Imbalanced Sentiment Classification

Shoushan Li[†] Guodong Zhou[†] Zhongqing Wang[†] Sophia Yat Mei Lee[‡] Rangyang Wang[†]

[†]Natural Language Processing Lab
Soochow University, Suzhou, China
{shoushan.li, wangzq870305,wangrongyang.nlp} @gmail.com, gdzhou@suda.edu.cn

[‡] Language Centre
Hong Kong Baptist University, Hong Kong
sophiaym@gmail.com

ABSTRACT

Sentiment classification has undergone significant development in recent years. However, most existing studies assume the balance between negative and positive samples, which may not be true in reality. In this paper, we investigate imbalanced sentiment classification instead. In particular, a novel clustering-based stratified under-sampling framework and a centroid-directed smoothing strategy are proposed to address the imbalanced class and feature distribution problems respectively. Evaluation across different datasets shows the effectiveness of both the under-sampling framework and the smoothing strategy in handling the imbalanced problems in real sentiment classification applications.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Text analysis*

General Terms

Algorithms, Experimentation

Keywords

Opinion Mining, Sentiment Classification, Imbalanced Classification

1. INTRODUCTION

Sentiment classification aims to predict sentiment polarity of a text [8] and it plays a critical role in many NLP applications However, most existing studies on sentiment classification assume the balance between the numbers of positive and negative samples, which may not hold in practice. Actually, many sentiment classification applications involve imbalanced class distributions in that the sample number of one class in the training data is much larger than the other class. We call this specific kind of sentiment classification as imbalanced sentiment classification, in which the class with a larger amount of samples is referred to as *majority class* and the other class with a smaller amount of samples is referred to as *minority class*.

In fact, imbalanced classification has been proven challenging in the machine learning research community [4]. Many approaches have been proposed to deal with the imbalanced class distribution problem, such as re-sampling [2], one-class classification [3], and cost-sensitive learning [14]. Unfortunately, none of the above approaches can be readily applied to imbalanced sentiment classification due to its specific characteristics.

In imbalanced classification, *majority class* normally contains more kinds of occurring features than *minority class*. For simplicity, we refer to this phenomenon as imbalanced feature distribution. Such phenomenon becomes worse in imbalanced sentiment classification since sentiment classification often involves a small number of positive and negative samples. It further worsens due to the sparseness of effective sentimental features. On one hand, sentiment classification faces the same challenge of high feature dimension as text categorization. On the other hand, the effective sentimental features in a sample are rather rare in sentiment classification, considering infrequent occurrence of sentimental words in text. For example, while the feature dimension of a typical sentiment classifier may be up to tens of thousands, there are only dozens of effective sentimental features (e.g., sentimental words) in a sample.

The imbalanced feature distribution problem can cause severe problems in the training process of imbalanced sentiment classification. Normally, the features that merely occur in the *majority class* (not occurring in *minority class*, called majority unique features) can be a strong distinguishing clue in the classifier. Nevertheless, considering that the number of effective sentimental features (e.g., sentimental words) is significantly fewer than that of other features (e.g., those words about facts) in sentiment classification, most of the majority unique features will contribute abnormally. As a result, if we use all the training samples to train a classifier, the classifier will have a strong tendency to wrongly predict a sample from the *minority class* as the *majority class*. This indicates the necessity of dealing with the imbalanced feature distribution problem in imbalanced sentiment classification.

In this paper, we propose a clustering-based stratified under-sampling framework to overcome the imbalanced class distribution problem in imbalanced sentiment classification. Under this framework, the samples in the *majority class* are first grouped into several clusters and then a suitable number of samples are selected from each cluster to form the training

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10…$10.00.

samples of *majority class*. The intuition is that these selected samples using the stratified under-sampling framework should be more representative than those by random selection. Moreover, a centroid-directed smoothing strategy is proposed to overcome the imbalanced feature distribution problem by linearly interpolating a sample with the centroid of the cluster to which this sample belongs. Since the centroid represents the average feature distribution of all occurring features in the cluster, our smoothing strategy can greatly increase the sample robustness and reduce its feature sparseness.

2. RELATED WORK

Early studies on sentiment classification mainly focus on unsupervised learning methods, which build a sentiment classifier without any labeled data. In such methods, the relationship between two words (e.g., a seed word and any other word) is usually first extracted from some knowledge resources, such as WordNet and unlabeled data. Then, such relationship is used to compute the semantic orientation of a word or even the sentiment polarity of a text [12]. In general, the performance of unsupervised learning methods is too low to meet the requirements of real applications.

Compared to unsupervised leaning, supervised learning methods often perform much better due to the availability of labeled data and become more popular since the pioneer work on sentiment classification by Pang et al. [8]. In particular, various kinds of information have been explored to improve the bag-of-words model [5][6][11]. Unfortunately, the performance of a supervised learning method drops dramatically when adapted to a new domain. This arouses wide interests on the research of domain adaptation in sentiment classification [1].

Besides domain adaptation, the imbalanced class distribution problem is another major reason which hurts the wide application of sentiment classification. To the best of our knowledge, our work is the first study on imbalanced sentiment classification.

3. CLUSTERING-BASED STRATIFIED UNDER-SAMPLING FRAMEWORK

3.1 Overview

Just as described in the introduction, imbalanced feature distribution in imbalanced sentiment classification is much due to the conflict between the high feature dimension problem (the high number of possible features in sentiment classification) and the feature sparseness problem (infrequent occurrence of sentimental words in a sample). Such imbalance in the feature distribution becomes even worse due to the imbalanced class distribution since the number of occurring features in the *minority class* would be much fewer than that in the *majority class*.

To have a better understanding of the imbalanced feature distribution phenomenon in imbalanced sentiment classification, Table 2 gives the statistics over two typical domains on the number of features occurring in the positive and negative classes, denoted as n_+ and n_- respectively, with the ratios of n_+ / n_- being around 2.

Table 1: Feature distributions on the number of occurring features in the positive and negative classes across two typical domains

Domain	n_+	n_-	n
Beauty	7,315	4,364	8,945
Computer	12,646	5,527	14,465

3.2 Stratified Under-sampling

As a popular sampling method in statistics, stratified sampling first groups the members of a population into a few relatively homogeneous subgroups (i.e. strata) according to one certain property and then selects samples from each stratum. It is believed that stratified sampling is able to select better samples to represent the distribution of the whole dataset. Previous work justifies its effectiveness theoretically and empirically in both general applications [7] and specific NLP applications such as semantic relation extraction between named entities [9] [10].

The basic motivation of our using clustering-based stratified sampling is to select some "representative" samples from the *majority class*. In particular, the same number of “representative” samples is selected from the *minority class*. Therefore, our sampling approach is basically a non-random under-sampling approach. The reason why we adopt under-sampling instead of over-sampling is basically due to its better performance. Please refer to Figure 2 in Section 6.2 for details.

Clustering groups the samples in the *majority class* into several strata. Considering that the strata may be skewed, the number of selected samples from each cluster is tuned according to the size of each stratum. Given N_{MA} samples in the *majority class* and N_{MI} samples in the *minority class*, the number of samples selected from the *i-th* stratum S_i should be $N_i = \frac{N_{MI}}{N_{MA}} \times |S_i|$.

Input: The training data and the number of strata being clustered, denoted as K

Output: Balanced training data

Algorithm:

1) Cluster the samples in the *majority class* into K strata using a clustering algorithm.
2) Calculate the number of samples being sampled for each stratum S_i , $i = \{1, 2, ..., K\}$
3) Perform intra-strata sampling in each stratum.
4) Combine the selected *majority class* samples from all the strata to form the *majority class* training data
5) Merge the *majority class* training data and all *minority class* data to obtain the balanced training dataset.

Figure 1: Clustering-based stratified under-sampling

3.3 Intra-stratum Sampling

Given the strata and the number of selected samples from each stratum, a natural question arises as to how to select the samples from each stratum. This can be viewed as intra-stratum sampling, which chooses a certain amount of samples from inside individual stratum [10].

In particular, we employ a diversity-motivated scheme to perform intra-stratum sampling with the objective to maximize the training utility of all the samples from a stratum. That is, those samples with high variance to each other are preferred, avoiding similar samples from a stratum. In particular, we first select a random candidate sample and then exclude its nearest two samples. This process repeats until enough samples are obtained. Figure 1 illustrates the clustering-based stratified under-sampling framework with intra-stratum sampling.

4. CENTROID-DIRECTED SMOOTHING

In this paper, a centroid-directed smoothing strategy is proposed to alleviate the imbalanced feature distribution problem in imbalanced sentiment classification.

Although under-sampling can balance both class and feature distributions by eliminating many samples from the *majority class* to keep the balance between positive and negative samples, a lot of *majority class* features are excluded. Without these excluded features, the selected samples may not be able to well represent the feature distribution of the whole dataset in the *majority class*, even when the clustering-based stratified under-sampling is used.

The centroid-directed smoothing strategy merges the feature vector of each sample in the *majority class* with that of the centroid of the corresponding cluster it belongs to. Accordingly, the feature vector of each sample in the *minority class* is extended with itself. As the number of non-zero elements in the centroid is much larger than the number of the non-zero elements in a *minority class* sample, the feature imbalanced distribution problem still exists. However, the centroid-directed smoothing strategy actually introduces another imbalanced factor in imbalanced feature weights since the feature weights in the centroid are usually much lower than the Boolean weights in a sample in the *minority class*. Therefore, these two kinds of imbalanced factors result in a slightly more balanced classifier classifying the samples from both the positive and negative classes.

Formally, the centroid feature vector c_i of the cluster S_i is calculated as the mean of feature vectors of all the samples in the *i-th* cluster $c_i = \left(\sum_{x \in S_i} x / |S_i| \right)$.

In summary, the centroid-directed smoothing strategy maps the feature vector of a sample x to a new feature vector x^{new} as follows,

$$\begin{cases} x^{new} = < x, c_i > \text{if } x \in S_i \text{ and } x \in X_{MA} \\ x^{new} = < x, x > \text{ if } x \in X_{MI} \text{ or } x \in X_{Test} \end{cases}$$

Where X_{MA} represents the samples in the *majority class*, X_{MI} represents the samples in the *minority class*, and X_{Test} represents the samples in the whole test data, regardless of what class they belong to.

5. EXPERIMENTAL STUDIES

In this paper, two datasets are used to investigate the performance of our approach on imbalanced sentiment classification. The first one[1] is a widely-used public dataset collected by Blitzer et al. [1] which consists of four domains: Book, DVD, Electronic, and Kitchen. For our experiment, each domain contains 400 negative samples (randomly selected from the 1000 original negative samples) and 1000 positive samples.

We adopt the popular geometric mean (*G-mean*), defined as $G - mean = \sqrt{TP_{rate} \times TN_{rate}}$, where TP_{rate} is the true positive rate (also called positive recall or sensitivity) and TN_{rate} is the true negative rate (also called negative recall or specificity) [4].

Finally, all the classifiers adopt the Maximum Entropy (ME) algorithm available with the *Mallet*[2] *tool*, and the same Boolean-weighted unigram. For thorough comparison on imbalanced sentiment classification, various settings are explored:

1) Full training (FullT): directly throwing all the training data for training.

2) Random over-sampling (OverS): performing over-sampling by randomly selecting the samples from the *minority class*.

3) Random under-sampling (UnderS): performing under-sampling by randomly selecting the samples from the *majority class*.

4) One-class classification (OneClass): performing one-class classification as proposed by Juszczak and Duin [3] using the *lib-SVM tool*[3].

5) Cost-sensitive classification (CostSensitive): performing cost-sensitive classification as proposed by Zhou and Liu [14] using the *lib-SVM tool*. Here, the cost weight for a *majority-class* sample is set to the imbalanced ratio between the *minority class* and *majority class* samples in each domain while the cost weight for a *minority-class* sample is 1.

6) Clustering-based under-sampling (ClusterU): performing clustering-based stratified under-sampling.

7) Clustering-based under-sampling plus centroid-directed smoothing (ClusterUC): performing both clustering-based stratified under-sampling and centroid-directed smoothing, as proposed in this paper.

Since most of the above settings involve random selection of samples, we run 20 times for each setting and report the average performance.

Table 2 compares the seven settings. It shows that both random over-sampling and random under-sampling significantly outperform full training due to balance keeping between positive and negative samples. It also shows that random under-sampling significantly outperforms random over-sampling largely due to the ignorance of imbalanced feature distribution by random over-sampling. Furthermore, it shows that one-class classification does not fit our task at all and that random under-sampling is rather difficult to beat. Generally, random under-sampling performs slightly better than cost-sensitive

[1] http://www.cs.jhu.edu/~mdredze/datasets/sentiment/index2.html

[2] http://mallet.cs.umass.edu/

[3] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Table 2: Performances of the seven imbalanced classification algorithms

Domain	FullT	OverS	UnderS	OneClass	CostSensitive	ClusterU	ClusterUC
Book	0.575	0.631	0.697	0.492	0.663	0.698	**0.713**
DVD	0.574	0.621	0.706	0.527	0.712	0.714	**0.731**
Electronic	0.675	0.675	0.782	0.546	0.748	0.789	**0.797**
Kitchen	0.677	0.677	0.770	0.861	0.773	0.782	**0.803**

classification. Although clustering-based under-sampling (ClusterU) employs cleverer selection strategies, they can only achieve comparable performances with random under-sampling. Observation on the features occurring in the selected samples shows that only about half of the features remain regardless of what kind of under-sampling (random or clustering-based) is used. Since half of the features can hardly well represent the feature distribution of the whole data, this justifies why cleverer under-sampling fails to improve the performance. It also shows that centriod-directed smoothing strategy significantly improves the performance of clustering-based under-sampling in all domains. This suggests the importance of resolving the imbalanced feature distribution problem and the effectiveness of our proposed centroid-directed smoothing strategy.

6. CONCLUSION

In this paper, we address the issue of imbalanced sentiment classification by taking into account both the imbalanced class and feature distribution problems. In particular, a clustering-based stratified under-sampling framework and a centroid-directed smoothing strategy are proposed to deal with the imbalanced class and feature distribution problems respectively. Evaluation shows the effectiveness of our approach.

Acknowledgments

The research work described in this paper has been partially supported by three NSFC grants, No. 61003155, No. 60873150 and No. 90920004 and Open Projects Program of National Laboratory of Pattern Recognition.

7. REFERENCES

[1] Blitzer J., M. Dredze, and F. Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification. In *Proceedings of ACL-07*. 440-447.

[2] Chawla N., K. Bowyer, L. Hall, and W. Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. 16(2002), 321–357.

[3] Juszczak P. and R. Duin. 2003. Uncertainty sampling methods for one-class classifiers. In *Proceedings of ICML-03, Workshop on Learning with Imbalanced Data Sets II*. 81–88.

[4] Kubat M. and S. Matwin. 1997. Addressing the curse of imbalanced training sets: one-sided selection. In *Proceedings of ICML-97*. 179–186.

[5] Li S., S. Lee, Y. Chen, C. Huang, and G. Zhou. 2010. Sentiment classification and polarity shifting. In *Proceedings of COLING-10*. 635-643.

[6] Nakagawa T., K. Inui, and S. Kurohashi. 2010. Dependency tree-based sentiment classification using CRFs with hidden. In *Proceedings of NAACL-10*. 786–794.

[7] Neyman J. 1934. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*. 97,4(1934), 558-625.

[8] Pang B., L. Lee, and S. Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of EMNLP-02*.79-86.

[9] Qian L., G. Zhou, F. Kong, and Q. Zhu. 2009. Semi-supervised learning for semantic relation classification using stratified sampling strategy. In *Proceedings of EMNLP-09*. 1437-1445.

[10] Qian L. and G. Zhou. 2010. Clustering-based stratified seed sampling for semi-supervised selation slassification. In *Proceedings of EMNLP-10*. 346-355.

[11] Riloff E., S. Patwardhan, and J. Wiebe. 2006. Feature subsumption for opinion analysis. In *Proceedings of EMNLP-06*. 440-448.

[12] Turney P. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL-02*. 417-424.

[13] Zhou Z. and X. Liu. 2006. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transaction on Knowledge and Data Engineering*. 18(2006), 63–77.