Knowledge-Based Systems 37 (2013) 451–461

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Sample cutting method for imbalanced text sentiment classification based on BRC

Suge Wang [a,b,*], Deyu Li [a,b], Lidong Zhao [a], Jiahao Zhang [c]

[a] School of Computer and Information Technology, Shanxi University, Taiyuan, 030006 Shanxi, China
[b] Key Laboratory of Computational Intelligence, and Chinese Information Processing of Ministry, of Education, Taiyuan, 030006 Shanxi, China
[c] School of Mathematics Science, Shanxi University, Taiyuan, 030006 Shanxi, China

ARTICLE INFO

Article history:
Received 17 October 2011
Received in revised form 25 July 2012
Accepted 11 September 2012
Available online 3 October 2012

Keywords:
Imbalanced text set
Text sentiment classification
Sample cutting algorithm
Boundary region
Feature weight

ABSTRACT

The vast subjective texts spreading all over the Internet promoted the demand for text sentiment classification technology. A well-known fact that often weakens the performance of classifiers is the distribution imbalance of review texts on the positive–negative classes. In this paper, we pay attention to the sentiment classification problem of imbalanced text sets. With regards to this problem, the algorithm BRC for clarifying the disorder boundary is proposed by cutting the majority class samples in the dense boundary region. The classifier is constructed based on Support Vector Machine. In order to find the better feature weight scheme, combination strategy of sample cutting, and parameters in BRC, three groups of experiments are designed on six text sets about five domains. The experimental results show that the feature weight scheme Presence has the best performance. And the combination strategy BRC + RS can give a tradeoff between the evaluation measures, Precision and Recall on two categories and make the synthetical evaluation measure Accuracy obtain a larger increase. It should be noted that the method of determining the parameters α and β in BRC is empirical.

Although the boundary region cutting algorithm BRC is aimed to text sentiment classification we believe that it is also suitable to any two-category classification problem with imbalanced sample data.

1. Introduction

With the rapid development of web technology, the Internet has become a very important source from which more and more people obtain information. At the same time, it is also rapidly becoming a platform for people to express their opinion, attitude, feeling and emotion. Text sentiment classification (TSC) aims to automatically judge what sentiment orientation, positive ('thumbs up') or negative ('thumbs down'), a subjective text is by mining and analyzing the subjective information in the text, such as standpoint, view, and attitude. However, the subjectivity texts on the Web, such as those on BBS, Blogs or forum websites are often non-structured or semi-structured. More than that, they are often with an imbalanced distribution on the positive and negative sentiment orientations. These two clear characteristics of subjective texts on the Web, non-structuration (or semi-structuration) and imbalance, bring some challenges to TSC. In general, the non-structuration of data can be solved by feature selecting and reexpressing technologies [1–11]. In recent years, by employing some machine learning techniques, e.g. Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machine (SVM), many important researches on TSC for English texts and Chinese texts have been reported [2–11]. One common unreasonable prior in the researches mentioned above is to assume the balance of training data. Nevertheless, as is well known, the performance of most machine learning methods for imbalanced data is not as good as it is for balanced data. This fact has been verified in the field of topic text classification. When standard classification algorithms are applied to imbalanced training data, they tend to be overwhelmed by the majority class and ignore the minor ones. The primary reasons are that the majority class is represented by a large portion of all the samples, while the minority class has only a small percentage. Finally, they result in the boundary fuzzification of the minority class texts, and influence the accuracy of the whole classification. So, many approaches and strategies have been proposed to deal with the imbalance problem in topic text classification, which are applied to the different stages of the training process, e.g., pre-training, in-training, and post-training stages respectively [12–19]. They mainly include sampling-based method [1,12], cost-sensitive learning [13,14], optimization in feature selection [15,16,20], ensemble approach [17], classifier specific improvements including complement Naive Bayes [21] and dynamical parameter tuning in *k*-NNs [22], term weighting approach [18] and so on.

In this paper, we focus our attention on the problem of pre-training stage for imbalanced text sentiment classification. We summarize our research contributions as follows.

* Corresponding author.
E-mail addresses: wsg@sxu.edu.cn (S. Wang), lidy@sxu.edu.cn (D. Li), 353718947@qq.com (L. Zhao), 526486713@qq.com (J. Zhang).

0950-7051/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2012.09.003

Table 1
Characteristics of six data sets.

LAN	DSet	WP	WN	WP/WN	PTr	NTr	PTe	NTe
Chinese	Book1	933	424	2.2/1	849	340	84	84
	Hotel	1000	400	2.5/1	920	320	80	80
English	DVD	1007	500	2/1	607	400	100	100
	Book2	1129	500	2.2/1	1029	400	100	100
	Electronics	756	500	1.5/1	656	400	100	100
	Kitchen	777	500	1.5/1	677	400	100	100

Table 2
Values of k and *DCnum* on six text sets with $\alpha = 1\%$ and $\beta = 10\%$.

Domains	Book1	Hotel	DVD	Book2	Electronics	Kitchen
k	6	7	7	8	6	6
DCnum	1	1	1	1	1	1

At first, some notions used to describe the local density of samples are introduced. Secondly, we propose an algorithm of boundary region cutting (BRC) to balance the two-class texts in the dense boundary region, and then to make the fuzzy class boundary region clear. Thirdly, the random sampling (RS) is adopted to balance the rest training samples after BRC, i.e., BRC + RS. At last, using SVM as the underlying classifier, we compared the proposed method BRC + RS with other three kinds of methods, without cutting samples (WCS), RS and BRC in six text sets in different domains. The experimental results show that BRC and BRC + RS can obviously clarify the chaotic class boundary, and significantly boost the performance of classifiers for imbalanced text sentiment classification.

The remainder of this paper is organized as follows: Section 2 introduces the related works. Section 3 describes the key steps in TSC. Section 4 introduces some basic concepts and the algorithm of boundary region cutting. Section 5 describes the procedures of text sentiment classification. Section 6 presents the experiment setup. Section 7 presents the experiments used to evaluate the effectiveness of the proposed method. Section 8 concludes the paper.

2. Related works

2.1. Text sentiment classification

Text sentiment classification can be achieved on multi-hierarchy linguistic granularity, such as word, collocation, sentence, or text. Relevant researches in this area have been conducted with lexicon-based or corpus-based approaches. Lexicon-based methods involve in deriving a sentiment measure based on sentiment lexica. Turney et al. [25] predicted the sentiment orientation of a review as the average semantic orientation of the phrases in the review that contain adjectives or adverbs, which is known as the semantic orientation method. Kim et al. [27] built three models to assign a sentiment category to a given sentence by combining the individual sentiment of sentiment-bearing words. Kennedy et al. [3] determined the sentiment orientation of a customer review by counting positive and negative terms and taking into account contextual valence shifters, such as negations and intensifiers. Devitt et al. [31] explored a computable metric of positive or negative polarity in financial news text. Taboada et al. [32] presented the semantic orientation CALculator (SO-CAL) which used dictionaries of words annotated with their semantic orientation (polarity and strength), and incorporated intensification and negation. SO-CAL was applied to the polarity classification task, the process of assigning a positive or negative label to a text that captures the text's opinion towards its main subject matter.

Corpus-based approaches consider the sentiment analysis task as a classification task and they used a labeled texts to train a text sentiment classifier. Since the work of Pang et al. [5], various classification models and linguistic features have been proposed to improve sentiment classification performance [6,5,29]. The effectiveness of machine learning techniques for sentiment classification tasks is evaluated in the pioneering research by Pang et al. [5]. The experimental results on the movie review data which are produced via NB, ME, and SVM are substantially better than those results obtained through human generated base lines. Dasgupta et al. [30] proposed a semi-supervised approach to sentiment classification. They firstly used spectral techniques to mine the unambiguous reviews and then classified the ambiguous reviews by a novel combination of active learning, transductive learning, and ensemble learning. Chinese text sentiment analysis have also been studied [22,33,28]. The proposed methods are similar to the lexicon-based or corpus-based methods mentioned above.

2.2. Imbalanced text classification

The most existing methods of sentiment classification mentioned in the previous subsection directly or indirectly assumed

Table 3
The number of positive and negative texts with three feature weight schemes after BRC and BRC + RS.

DSet	Num.	BRC			BRC + RS		
		TFIDF	TF	Pre	TFIDF	TF	Pre
Book1	PT	666	595	671	336	336	334
	NT	336	336	334	336	336	334
Hotel	PT	316	380	424	314	315	312
	NT	314	315	312	314	315	312
DVD	PT	848	801	830	398	396	393
	NT	398	396	393	398	396	393
Book2	PT	971	930	977	397	395	397
	NT	397	395	397	397	395	397
Electronics	PT	592	568	580	397	395	397
	NT	397	395	397	397	395	397
Kitchen	PT	619	597	608	398	394	397
	NT	398	394	397	398	394	397

Table 4
The mean/variance of six evaluation values of BRC and BRC + RS on Bookl and DVD.

		TFIDF Mean/variance	TF Mean/variance	Pre Mean/variance
Book1	BRC	0.7691/0.00373	0.8165/0.00247	0.8200/0.00685
	BRC + RS	0.7489/0.00008	0.7954/0.00004	0.8074/0.00010
DVD	BRC	0.6598/0.01094	0.6872/0.00578	0.7026/0.00834
	BRC + RS	0.6483/0.00056	0.6761/0.00004	0.6851/0.00012

Table 5
The average RN results of under the different α and β on Book1.

α (%)	β									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
1	0.698	0.669	0.669	0.669	0.669	0.669	0.676	0.676	0.676	0.676
2	0.695	0.695	0.695	0.702	0.702	0.702	0.688	0.681	0.676	0.676
3	0.695	0.695	0.702	0.702	0.688	0.676	0.676	0.676	0.676	0.676
4	0.695	0.702	0.702	0.681	0.676	0.676	0.676	0.676	0.676	0.676
5	0.695	0.688	0.674	0.662	0.662	0.662	0.662	0.662	0.662	0.676
6	0.695	0.688	0.662	0.662	0.662	0.662	0.662	0.676	0.676	0.676
7	0.695	0.683	0.671	0.671	0.671	0.671	0.686	0.686	0.686	0.676
8	0.702	0.683	0.671	0.671	0.671	0.686	0.686	0.686	0.676	0.676
9	**0.724**	0.688	0.688	0.688	0.688	0.702	0.702	0.693	0.681	0.676
10	0.724	0.688	0.688	0.688	0.702	0.702	0.693	0.693	0.676	0.676

Table 6
The average RN results of under the different α and β on DVD.

α (%)	β									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
1	0.560	0.566	0.564	0.564	0.564	0.558	0.558	0.552	0.552	0.552
2	0.560	0.556	0.548	0.554	0.554	0.554	0.554	0.558	0.552	0.552
3	0.566	0.554	0.552	0.556	0.556	0.560	0.554	0.554	0.552	0.552
4	0.562	0.562	0.566	0.564	0.572	0.572	0.570	0.566	0.552	0.552
5	0.562	0.562	0.566	0.572	0.572	0.570	0.560	0.552	0.552	0.552
6	0.572	0.562	0.576	0.576	0.576	0.566	0.558	0.558	0.556	0.552
7	0.558	0.574	0.568	0.568	0.564	0.570	0.570	0.568	0.564	0.552
8	0.572	0.580	0.568	0.564	0.570	0.570	0.568	0.564	0.552	0.552
9	0.572	0.568	0.568	0.572	0.570	0.568	0.564	0.552	0.552	0.552
10	**0.584**	0.566	0.562	0.566	0.566	0.566	0.554	0.554	0.554	0.552

Table 7
The average Acc results of under the different α and β on Book1.

α (%)	β									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
1	0.814	0.804	0.804	0.804	0.804	0.804	0.808	0.808	0.808	0.808
2	0.815	0.815	0.815	0.820	0.820	0.820	0.813	0.808	0.808	0.808
3	0.815	0.815	0.820	0.820	0.813	0.808	0.808	0.808	0.808	0.808
4	0.815	**0.820**	0.820	0.808	0.808	0.808	0.808	0.808	0.808	0.808
5	0.813	0.813	0.806	0.801	0.801	0.801	0.801	0.801	0.801	0.808
6	0.813	0.813	0.801	0.801	0.801	0.801	0.801	0.808	0.808	0.808
7	0.813	0.810	0.805	0.805	0.805	0.805	0.812	0.812	0.812	0.808
8	0.810	0.801	0.801	0.801	0.801	0.808	0.808	0.808	0.805	0.808
9	0.810	0.807	0.807	0.807	0.807	0.814	0.814	0.811	0.810	0.808
10	0.810	0.807	0.807	0.807	0.814	0.814	0.811	0.814	0.808	0.808

the balance between negative and positive samples in both of the labeled and unlabeled data. None of them consider a more common case where the class distribution is imbalanced, i.e., the number of positive samples is quite different from that of negative samples in the data. For clarity, the class with more samples is referred as the majority class and the other class with fewer samples is referred as the minority class. In fact, supervised learning on imbalanced classification is rather challenging learning problem, which has been widely studied in several research areas, such as machine learning [34], pattern recognition [35], and data mining [36,51,52], at either data level or algorithmic level. At the data level, the different forms of re-sampling, such as over-sampling and under-sampling, are proposed. Specifically, over-sampling aims to balance the class populations through replicating the minority class samples [38], while under-sampling aims to balance the class populations through eliminating the majority class samples [34,35,39]. Over-sampling method increases the training set size and thus requires longer training time. Furthermore, it tends to lead to overfitting since it repeats minority class examples [38,50]. One typical under-sampling method is random sampling (or undirected sampling) which refers to the process of randomly drawing a subset of training examples from the original set [45]. Many studies have shown that random sampling ignores potentially useful data [45]. At the algorithmic level, specific learning algorithms, such as cost-sensitive learning,

Table 8
The average Acc results of under the different α and β on DVD.

α (%)	β									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
1	0.699	0.705	0.705	0.705	0.705	0.705	0.705	0.700	0.700	0.700
2	0.694	0.694	0.693	0.695	0.695	0.698	0.698	0.701	0.700	0.700
3	0.700	0.696	0.697	0.697	0.697	0.700	0.699	0.699	0.700	0.700
4	0.696	0.701	0.701	**0.709**	0.707	0.707	0.708	0.709	0.700	0.700
5	0.696	0.701	0.701	0.707	0.707	0.708	0.703	0.700	0.700	0.700
6	0.692	0.698	0.709	0.708	0.706	0.703	0.700	0.700	0.700	0.700
7	0.690	0.701	0.702	0.700	0.703	0.708	0.708	0.708	0.708	0.700
8	0.698	0.707	0.702	0.703	0.708	0.708	0.708	0.708	0.700	0.700
9	0.698	0.701	0.700	0.709	0.708	0.708	0.708	0.700	0.700	0.700
10	0.691	0.697	0.694	0.703	0.703	0.704	0.696	0.696	0.696	0.700

Table 9
α, β, k, *DCnum* and the remaining PT and NT by BRC or BRC + RS on six text sets with the best average Acc.

DSet	Methods	α (%)	β (%)	k	DCnum	PT	NT
Book1	BRC	2	40	12	5	697	336
	BRC + RS	3	100	18	18	330	330
Hotel	BRC	1	30	7	3	615	312
	BRC + RS	8	70	50	35	316	316
DVD	BRC	4	40	40	12	833	395
	BRC + RS	5	80	33	27	400	400
Book2	BRC	2	10	15	2	947	397
	BRC + RS	5	10	36	4	396	396
Electronics	BRC	4	30	22	7	606	395
	BRC + RS	9	50	48	24	394	394
Kitchen	BRC	4	60	22	14	667	400
	BRC + RS	1	40	6	3	396	396

Table 10
Other evaluation measures corresponding to the best average Acc on six text sets with four cutting schemes.

DSet	Cutting scheme	RN	PN	FN	RP	PP	FP	ACC
Book1	WCS	0.676	0.919	0.778	0.940	0.745	0.831	0.808
	RS	0.771	0.825	0.797	0.836	0.787	0.810	0.804
	BRC	0.702	0.920	0.795	0.938	0.761	0.840	0.820
	BRC + RS	0.788	0.884	0.833	0.895	0.809	0.849	0.842
Hotel	WCS	0.630	0.887	0.737	0.920	0.713	0.804	0.775
	RS	0.768	0.864	0.812	0.878	0.791	0.832	0.823
	BRC	0.703	0.857	0.771	0.883	0.749	0.810	0.792
	BRC + RS	0.788	0.868	0.825	0.880	0.806	0.841	0.834
DVD	WCS	0.552	0.785	0.646	0.848	0.656	0.739	0.700
	RS	0.682	0.699	0.690	0.706	0.690	0.698	0.694
	BRC	0.576	0.783	0.663	0.842	0.667	0.744	0.709
	BRC + RS	0.690	0.719	0.704	0.730	0.703	0.716	0.710
Book2	WCS	0.468	0.730	0.569	0.828	0.610	0.702	0.648
	RS	0.610	0.662	0.635	0.688	0.638	0.662	0.649
	BRC	0.520	0.730	0.607	0.810	0.629	0.708	0.665
	BRC + RS	0.734	0.673	0.702	0.644	0.708	0.674	0.689
Electronics	WCS	0.626	0.747	0.681	0.788	0.679	0.729	0.707
	RS	0.696	0.699	0.696	0.696	0.696	0.695	0.696
	BRC	0.660	0.750	0.701	0.778	0.696	0.734	0.719
	BRC + RS	0.756	0.718	0.735	0.700	0.743	0.719	0.728
Kitchen	WCS	0.622	0.774	0.689	0.818	0.684	0.745	0.720
	RS	0.718	0.725	0.721	0.726	0.720	0.723	0.722
	BRC	0.638	0.777	0.700	0.816	0.693	0.749	0.727
	BRC + RS	0.712	0.762	0.736	0.776	0.729	0.752	0.744

one-class learning, and ensemble learning [17,40,41,45] are proposed. For more details, please refer to the comprehensive survey by He and Garcia [42]. However, all these effectiveness previous studies strategies have been mainly conducted on non-text domain (e.g., using UCI data sets [34–38,40–43]). Sun et al. [19] provided a comparative study of existing strategies proposed for imbalanced text classification using SVM through extensive experiments on multiple benchmark data sets. The importance of imbalanced text classification in real-world applications and the uniqueness of text classification tasks were discussed (e.g., high dimensionality, sparse feature spaces, and linearly separability in most tasks [16,18,46,44,52]). Combarro et al. [44] proposed a family of linear measures for feature selection and evaluated their effectiveness with SVM classifiers on two text data sets (i.e., Reuters-21578 and Ohsumed) and improvement on F1 was observed. Liu et al. [18] studied a probability scheme based on feature weighting for

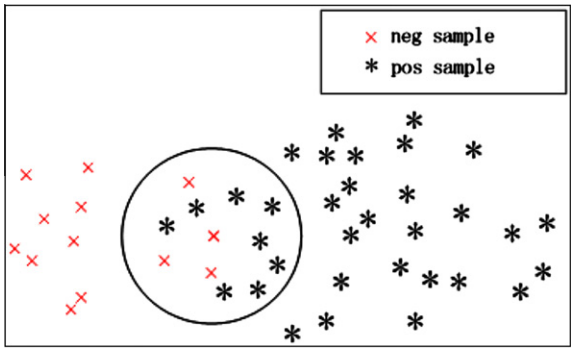

Fig. 1. A high density neighbor in the boundary without cutting.

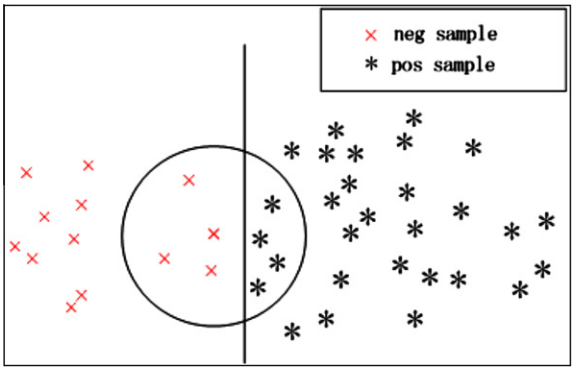

Fig. 2. The high density neighbor in the boundary after cutting.

imbalanced classification. A feature is assigned more weight if it appears more frequently in the positive training examples than in the negative ones. Zheng et al. [16] proposed a feature selection framework to select positive features that are most indicative of membership of target category and negative features that are most indicative of membership of non-target category separately. The positive and negative features are then combined and used to represent training texts. All of the research works mentioned above focus on imbalanced text topic classification. Only a few of literatures [47–49] considered the problem of imbalanced text sentiment orientation. These research works focus on learning algorithms and random under-sampling scheme.

3. Key steps of TSC

In this section, we briefly illustrate the techniques adopted in the key steps of TSC such as text representation, feature weight schemes, feature selection and classifier construction in this paper.

3.1. Text representation

Text representation is the precondition of text analysis. After word segmentation a Chinese text can be representation by the bag of words model which is one of most commonly employed models in modern text analysis and is widely used in information retrieval and text mining [24]. Words are counted in the bag, which differs from the mathematical definition of a set. Each word corresponds to a dimension in the resulting text representation space and every text then becomes a vector for which each dimension takes a non-negative value.

Let D be a set of review texts, T the set of distinct terms occurring in D. In particular, a term is refer to a word in this paper. Suppose the number of terms in T is m. A text is then represented as a m-dimensional vector

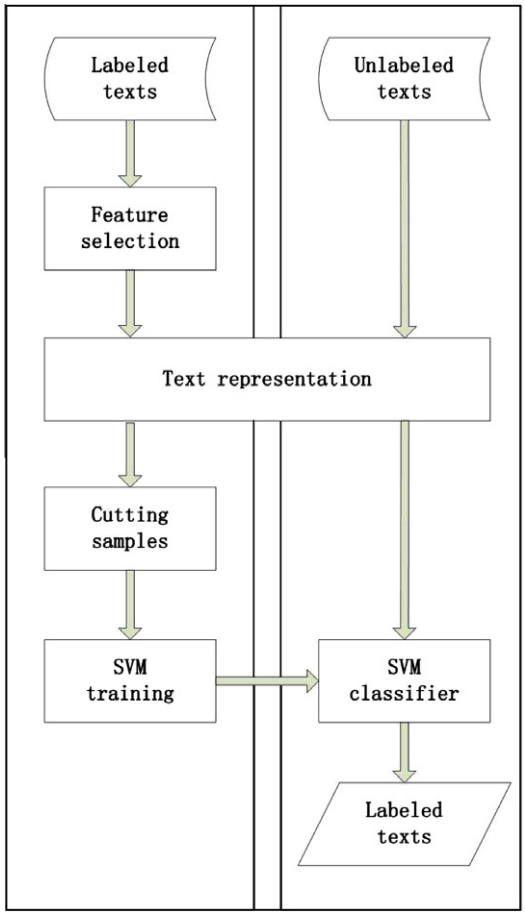

Fig. 3. Flow chart of the text sentiment classification procedures.

$$V(d) = ((t_1, w(d,t_1)), \ldots, (t_i, w(d,t_i)), \ldots, (t_m, w(d,t_m)))$$

where $d \in D$, $t_i \in T$ and $w(d,t_i)$ denotes the weight of t_i in text d.

3.2. Feature weight schemes

For TSC, the weight $w(d,t_i)$ is used to measure the significance of t_i to classifying contribution. In topic text classification, the term weighting TFIDF (term frequency TF, inverse text frequency IDF) has gained a great success. In sentiment classification, three kinds of weight schemes Presence [5], TF [28] and TFIDF [19] were attempted. In this paper, these three kinds of weight schemes will be adopted in our comparative experiments.

3.3. Feature selection

In general, the original number of terms in T is very large. If all of these terms are used to describe texts it will result in the high dimensionality of text representation space and the sparsity of text data which will worsen the performance of a classifier for TSC. Feature selection plays a very important role in reducing of high dimensionality and the sparsity of data. Those terms with larger classifying contribution should be selected as text representation features. There are many effective feature selection methods for measuring the classifying contribution of a term in machine learning, such as, Information Gain (IG), Fisher Discriminant Ratio (FDR), Document Frequency (DF), CHI and Mutual Information (MI) [7]. However, for text sentiment classification, only the words of the part of speech (POS) such as noun, adjective, verb, adverb and idioms are usually with the positive or negative sentiment orientation [7,8,11,25,26]. The Chinese words and domain idioms similar to the situation in

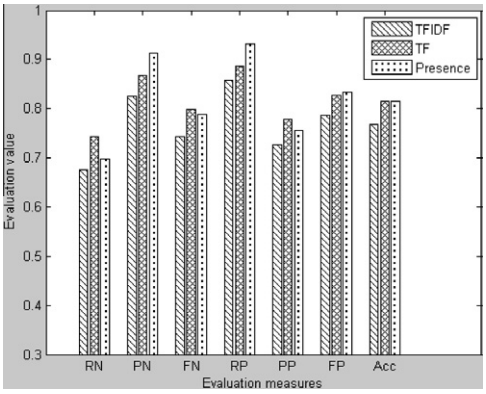

Fig. 4. The results on Book1 text set using BRC with three weights schemes.

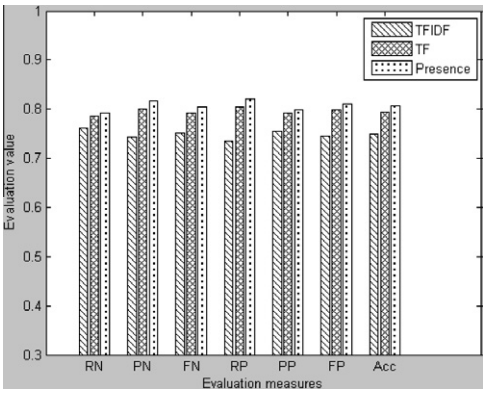

Fig. 5. The results on Book1 text set using BRC + RS with three weights schemes.

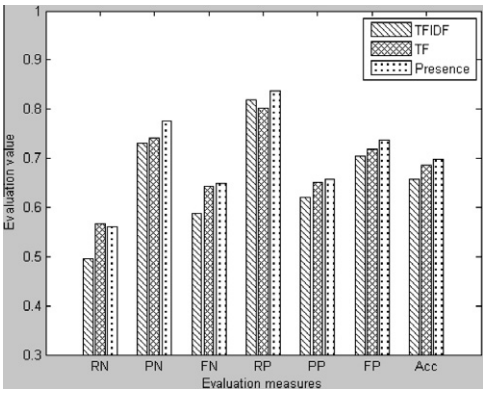

Fig. 6. The results on DVD text set using BRC with three weights schemes.

English text sentiment classification are also often selected to represent texts in Chinese text sentiment classification. Some Chinese idioms have obvious sentiment orientations. For example, some Chinese idioms in the domain of computer book reviews are listed below. "浅显易懂 (clear and easy to understand), 可视化操作 (visualized operations), 物超所值 (excellent quality and reasonable price), 不容错过 (not to be missed), 易懂 (understandability), 无可挑剔 (impeccable), 珍藏 (enshrinement), 晦涩难懂 (obscure), 必经之路 (the only way which must be passed), 扛鼎之作 (World-renowned works), 晦涩 (obscure), 不知所云 (not know what is said), 通俗易懂 (easily understood), 漏洞百出 (full of loopholes), 必读之书 (books which must be read)".

3.4. Classifier based on support vector machine

As a relatively new machine learning method, SVM developed by Vapnik (1995) [53] embodies the VC-dimension theory and the

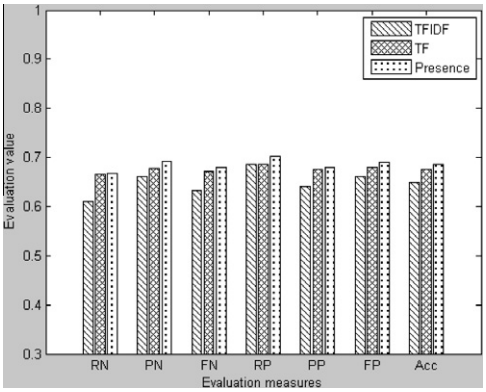

Fig. 7. The results on DVD text set using BRC + RS with three weights schemes.

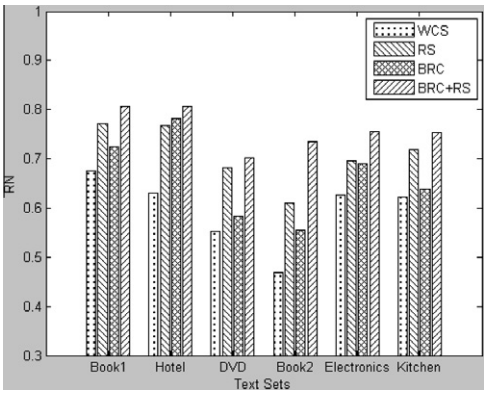

Fig. 8. The best negative average recall RN on six text sets with four cutting schemes.

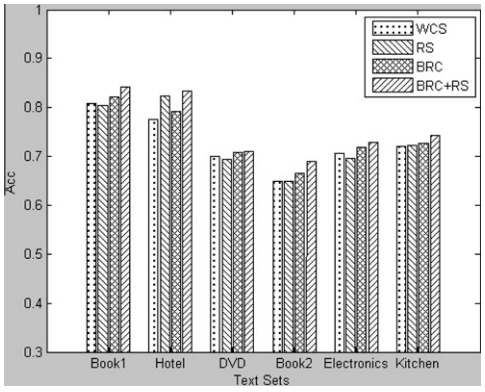

Fig. 9. The best whole average accuracy Acc on six text sets with four cutting schemes.

structural risk minimization principle. It seeks a decision hyperplane to separate the training data points into two classes and makes decisions based on the support vectors that are selected as the only effective elements in the training set. Up to now, it is verified that SVM possesses the best performance for text sentiment classification [5,7,8,11,23] in balanced training set. This paper focuses on the problem of eliminating the imbalance of a training data set. The treated training data are balanced. Therefore, we adopt SVM to construct the classifier in this paper for imbalanced TSC.

4. Sample cutting for imbalanced text set

In order to balance the training data, one straightforward way is re-sampling method, which balances the positive and negative

classes either by over-sampling the minority class samples or by under-sampling the majority class samples. However, many existing studies shown that under-sampling is an efficient strategy to deal with class imbalance [34,38,39,43,19,47,50]. This method uses a subset of the majority class to train the classifier. Since many majority class examples are ignored, the training set becomes more balanced and training process becomes faster. To improve the classification accuracy rate and recall value of the negative texts (usually the minority class) in an imbalanced text sets, in this paper, we attempt a method of boundary region cutting (BRC) to clarify the disorder class boundary region and to balance training set for TSC.

4.1. Some basic concepts

In order to show the proposed cutting method clearly, we introduce several basic concepts.

Definition 1. Let D be a text set, $d \in D$, $\varepsilon > 0$ a constant number. We call

$$N_\varepsilon(d) = \{d_y \in D | Dist(d, d_y) \leqslant \varepsilon\}$$

as the ε-neighbor of d in D, where $Dist(\cdot,\cdot)$ is a distance measure in the text representation space.

Definition 2. Let $D = C_1 \cup C_2$ be a review text set, $C_1 \cap C_2 = \emptyset$ and $d \in D$. $C_i (i = 1, 2)$ denotes the ith category text set. A text d is called a within-class text if $N_\varepsilon(d) \subseteq C_1$ or $N_\varepsilon(d) \subseteq C_2$, and d is called a boundary text otherwise.

It should be noted that whether a text d is a within-class text is often related to the neighbor radius ε.

Definition 3. Let d be a text in D. By $N^k(d)$, we denotes the set of the nearest k texts to d in D. And we call $R^k(d) = \frac{1}{k}\sum_{d_x \in N^k(d)} Dist(d_x, d)$ the average radius of k-neighborhood of d. For a given review text set D, $\varepsilon^k(D) = \frac{1}{|D|}\sum_{d \in D} R^k(d)$ is called the average k-neighborhood radius of D, where $|\cdot|$ denotes the cardinality of a set.

Definition 4. Let $d \in D$. We call d a high density region text if the text number in $N_{\varepsilon^k(D)}(d)$ is larger than a preselected threshold.

Definition 5. Let d_x, $d \in D$ be two high density region texts. If $d_x \in N_{\varepsilon^k(D)}(d)$ (or equivalently $d \in N_{\varepsilon^k(D)}(d_x)$), d_x is then called to be high-density reachable from d. By $HDR^k(d)$, we denote all high-density reachable texts from d.

4.2. Boundary region cutting algorithm BRC

The main idea of the algorithm BRC is as follows: For every high-density neighbor in the boundary region, we cut some majority class texts from it to clarify the disorder boundary region and to balance two-class texts when the majority class texts are much more than the minority class texts. If all other samples in the high-density neighbor of a minority text are majority class texts we consider it as a noise or an outlier, and put it away. The main idea of BRC can be illustrated by Figs. 1 and 2.

There are two parameters α and β in Algorithm BRC that need to be preassigned. The parameter α is used to tune the text number k in a neighborhood, which shows how large the ratio of k to the average text number of two class we expect. The parameter β is used to control the least number $DCnum$ of texts in a high density neighborhood which shows how large the ratio of $DCnum$ to k we expect. For the given parameters α and β, k and $DCnum$ are determined by the following formulas. In this paper, $\alpha \in [1\%, 10\%]$ and $\beta \in [10\%, 100\%]$.

$$k = \frac{\text{the number of training texts}}{\text{the number of categories}} \times \alpha \quad (1)$$

$$DCnum = k \times \beta \quad (2)$$

Let $d_x \in D_N$. For the given k and $DCnum$, by $k\_n(d_x)$ we denote the text number in the neighborhood $N_{\varepsilon^k(D_N)}(d_x)$, i.e., $k\_n(d_x) = |N_{\varepsilon^k(D_N)}(d_x)|$. By $DCnum\_n(d_x)$ we denote the text number in the neighborhood $N_{\varepsilon^{DCnum}(D_N)}(d_x)$, i.e., $DCnum\_n(d_x) = |N_{\varepsilon^{DCnum}(D_N)}(d_x)|$. We call $N_{\varepsilon^k(D_N)}(d_x)$ a high density neighborhood, if $k\_n(d_x) > DCnum\_n(d_x)$. The task of the boundary region cutting algorithm BRC is to cut those high-density reachable positive texts of negative texts. Algorithm BRC is described as follows.

Algorithm 1. Boundary Region Cutting Algorithm (BRCA)

Input: training text set $D = D_P \cup D_N$; D_P-positive training text set; D_N-negative training text set; $Cnum$-the number of categories; α; β.
Output: D'-new training set.

$D' = \emptyset$; $k = \frac{|D|}{Cnum} \times \alpha$; $DCnum = k \times \beta$;
for every $d_x \in D_N$ **do**
  $R^k(d_x) = \frac{1}{k}\sum_{d_y \in N^k(d_x)} Dist(d_y, d_x)$;
  $R^{DCnum}(d_x) = \frac{1}{DCnum}\sum_{d_y \in N^{DCnum}(d_x)} Dist(d_y, d_x)$;
end for
$\varepsilon^k(D_N) = \frac{1}{|D_N|}\sum_{d_x \in D_N} R^k(d_x)$;
$\varepsilon^{DCnum}(D_N) = \frac{1}{|D_N|}\sum_{d_x \in D_N} R^{DCnum}(d_x)$;
for every $d_x \in D_N$ **do**
  $R(d_x) = \{d_y \in D_N | d_y \in N_{\varepsilon^k(D_N)}(d_x)\}$;
  $R^{\sim}(d_x) = \{d_y \in D_P | d_y \in N_{\varepsilon^k(D_N)}(d_x)\}$;
  $k\_n(d_x) = |N_{\varepsilon^k(D_N)}(d_x)|$; $DCnum\_n(d_x) = |N_{\varepsilon^{DCnum}(D_N)}(d_x)|$;
  if $R^{\sim}(d_x) = \emptyset$ **then**
    $ratio(d_x) = 0$;
  else
    $ratio(d_x) = \frac{|R(d_x)|}{|R^{\sim}(d_x)|}$;
  end if
end for
for every item d_x of D_N in non-increasing order according the value of $ratio(d_x)$ **do**
  if $R^{\sim}(d_x) \neq \emptyset$ and $k\_n(d_x) > DCnum\_n(d_x)$ **then**
    if $|R(d_x)| = 1$ and $|R^{\sim}(d_x)| > |R(d_x)|$ **then**
      $D_N = D_N - \{d_x\}$;
      if $|D_N| = |D_P|$ **then**
        goto Step 40;
      end if
    else
      while $|R^{\sim}(d_x)| > |R(d_x)|$ and $k\_n(d_x) > DCnum\_n(d_x)$ **do**
        $d_m = \arg\max_{d_z \in R^{\sim}(d_x) \bigcap HDR^k(d_x)} |N_{\varepsilon^k(D_N)}(d_z)|$;
        $D_P = D_P - \{d_m\}$;
        if $|D_N| = |D_P|$ **then**
          goto Step 40;
        end if
        for every $d_y \in D_N$ and $d_m \in R^{\sim}(d_y)$ **do**
          $R^{\sim}(d_y) = R^{\sim}(d_y) - \{d_m\}$;
          $k\_n(d_y) = k\_n(d_y) - 1$;
        end for
      end while
    end if
  end if
end for
$D' = D_N \cup D_P$;

5. Procedures of text sentiment classification

The procedures of text sentiment classification are briefly described as follows:

Step 1. Data pre-processing
Deleting non-textual information: By scanning text set, a lot of messy codes and tags are deleted.
Chinese words segmentation and POS tagging: Process the texts using the software for Chinese words segmentation and POS tagging.[1]

Step 2. Data set partitioning
In order to test the performance of Algorithm BRC, we perform fivefold cross-validation on six data sets in five domains. For each data set we randomly split it into five text subsets. During each time of the fivefold cross-validation process, a single text subset is retained as the testing text set, and the other four text subsets are merged as the training text set.

Step 3. Feature selection and text representation
As discussed in Section 3.3, the Chinese words of the part of speech (POS) such as noun, adjective and domain idioms are selected as features to represent texts. For comparison, three kinds of features weight measure schemes, i.e., TFIDF, TF and Presence are used to construct text representation vectors respectively.

Step 4. Sample cutting
Some positive training texts are cut in two ways, i.e., BRC and BRC + RS before a classifier is trained. Where, BRC refers to cutting texts only using Algorithm BRC. BRC can cut the high density texts, and then balance the text numbers of two classes only in the boundary region. The complimentary random sampling method (RS) to BRC is for further cutting some positive texts to balance the entire training set after BRC. In this paper, Euclidean distance is adopted to calculate the distance between two texts.

Step 5. Training a classifier
In this paper, we train a classifier by using Support Vector Machine (SVM)[2] on the cut training text set obtained by BRC or BRC + RS.

Step 6. Classing unlabeled texts
For an unlabeled text, represent it by using the methods in Step3, and present it to the classifier. The classifier then returns a labeled result to it.

The flow chart about the text sentiment classification procedures is shown in Fig. 3.

6. Experiment setup

Our experiments are conducted on six text sets, i.e., book review and hotel review in Chinese, and book, DVD, electronics, kitchen subjective texts in English. In order to exam the effectiveness of the proposed cutting method, we conduct three group experiments. In the first group, we compare the impact of varying feature weight schemes of text representation. In the second group, we study the impact of varying parameters in Algorithm BRC. In the third group, we compare four pre-processing schemes which include WCS, RS, BRC and BRC + RS on six data sets.

6.1. Cutting schemes

In our experiments, the three kinds of data pre-processing schemes RS, BRC and BRC + RS are respectively used to cut training samples. On the basis of the cut training text sets, we adopt SVM with linear kernel as the underlying classifier.

WCSs (Without Cutting Samples) refers to training the SVM classifier on the original training text set.

RS (Random Sampling) indicates an under-sampling method which randomly selects some positive training texts and cut them until the training data set is balanced.

BRC and BRC + RS have been illustrated in Section 5.

6.2. Datasets collection

We collected a large number of Chinese subjective texts about book review and hotel review and English subjective texts about book, DVD, electronics, kitchen from related websites[3] and some research institutes.[4]

Chinese book review text set consists of 1357 Chinese book review texts about 410 computer-related books, in which there are 933 positive texts and 424 negative texts. For this text set, two subjects participated in the annotation procedure. The polarity tags of the reviews were first annotated by one subject and then checked by the other subject. The conflicts were resolved by discussion. The book reviews are generally more brief and their positive–negative orientation have various factors, such as content, price, quality, and after-sales service. Here are two examples of Chinese reviews:

Doc1.
还没来得及仔细看，不过翻了翻觉得确实是一本经典的好书

(There's no time to read it carefully, but with glancing over, feel that it is indeed a classic book.)

Doc2.
书里面有很多错误的地方或者说是一本考研用书不到位的地方，很都是编排错误的，幸亏先看书再看这本参考书，要不都得误导自己了。

(There are many mistakes in this book, in other words, many knowledge points did not come up to reader's expectations, many of them are editing-phototypesetting wrongs. Fortunately, I had read the textbook before reading it, if not, this reference book have to mislead myself.)

Chinese hotel review text set is from the hotel sentiment corpus of Songbo Tan research group of computation technology institute of Chinese academy of sciences. We selected 1000 positive texts with document names from pos0.txt to pos999.txt and 400 negative texts with document names from neg0.txt to neg399.txt.

English data sets are from the corpora constructed by Li et al. [47].

Some characteristics of six data sets are shown in Table 1, where

LAN	Language;
DSet	Text set;
WP	Number of whole positive texts;
WN	Number of whole negative texts;
PTr	Number of positive training texts;
NTr	Number of negative training texts;
PTe	Number of positive testing texts;
NTe	Number of negative testing texts.

[1] http://ictclas.org/.
[2] http://www.csie.ntu.edu.tw/cjlin/libsvm
[3] http://www.dangdang.com.
[4] http://www.am-azon.cn.

6.3. Evaluation measures

In this paper, four classical evaluation measures, Precision, Recall, F1-measure[7–11], Accuracy [45,47] are adopted to test the effectiveness of data pre-processing schemes. By RN (RP), PN (PP), FN (FP) and Acc we denote Recall, Precision and F1-measure of positive (negative) review texts and Accuracy respectively. These evaluation measures can be calculated by the following formulas respectively.

$$RN(\text{Recall}) = \frac{d}{b+d} \tag{3}$$

$$PN(\text{Precision}) = \frac{d}{c+d} \tag{4}$$

$$FN(\text{F1-measure}) = \frac{2 \times RN \times PN}{RN + PN} \tag{5}$$

$$RP(\text{Recall}) = \frac{a}{a+c} \tag{6}$$

$$PP(\text{Precision}) = \frac{a}{a+b} \tag{7}$$

$$FP(\text{F1-measure}) = \frac{2 \times RP \times PP}{RP + PP} \tag{8}$$

$$Acc(\text{Accuracy}) = \frac{a+d}{a+b+c+d} \tag{9}$$

where a (true positive) denotes the number of test texts that the true class of them is positive, and were classified into the positive class; b (false positives) denotes the number of testing texts that the true class of them is negative, but were classified into the positive class; c (false negatives) denotes the number of testing texts that the true class of them is positive, but were classified into the negative class; d (true negatives) denotes the number of testing texts that the true class of them is negative, and were classified into the negative class.

We tested an experimental configuration five times with different random training set and testing set that split each of six text sets, i.e., fivefold cross-validation. Then, the classifiers trained by four ways, i.e., WCS, RS, BRC, BRC + RS are applied to each of the corresponding testing text sets, so that we could compare the average of these results.

7. Experimental results and analysis

According to the procedures of text sentiment orientation classification and the experimental setup described in Sections 5 and 6 respectively, we conduct the following three groups of experiments on six text sets. It should be noted that the testing text sets are balanced, and all of the experimental results are in the case of fivefold cross-validation.

7.1. Comparison of feature weight schemes

In general, text representation method including feature selection and feature weight scheme would influence classification results. In this group of experiments, we compare the results of three kinds of feature weight schemes TFIDF, TF and Presence for sentiment orientation classification on six text sets.

The values of k and $DCnum$ obtained by using Formulas (1) and (2) under $\alpha = 1\%$ and $\beta = 10\%$ are given in Table 2.

The number of the remaining positive and negative training texts in six text sets with three feature weight schemes after using BRC and BRC + RS are shown in Table 3. Where, by Dset, Pre, PT and NT we denote text set, Presence, positive text, negative text respectively.

Owing to the reasons of analogous experimental conclusion and lack of space, we only show the experimental results in Figs. 4–7 on Book1 text set in Chinese and DVD text set in English here.

From Figs. 4–7 we can see that the weight scheme Presence has the best classification results in almost all experiments on two text sets under two cutting methods BRC and BRC + RS except the RN, FN and PP on Book1 text set under the cutting method BRC.

We guess that the reason of the better performance of Presence than TFIDF and TF is the shorter text length in our text sets and the very low frequency of each feature in a text. Therefore, the weight scheme presence is adopted in the other experiments shown in Sections 7.2-7.3.

In order to compare the whole stability in six evaluation values of BRC and BRC + RS, we count the means and variances of BRC and BRC + RS in RN, PN, FN, RP, PP, FP on Book1 and DVD. The result is shown in Table 4.

From Table 4, we can see that the corresponding means of BRC and BRC + RS are of flat, however, the variances of BRC + RS are two orders of magnitude smaller than the corresponding variances of BRC on two text sets Book1 and DVD. This indicates that adding RS to BRC can indeed achieve the trade-off effect of classification performance between the negative and positive samples. In other words, RS can promote the proportionality of BRC in all six evaluation measures.

7.2. Parameter determining

As mentioned in Section 4.2, two parameters α and β in the algorithm BRC need to be preassigned. To study their impact to classification performance, in this group of experiments, we will vary α from 1% to 10% and β from 10% to 100% with the feature weight scheme Presence. The experimental results about RN and Acc on six text sets are conducted. Owing to the reasons of analogous experimental conclusion and lack of space, we only show the experimental results in Tables 5–8 on Book1 text set in Chinese and DVD text set in English here.

In the view of negative average recall RN, by only using BRC, their best average values in two text sets Book1, DVD are 0.724, 0.584 with the corresponding (α,β) values (9%,10%), (10%,10%) from Tables 5 and 6. In Hotel, Book2, Electronics and Kitchen, their average values are 0.783, 0.554, 0.69, 0.686 with the corresponding (α,β) values (1%,10%), (10%,10%), (7%,10%), (5%,10%) from the experiment results.

In the view of whole average accuracy Acc, by only using BRC, their best average values in two text sets Book1, DVD are 0.820, 0.709 with the corresponding (α,β) values (4%,20%), (4%,40%) from Tables 7 and 8. In Hotel, Book2, Electronics and Kitchen, their average values are 0.792, 0.665, 0.719, 0.727 with the corresponding (α,β) values (1%,30%), (2%,10%), (4%,30%), (4%,60%) from the experiment results.

The analysis above indicates that, in practical applications of BRC, we should select a smaller β if users hope a higher negative recall RN; we should select smaller α and β if users hope a higher whole accuracy Acc.

The values of α, β, k, $DCnum$ and the remaining PT and NT by BRC or BRC + RS on six text sets with the best Acc are given in Table 9.

Table 9 gives us the following enlightenment. In practical applications of BRC + RS, because RS cuts some positive samples again after BRC, the control parameter of a neighborhood α should be selected a bigger value, and the density control parameter of a neighborhood β should be selected a bigger value than that only of BRC if users hope a higher whole accuracy Acc.

7.3. Comparison of different cutting strategies

In this group of experiments, we compare four pre-processing schemes WCS, RS, BRC and BRC + RS on six text sets.

The best values of negative average recall RN on 6 text sets with four cutting schemes WCS, RS, BRC and BRC + RS are given in Fig. 8. And the best values of whole average accuracy Acc on six text sets with four cutting schemes WCS, RS, BRC and BRC + RS are also given in Fig. 9.

The other evaluation measures RN, PN, FN, RP, PP and FP corresponding to the best average Acc on six text sets with four cutting schemes WCS, RS, BRC and BRC + RS are given in Table 10.

From Figs. 8 and 9, we can see that the superior order is BRC + RS, RS, BRC and WCS with respect to RN and Acc. This implies that a cutting strategy of the positive samples should be added into the data pre-processing for TSC. The reason of the weaker performance of BRC than RS may be caused by the imbalance of the training data set cut only by BRC. However, adopting RS after using BRC can overcome this drawback, it can further cut some positive samples to balance the entire training data set. So BRC + RS has the best average performance among four cutting schemes.

8. Conclusion and future work

In this paper,we proposes a kind of sample cutting method for text sentiment classification on imbalanced two-class data. The main idea of the method is cutting some majority class texts in the high density boundary region to balance two-class texts. For this purpose, some notions such as high density neighbor, high density reachability between two samples are introduced. And the boundary region cutting algorithm BRC is proposed as well.

To check the validity of the proposed method three groups of experiments are designed on six text sets about hotel review and book review in Chinese, and about book, DVD, electronics, kitchen in English respectively. Three kinds of often-used feature weight schemes, TFIDF, TF and Presence are tested in a group of experiments. The experimental results show that Presence has the best average performance.

The second group of experiments are designed to observe the impact of the parameters α and β in BRC to text sentiment classification. Our experiments show that, in practical applications of BRC, we should select a smaller α if users hope a higher negative recall RN; we should select smaller α and β if users hope a higher whole accuracy Acc. And, in practical applications of BRC + RS, because RS cuts some positive samples again after BRC, the control parameter of a neighborhood α should be selected a bigger value, and the density control parameter of a neighborhood β should be selected a bigger value than that only of BRC if users hope a higher whole average accuracy Acc.

The third group of experiments are conducted to check the performance of two kinds of different sample cutting strategies BRC and BRC + RS. The experimental results indicate that BRC can indeed enhance the recall value of negative texts (i.e. the minority category). However the recall value of positive texts and the precision of negative texts will be reduced to a certain extent. In the view of combination strategy, BRC + RS can give a tradeoff between the evaluation measures, Precision and Recall on two categories and make the synthetical evaluation measure F1 obtain a larger increase. As a whole, BRC + RS has the better performance than BRC. It is because RS can further balance entire training text set after BRC clarifies the boundary region. So combining appropriate text cutting methods is needed.

It should be pointed out that determining the values of parameters α and β in BRC is empirical. We will explore the automatical method to determine them according to the data distribution in our future works.

Although the boundary region cutting algorithm BRC is aimed to text sentiment classification we believe that it is also suitable to any two-category classification problem with imbalanced sample data.

Acknowledgements

This work was supported by the National Natural Science Foundation (Nos. 61175067, 60970014, 61272095, 71031006), Natural Science Foundation of Shanxi Province (No. 2010011021-1), Shanxi Foundation of Tackling Key Problem in Science and Technology (No. 20110321027-02), the Foundation of Doctoral Program Research of Ministry of Education of China (No. 200801080006). The authors of this paper would like to thank the reviewers for their valuable comments.

References

[1] Y. Yang, Sampling strategies and learning efficiency in text categorization, in: Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access, 1996, pp. 88–95.
[2] P. Chaovalit, L.N. Zhou, Movie review mining: a comparison between supervised and unsupervised classification approaches, in: Proceedings of the 38th Hawaii International Conference on System Sciences, Big Island, Hawaii, pp. 1–9.
[3] A. Kennedy, D. Inkpen, Sentiment classification of movie and product reviews using contextual valence shifters, in: Workshop on the Analysis of Informal and Formal Information Exchange During Negotiations, 2005.
[4] G. Michael, Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis, in: Proceedings the 20th International Conference on Computational Linguistics, 2004.
[5] B. Pang, L. Lee, S. Vaithyanathan, Thumbs Up? Sentiment Classification Using Machine Learning Techniques, EMNLP, 2002.
[6] T. Mullen, C. Nigel, Sentiment analysis using support vector machines with diverse information sources, in: Dekang Lin, Dekai Wu (Eds.), in: Proceedings of EMNLP, 2004, Barcelona, Spain. 2004, pp. 412–418.
[7] S.B. Tan, J. Zhang, An empirical study of sentiment analysis for Chinese documents, Expert Systems with Application 34 (4) (2008) 2622–2629.
[8] S.G. Wang, Y.J. Wei, W. Zhang, D.Y. Li, W. Li, A hybrid method of feature selection for Chinese text sentiment classification, in: Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery IEEE Computer Society, 2007, pp. 435–439.
[9] Q. Ye, B. Lin, Y.J. Li, Sentiment classification for Chinese reviews: a comparison between SVM and semantic approaches, in: the 4th International Inference on Machine Learning and Cybernetics ICMLC, 2005.
[10] Q. Ye, Z.Q. Zhang, L. Rob, Sentiment classification of online reviews to travel destinations by supervised machine learning approaches, Expert System with Application 36 (3) (2009) 6527–6535.
[11] S.G. Wang, D.Y. Li, X.L. Song, Y.J. Wei, H.X. Li, A feature selection method based on improved fisher's discriminant ratio for text sentiment classification, Expert Systems with Applications 38 (7) (2009) 8696–8702.
[12] A. Nickerson, N. Japkowicz, E. Milios, Using unsupervised learning to guide resampling in imbalanced data sets, in: Proceedings of the Eighth International Workshop on AI and Statistics, 2001, pp. 261–265.
[13] C. Elkan, The foundations of cost-sensitive learning, in: Proceedings of the 17th International Joint Conference on Artificial Intelligence, 2001, pp. 973–978.
[14] G.M. Weiss, F. Provost, Learning when training data are costly: the effect of class distribution on tree induction, Journal of Artificial Intelligence Research 19 (2003) 315–354.
[15] D. Mladenic, M. Grobelnik, Feature selection for unbalanced class distribution and Naive Bayes, in: Proceedings of the 16th International Conference on Machine Learning, 1999, pp. 258–267.
[16] Z. Zheng, X. Wu, R. Srihari, Feature selection for text categorization on imbalanced data, ACM SIGKDD Explorations Newsletter 6 (1) (2004) 80–89.
[17] H. Guo, H.L. Viktor, Learning from imbalanced data sets with boosting and data generation: the DataBoost-IM approach, ACM SIGKDD Explorations Newsletter 6 (1) (2004) 30–39.
[18] Y. Liu, H.T. Loh, A. Sun, Imbalanced text classification: a term weighting approach, Expert System with Applications 36 (2009) 690–701.
[19] A.X. Sun, E.P. Lim, Y. Liu, On strategies for imbalanced text classification using SVM: a comparative study decision support systems 48 (2009) 191–201.
[20] Y.M. Yang, J.O. Pedersen, A comparative study on feature selection in text categorization, ICML (1997) 412–420.
[21] J.D.M. Rennie, L. Shih, J. Teevan, D.R. Karger, Tackling the poor assumptions of Naive Bayes text classifiers, in: Proceedings of the 20th International Conference on Machine Learning, Washington, DC, USA, 2003, pp. 616–623.
[22] B.L. Li, Q. Lu, S.W. Yu, An adaptive *k*-nearest neighbor text categorization strategy, ACM Transactions on Asian Language Information Processing (TALIP) 3 (4) (2004) 215–226.
[23] M. Gamon, Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis, in: Proceedings of the 20th International Conference on Computational Linguistics, 2004, pp. 841–846.
[24] G. Salton, Term-weighting approaches in automatic text retrieval, Information Processing and Management 24 (5) (1988) 513–523.

[25] P.D. Turney, M.L. Littman, Measuring praise and criticism: inference of semantic orientation from association, ACM Transactions on Information Systems 21 (4) (2003) 315–346.
[26] V. Hatzivassiloglou, R. Kathleen Mckeown, Predicting the semantic orientation of adjectives, in: Proceeding of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the ACL, Association for Computational Linguistics, New Brunswick, 1997, pp. 174–181.
[27] S.M. Kim, E. Hovy, Determining the sentiment of opinions, in: Proceedings of the 20th International Conference on Computational Linguistics (COLING), 2004, pp. 1367–1373.
[28] X.J. Wan, Bilingual co-training for sentiment classification of Chinese product reviews, Computational Linguistics 37 (3) (2011) 587–616.
[29] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, in: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP), 2005, pp. 347–354.
[30] S. Dasgupta, N. Vincent., Mine the easy, classify the hard: A semi-supervised approach to automatic sentiment classification. in: Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP (ACL-IJCNLP), Suntec. 2009, pp. 701–709.
[31] A. Devitt, A. Khurshid, Sentiment polarity identification in financial news: a cohesion-based approach, in: Proceedings of the 45th Annual Meeting of the Association of Stanford, CA. Computational Linguistics (ACL), 2007, pp. 984–991.
[32] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Computational Linguistics 37 (2) (2011) 267–307.
[33] R. Xia, C.Q. Zong, S.S. Li, Ensemble of feature sets and classification algorithms for sentiment classification, Information Sciences 181 (2011) 1138–1152.
[34] M. Kubat, S. Matwin, Addressing the curse of imbalanced training sets: One-sided selection, in: Proceedings of ICML-97, 1997, pp. 179-186.
[35] R. Barandela, J. Sanchez, V. Garcia, E. Rangel, Strategies for learning in class imbalance problems, Pattern Recognition 36 (2003) 849–851.
[36] N. Chawla, N. Japkowicz, A. Kotcz, Editorial:special issue on learning from imbalanced data sets, SIGKDD Exploration Newsletter 6 (1) (2004) 1–6.
[37] R. Akbani, S. Kwek, N. Japkowicz, Applying support vector machines to imbalanced datasets, in: Proceedings of ECML'04, 2004, Pisa, Italy, pp. 39–50.
[38] N. Chawla, K. Bowyer, L. Hall, W. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16 (2002) 321–357.
[39] S. Yen, Y. Lee, Cluster-Based under-sampling approaches for imbalanced data distributions, Expert Systems with Applications 36 (2009) 5718–5727.
[40] P. Juszczak, R. Duin, Uncertainty sampling methods for one-class classifiers, in: Proceedings of ICML-03, Workshop on Learning with Imbalanced Data Sets II, 2003, pp. 81–88.
[41] Z. Zhou, X. Liu, Training cost-sensitive neural networks with methods addressing the class imbalance problem, IEEE Transaction on Knowledge and Data Engineering 18 (2006) 63–77.
[42] H. He, E. Garcia, Learning from imbalanced data, IEEE Transactions on Knowledge and Data Engineering in Knowledge and Data Engineering 21 (9) (2009) 1263–1284.
[43] N. Japkowicz, S. Stephen, The class imbalance problem: a systematic study, Intelligent Data Analysis 6 (2001) 429–450.
[44] E.F. Combarro, E. Montanes, I. Diaz, J. Ranilla, R. Mones, Introducing a family of linear measures for feature selection in text categorization, IEEE Transactions on Knowledge and Data Engineering (TKDE) 17 (9) (2005) 1223–1232.
[45] X.Y. Liu, J.X. Wu, Z.H. Zhou, Exploratory under-sampling for class-imbalance learning, IEEE Transaction on Systems Man and Cybernets – Part B: Cybernetics 39 (2) (2009) 539–550.
[46] A. Sun, E.P. Lim, B. Benatallah, M. Hassan, FISA: feature-based instance selection for imbalanced text classification, in: Proceedings of PAKDD'06, 2006, Singapore, pp. 250–254.
[47] S.S. Li, Z.Q. Wang, G.D. Zhou, S.Y.M. Lee, Semi-supervised learning for imbalanced sentiment classification, in: Proceedings of the Twenty-Second International Joint Conference on Artificial intelligence, 2011, pp. 1826–1831.
[48] Z.Q. Wang, S.S. Li, G.D. Zhou, P.F. Li, Q.M. Zhu, Imbalanced sentiment classification with multi-strategy ensemble learning. in: Proceedings of IALP-2011, 2011, pp. 131–134.
[49] S.S. Li, G.D. Zhou, Z.Q. Wang, S.Y.M. Lee, R.Y. Wang, Imbalanced sentiment classification, in: Proceedings of CIKM-2011.
[50] C. Drummond, R.C. Holte, C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling, in: Proceedings of Working Notes ICML Workshop Learn, Imbalanced Data Sets, Washington DC, 2003.
[51] D. Ali, Using time topic modeling for semantics-based dynamic research interest finding, Knowledge-Based Systems 26 (2012) 154–163.
[52] F. Jacquenet, C. Largeron, Discovering unexpected documents in corpora, Knowledge-Based Systems 22 (6) (2009) 421–429.
[53] V.N. Vapnik. The nature of statistical learning theory. Springer-Verlag, New York, 1995.