(3,1)	(3,2)	(3,3)	(3,4)
4	(4,1)	(4,2)	(4,3)	(4,4)

Each unit of BKS contains three kinds of data: (1) the total number of input examples with each class $m \in \Lambda$, denoted by $n_{e(1),...,e(K)}(m)$, (2) the total number of input examples $T_{e(1),...,e(K)}$, and (3) the best representative class $R_{e(1),...,e(K)}$. The last two can be calculated using the first one.

$$T_{e(1),...,e(K)} = \sum_{m=1}^{M} n_{e(1),...,e(K)}(m) \tag{2}$$

$$R_{e(1),...,e(K)} = \{l | n_{e(1),...,e(K)}(l) = \max_{1 \le m \le M} n_{e(1),...,e(K)}(m)\} \tag{3}$$

For each unit, the best representative class is unique. Given such a BKS, the belief value of an input example belonging to one class is computed by Eq.4.

$$BEL(m) = \frac{n_{e(1),...,e(K)}(m)}{T_{e(1),...,e(K)}} \tag{4}$$

There are two stages in BKS method: knowledge modeling and operation. The knowledge modeling stage is to construct a BKS using training examples, mainly compute the values of T and R for each unit. The operation stage is to make final decisions for new examples according to the decision classes of basic classifiers and the following decision rule,

$$D(x) = \begin{cases} R_{e(1),\ldots,e(K)} & \text{when } T_{e(1),\ldots,e(K)} > 0 \text{ and } BEL(R_{e(1),\ldots,e(K)}) \geq \alpha \\ M+1 & \text{otherwise} \end{cases} \quad (5)$$

where $\alpha (0 \leq \alpha \leq 1)$ is the threshold which controls the reliable degree of the decision.

3 Our Method

3.1 Problem Statement

Here, document-level sentiment classification can be regarded as two-class text classification, with positive and negative labels. Four basic classifiers are used: the simple weighted sum of sentiment words (called SWS) [10], the weighted sum of sentiment words based on concepts (called WSC) [11], support vector machine (SVM) [12] and k-nearest neighbors (KNN)[13]. The former two are unsupervised learning algorithms, and the latter two are supervised learning algorithms. Thus, we get $K = 4$ and $M = 2$ in BKS method. Our ensemble method is indicated in Fig.2.

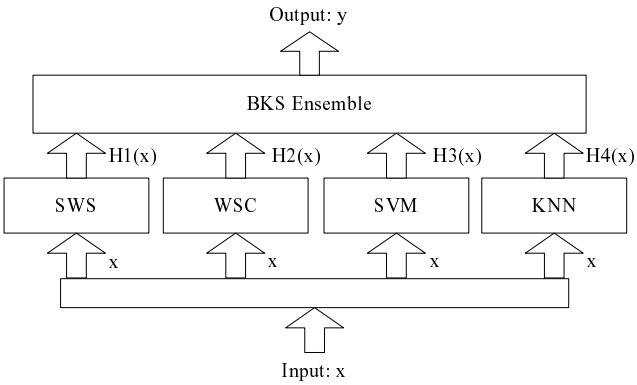

Fig. 2. BKS Ensemble Method for Document-level Sentiment Classification. Here, $H1(x)$, $H2(x)$, $H3(x)$, $H4(x)$ and y are all from {Positive, Negative, Reject}.

3.2 Method Illustration

Our ensemble method is roughly illustrated in Fig.3. In this diagram, the top half is about training and the bottom half is about testing. In training procedure, a BKS ensemble classifier is established based on a training set. In testing procedure, the BKS ensemble classifier makes decisions for a testing set. It is necessary to state that, SWC and WSC directly give the outputs of training examples or testing examples, but SVM and KNN give the outputs of testing examples after learning from training examples.

Algorithm 1 describes the proposed method in detail. The settings of these four basic classifiers are similar to that described in their corresponding references.

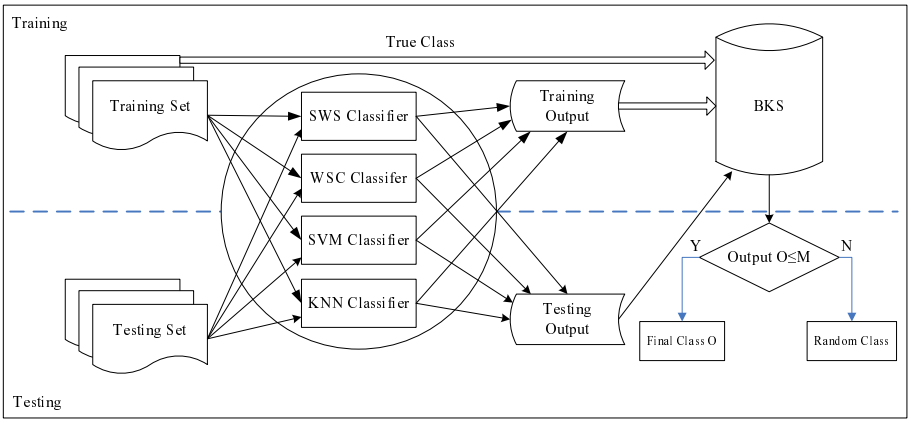

Fig. 3. Diagram of BKS Ensemble for Document-level Sentiment Classification

Algorithm 1. Document-level Sentiment Classification Based on BKS

Input:

Set of training examples, TS;
Set of testing examples, NS;
Reliability threshold for decision, α;

Output:

Decision classes of testing examples, D;

1: Building up two supervised learning classifiers, SVM and KNN, using the training set TS;
2: Obtaining the decision classes by four individual classifiers for each example in TS, denoted by $TID_{|TS|\times 4}$;
3: From $TID_{|TS|\times 4}$ and the true classes in TS, computing T and R for each unit of BKS by Eq.2 and Eq.3;
4: Obtaining the decision classes by four individual classifiers for each example in NS, denoted by $NID_{|NS|\times 4}$;
5: From $NID_{|NS|\times 4}$ and BKS, making decisions for all testing examples by Eq.5, denoted by D';
6: For all the i-th testing example in NS, if $D'(i) < M+1$, then $D(i) = D'(i)$, otherwise $D(i)$ is randomly generated from $\{1, ..., M\}$; $(M = 2)$
7: **return** D.

4 Experiments

4.1 Experiment Settings

The experiments are carried out on ChnSentiCorp-Htl-ba-4000 from ChnSentiCorp[1] plus 5-fold cross validation. The corpus contains 4000 texts about hotel, 2000 positive and 2000 negative. In BKS method, the reliability threshold is set to be 0.55 according to experiences. Certainly, the optimal threshold can be found automatically [18], but more time is needed.

Two metrics, precision and recall, are used to evaluate the performance respectively in positive examples and negative examples, denoted by PP (precision for positive), RP (recall for positive), PN (precision for negative) and RN (recall for negative). F_1 considers both precision and recall, calculated by the following formula. (In fact, here is Macro F_1)

$$F_1 = \frac{(PP + PN) \times (RP + RN)}{PP + PN + RP + RN} \tag{6}$$

Our experiments include two parts. First, we explain the effectiveness of BKS ensemble on the basis of the performances of four basic classifiers and their diversity. Second, the performance of our method is demonstrated via comparison with four basic classifiers and the other two ensemble methods.

4.2 Effectiveness of BKS Ensemble

As known, there are two conditions to guarantee the effectiveness of ensemble learning. The experiment results of four basic classifiers are shown in Table 3.

Table 3. Results of Four Basic Classifiers

Classifier	PP	RP	PN	RN	F_1
SWS	0.850	0.852	0.851	0.850	0.851
WSC	0.681	0.778	0.748	0.631	0.709
SVM	0.898	0.826	0.839	0.905	0.867
KNN	0.857	0.878	0.875	0.854	0.866

Obviously seen in Table 3, the accurate rate of each basic classifier is higher than 0.5, even higher than 0.6. One condition of effective ensemble learning is satisfied.

The diversity is measured by Eq.1. Table 4 displays the diversity measure of four basic classifiers used in our method. Because the diversity measure of two same classifiers is equal to 1, all values of the diagonal direction are useless and marked as "—".

[1] `http://www.searchforum.org.cn/tansongbo/corpus-senti.htm`

Table 4. Diversity Measure of Four Basic Classifiers

	SWS	WSC	SVM	KNN
SWS	—	0.51	0.61	0.45
WSC	0.51	—	0.24	0.26
SVM	0.61	0.24	—	0.90
KNN	0.45	0.26	0.90	—

SVM classifier and KNN classifier are close, because their diversity measure is up to 0.9. WSC classifier is significantly different from the other three classifiers. The difference between SWS classifier and the others is medium. In a word, the diversity of these four classifiers is guaranteed so that our ensemble method is effective.

4.3 Performance of BKS Ensemble

Document-level sentiment classification based on BKS is compared with four basic classifiers, as shown in Fig.4.

In general, BKS ensemble method outperforms four individual classifiers, F_1 is up to 92.5%. Besides, the difference of performances in positive and negative examples is not significant, which is wanted. BKS ensemble method can overcome its basic classifier's flaws and improve the accuracy of classification system.

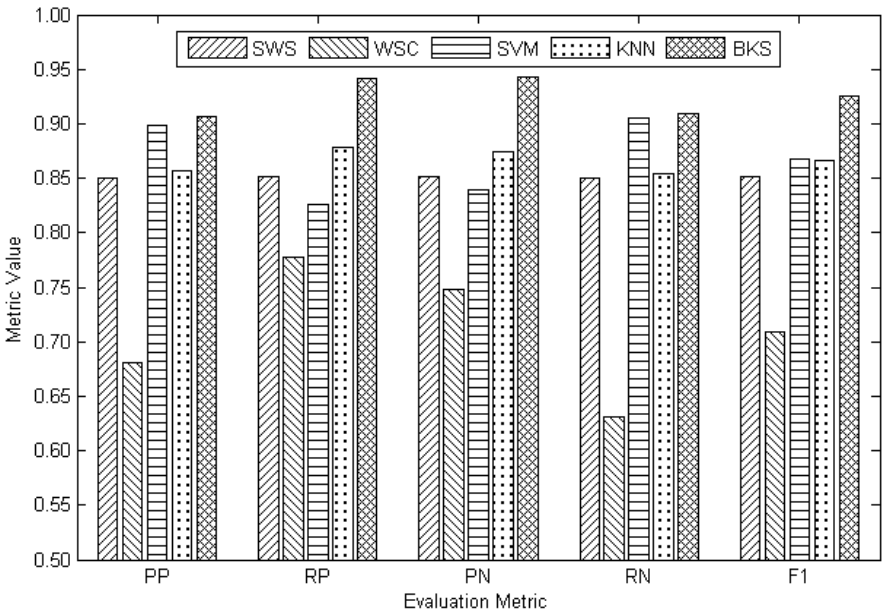

Fig. 4. Comparison Results between BKS Ensemble and Four Basic Classifiers

BKS ensemble method is also compared with the other two ensemble methods, Voting and Bayes' Rule (see Table 5).

Table 5. Comparison Results of Three Ensemble Methods

Ensemble Method	PP	RP	PN	RN	F_1	Random Times
Voting	0.884	0.883	0.883	0.884	0.884	162
Bayes' Rule	0.88	0.909	0.911	0.884	0.896	0
BKS	0.906	0.941	0.943	0.909	0.925	14

BKS is better than the other two ensemble methods. The reason is that BKS does not need the assumption of independence among basic classifiers, while Bayes' Rule needs. BKS does not treat basic classifiers equally, but Voting does. The last column in Table 5 indicates the average times of random decisions. Voting needs more times to decide the classes of examples randomly. Bayes' Rule can make decisions definitely. BKS needs few times to make random decisions, which is acceptable. If the number of basic classifiers increases, the times of random decisions will become more. In this situation, some units of BKS are not assigned more training examples and it is not useful to predict the class of a new example. Thereby, this phenomenon tells us BKS ensemble method is more suitable for document-level sentiment classification.

5 Conclusions

An ensemble of two unsupervised learning methods and two supervised learning methods is proposed for document-level sentiment classification, which is different from previous work. The ensemble classifier based on Behavior-Knowledge Space method is effective and outperforms each basic classifier. Meanwhile, our method is better than the other two ensemble methods, Voting and Bayes' Rule. In ensemble learning, "No free lunch" is always tenable. We will further carry out research on BKS ensemble method with more corpus and more classification algorithms.

Acknowledgments. The authors are grateful to the anonymous referees for their valuable comments and suggestions. This work is partially supported by the National Natural Science Foundation of China (No. 60970061, No. 61075056 and No. 61103067) and the Fundamental Research Funds for the Central Universities.

References

1. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135 (2008)
2. Rao, D., Ravichandran, D.: Semi-supervised Polarity Lexicon Induction. In: EACL 2009, pp. 675–682. ACL, Morristown (2009)

3. Kim, S.M., Hovy, E.: Automatic Detection of Opinion Bearing Words and Sentences. In: IJCNLP 2005, pp. 61–66. ACL, Morristown (2005)
4. Fu, X., Liu, G., Guo, Y., Guo, W.: Multi-aspect Blog Sentiment Analysis Based on LDA Topic Model and Hownet Lexicon. In: Gong, Z., Luo, X., Chen, J., Lei, J., Wang, F.L. (eds.) WISM 2011, Part II. LNCS, vol. 6988, pp. 131–138. Springer, Heidelberg (2011)
5. Wiebe, J.: Learning Subjective Adjectives from Corpora. In: AAAI 2000, pp. 735–740. AAAI Press, Menlo Park (2000)
6. Turney, P., Littman, M.L.: Measuring Praise and Criticism: Inference of Semantic Orientation from Association. ACM Trans. on Information Systems 21(4), 315–346 (2003)
7. Hu, M.Q., Liu, B.: Mining and Summarizing Customer Reviews. In: KDD 2004, pp. 168–177. ACM Press, New York (2004)
8. Zhao, J., Liu, K., Wang, G.: Adding Redundant Features for CRFs-based Sentence Sentiment Classification. In: EMNLP 2008, pp. 117–126. ACL, Morristown (2008)
9. Pang, B., Lee, L.: A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In: ACL 2004, pp. 271–278. ACL, Morristown (2004)
10. Turney, P.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: ACL 2002, pp. 417–424. ACL, Morristown (2002)
11. Chen, Y.F., Miao, D.Q., Li, W., Zhang, Z.F.: Semantic Orientation Computing Based on Concepts. CAAI Trans. on Intelligent Systems 6(6), 489–494 (2011) (in Chinese)
12. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs Up? Sentiment Classification Using Machine Learning Techniques. In: EMNLP 2002, pp. 79–86. ACL, Morristown (2002)
13. Tan, S.B., Zhang, J.: An Empirical Study of Sentiment Analysis for Chinese Documents. Expert Systems with Applications 34, 2622–2629 (2008)
14. Dong, Y.S., Han, K.S.: A Comparison of Several Ensemble Methods for Text Categorization. In: SCC 2004, pp. 419–422. IEEE Computer Society, Washington (2004)
15. Xia, R., Zong, C.Q., Li, S.S.: Ensemble of Feature Sets and Classification Algorithms for Sentiment Classification. Information Sciences 181, 1138–1152 (2011)
16. Wan, X.J.: Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis. In: EMNLP 2008, pp. 553–561. ACL, Morristown (2008)
17. Dietterich, T.G.: Ensemble Methods in Machine Learning. In: Kittler, J., Roli, F. (eds.) MCS 2000. LNCS, vol. 1857, pp. 1–15. Springer, Heidelberg (2000)
18. Huang, Y.S., Suen, C.Y.: The Behavior-Knowledge Space Method for Combination of Multiple Classifiers. In: CVPR 1993, pp. 347–352. IEEE Computer Society, Washington (1993)
19. Dietterich, T.G.: Machine Learning Research: Four Current Directions. AI Magazine 18(4), 97–136 (1997)
20. Kuncheva, L.I., Whitaker, C.J.: Measures of Diversity in Classifier Ensembles. Machine Learning 51, 181–207 (2003)