Two Methods for Classifying Bloggers' Sentiment

Long-Sheng Chen*, Li-Wei Lin

***Abstract*—On lots of commercial blogs, bloggers' negative comments about enterprisers' images or harmful evaluations of products spread quickly in the cyber space. Moreover, an exposure of an inside story in the blogosphere may influence a company's reputation. Therefore, to identify bloggers' sentiment effectively is extremely important for enterprisers. These negative comments often bring great damage to enterprises. Recently, researchers proposed lots of machine learning techniques to efficiently detect customers' negative emotions for helping companies to carefully response customers' comments. However, they don't consider the class imbalance problem which lots of bloggers' comments are positive and far fewer comments are negative. A classifier induced from an imbalanced data set has high classification accuracy for the majority class, but an unacceptable error rate for the minority class. Therefore, this study proposed two new methods, MCBS and VS to provide a possible solution. Finally, a real case of bloggers' comments regarding MP3 products will be employed to illustrate the effectiveness of our proposed methods.**

***Index Terms*—Sentiment Classification, Class Imbalance Problems, Blogs, Data Mining.**

I. Introduction

Nowadays, blogs have been considered as one of the fastest growing sections of the Internet and are emerging as an important communication mechanism that is used by an increasing number of people [1][2]. On lots of commercial blogs, bloggers' negative comments about enterprisers' images or harmful evaluations of products spread quickly in the cyber space. Moreover, an exposure of an inside story in the blogosphere may influence a company's reputation. Therefore, to identify bloggers' sentiment effectively is extremely important for enterprisers. These negative comments often bring great damage to enterprises.

Recently, researchers have paid much attention on sentiment classification [3] to identify customers' negative emotions for helping companies to carefully response customers' comments. In related works, two popular approaches, machine learning methods and information retrieval techniques, have been employed to address this problem [4] . It's reported that machine learning methods have better classification performances than information retrieval techniques. Hence, machine learning methods have become one of main streams in sentiment classification.

In machine learning methods, several approaches have been developed for classifying sentiments. For example, Abbasi et al. [5] proposed SVRCE approach to analyze emotional states. Pang et al [6] investigate several supervised machine learning methods to semantically classify movie reviews. Dave et al [7] develop a method for automatically classifying positive and negative reviews and experiment several methods related to feature selections and scoring. In the work of Chaovalit and Zhou [4], machine learning methods and semantic orientation index have been presented to classify movie reviewers' comments. The experimental results indicated that machine learning techniques have better performance, but they need additional time to be trained.

However, when applying machine learning techniques to bloggers' sentiment classification, researchers don't consider the class imbalance problem (lots of bloggers' comments are positive and far fewer comments are negative). A classifier induced from an imbalanced data set has high classification accuracy for the majority class, but an unacceptable error rate for the minority class. Therefore, this study proposed two new methods called Modified Cluster Based Sampling (MCBS) and BPN based Voting Scheme (VS), to provide possible solutions. Finally, a real case of bloggers' comments regarding MP3 products will be employed to illustrate the effectiveness of our proposed methods.

II. Class Imbalance Problems

Generally speaking, there are two major groups of methods for solving class imbalance problems. They are algorithm/models oriented approaches and data manipulation techniques. The former aims to propose new learning mechanism or modify existing methods. This group includes Support Vector Machines (SVM), one class learning, information granulation based methods, and so on. One class learning techniques are to detect rare events involves an unsupervised framework, i.e. outlier detection or one-class classification [8]. Initially, minority examples are completely ignored and a model is trained by using all examples from the majority class (target class). Then, the outliers are detected as the data points with low probability of occurrence, small number of neighboring examples. Moreover, SVM is usually used to tackle class imbalance problem [9]. In information granulation based methods [10][11][12], objects are considered as "information granules" . Through the process of information granulation, the class imbalance situations

Manuscript received December 25, 2009. This work was supported in part by the National Science Council of Taiwan, R.O.C. (Grant No. NSC 98-2410-H-324-007-MY2).

*L.-S. Chen is with the Chaoyang University of Technology, Taichung 41349, Taiwan (phone: 886-4-23323000ext7752; fax: 886-4-23304902; e-mail: lschen@cyut.edu.tw).

L.-W. Lin is with the Chaoyang University of Technology, Taichung 41349, Taiwan (e-mail: bow1221@cyut.edu.tw).

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

will be improved.

The second group involves several re-sampling techniques. They are: (1)under-sampling, methods in which the minority population is kept intact, while the majority population is under-sampled, (2) oversampling, methods in which the minority examples are over-sampled so that the desired class distribution is obtained in the training set, (3) cluster based sampling, methods in which the representative examples are randomly sampled from clusters [13], (4) adjust misclassification costs matrices, methods in which the prediction accuracy is improved by adjusting the cost (weight) for each class [14], (5) Mahalanobis Distance (MD) based two phase learning, methods in which MD is firstly used to screen the majority examples which we are 100% confidence to ensure they are truly majority [15], and (6) SOM based method (SWAI), methods in which the weights of SOM to represent the constructed clusters [16].

Traditionally, the simplest approaches are re-sampling techniques. Such re-sampling will modify the class distributions of the training data. However, over-sampling cannot gain new information about the minority class, since under-sampling may lose useful information about the majority class [17]. To enhance over-sampling and under-sampling, cluster based sampling technique [13] has been proposed. But, the number of clusters and how to choose representative examples are difficult to be determined. Moreover, these supervised methods lack a rigorous and systematic treatment on imbalanced data [18] and they still have some drawbacks.

In order to solve the problems of cluster based sampling, this study proposes two novel solutions, MCBS and VS for imbalanced data. A real case of bloggers' comments regarding MP3 products will be employed to evaluate the effectiveness of the proposed methods.

III. Proposed Methodologies

In this section, we will introduce the implemental steps of the proposed two methods, MCBS and VS, to tackle class imbalance problems. Decision Tree algorithm (DT) has been employed as the basic learner in this study.

A. Modified cluster based sampling (MCBS)

How to select representative examples and how to determine the number of clusters is the main problems in cluster based sampling technique. The major purpose of MCBS is to enhance the advantages of "cluster based sampling" technique by providing possible solutions. MCBS involves three phases. They are "clustering", "selecting representative examples in constructed clusters", and "learning". Detailed implemental steps can be described as below.

Phase I: Clustering

Step 1: Separate majority and minority examples into two groups. The minority population is kept intact.

Step 2: Build clusters from the majority examples.

In this study, K-mean algorithm has been employed to cluster objects. In order to determine number of clusters, two objective indexes, Entropy and Purity, are introduced. These two indexes can be defined as equations (1)&(2).

$$Entropy = \sum_{r=1}^{k} \frac{n_r}{n} E(C_r) \quad (1)$$

$$Purity = \sum_{r=1}^{k} \frac{n_r}{n} P(C_r) \quad (2)$$

where,

$$E(C_r) = -\frac{1}{\log q} \sum_{i=1}^{q} \frac{n_r^i}{n_r} \log \frac{n_r^i}{n_r} \quad (3)$$

$$P(C_r) = -\frac{1}{n_r} \max_i (n_r^i) \quad (4)$$

And q denotes the number of class labels, n_r^i menas the percentage of class label r in cluster i, k is number of constructed clusters.

Phase II: Selecting representative examples in constructed clusters

Step 3: Calculate the distance between objects and the central point of cluster.

Step 4: Select those examples that are close to the central point to be the representative of the constructed clusters.

Phase III: Learning

Step 5: Join those representative majority examples and the original minority examples together to be the training data set.

Step 6: Build a classifier by implementing DT.

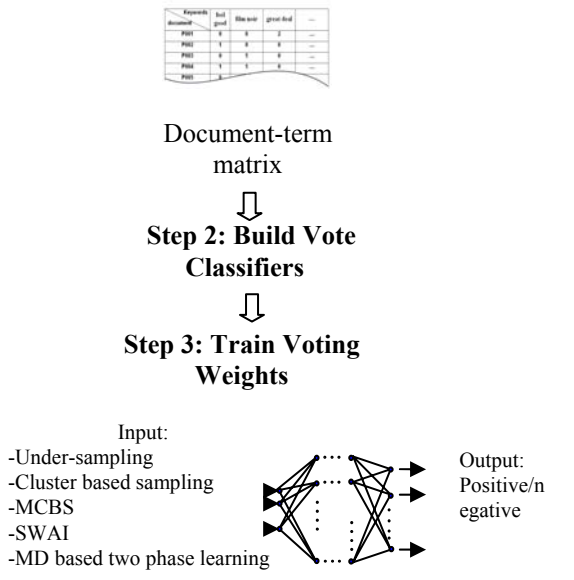

Figure 1 The implemental procedure of the proposed VS method

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

B. BPN based voting scheme

The main purpose of the proposed BPN based Voting Scheme (VS) is to construct a multi-classifiers decision scheme. In VS method, based on five methods for dealing with the class imbalanced data, including "under-sampling", "cluster based sampling", "MCBS", "SWAI", and "MD based two phase learning", we train the BPN to adjust the voting weights of these methods. Then, depending on the result of 5 classifier's vote, we can find the best classification performance among these methods. VS method involves 2 major phases, constructing voting classifiers (phase 1) and vote weights adjustment by BPN training (phase 2). The detailed implemental algorithm can be found as Figure 1.

IV. Implementation

A. The Employed Data

To evaluate the effectiveness of our approaches, we employ an actual case of the bloggers' comments regarding MP3 products from a real world blog, "reviewcenter (www.reviewcentre.com)". 338 comments have been collected. In addition, 10-star rating system of "reviewcenter" website has been used to define the class label. For example, if the rate of one comment is above 7-star (below 4-star), this comment will be defined as positive (negative). Among 338 collected data, 34 of them are negative and 304 are negative comments.

In addition, we employed the shareware Rubryx which can be downloaded at the website (http://www.sowsoft.com/rubryx) to segment words in this study. Rubryx segments words based on n-gram (unigram, bigrams, and tri-grams) features. Before extracting n-gram key words, some frequently used stop words should be removed. Readers can find a useful stop word list at http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/ stop_ words. Finally, we extracted 300 keywords to describe the collected data. A 338×300 document-term matrix can be constructed after data preparation phase. Moreover, 4 fold cross validation experiment has been used in this study.

B. Performance Measurement Index

When learning from imbalanced data, we should discuss the effectiveness of performance index. Traditionally, the easiest way to evaluate the classification performance is based on the confusion matrix shown as Table 1.

Table 1. Confusion matrix for binary class problem

	Predicted Positive	Predicted Negative
Actual Positive	*TP* (the number of True Positive)	*FN* (the number of False Negative)
Actual Negative	*FP* (the number of False Positive)	*TN* (the number of True Negative)

The most popular performance index is overall accuracy. However, when handling imbalanced data, this measure is often not enough [10][11][12]. Another fact is the index considers different misclassification errors to be equally important. But as we know, a highly skewed class situation does not have equal error costs that favor the minority class, which is often the class of primary interest. Therefore, following the available studies [10][11][12][19][20][21][22], we use Overall Accuracy (OA), Positive Accuracy (PA), Negative Accuracy (NA), and Geometric Mean of PA and NA (G-Mean) to evaluate classifiers. Overall Accuracy is defined as

$$\text{Overall Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

In this study, PA and NA represent the ability of detecting the positive (majority) and negative (minority) examples, respectively. They are defined as

$$PA = TP \big/ TP + FN \tag{6}$$

$$NA = TN \big/ FP + TN \tag{7}$$

Besides, we should consider another integrated index, G-mean which is defined as

$$G - mean = \sqrt{\text{PA} \times NA} \tag{8}$$

This measure is to maximize the accuracy on each of two classes while keeping these accuracies balanced. For instance, a high PA by a low NA will result in a poor G-mean.

C. The Results

In addition to MCBS and VS, we also implemented original DT without doing nothing regarding imbalanced data and traditional re-sampling techniques including over-sampling (OS) and traditional cluster based sampling (CBS) for the purpose of comparison. Table 2 shows the computational results of all implemented techniques. In this table, PA and NA represent the abilities of detecting majority and minority examples, respectively. From the results, we can find the original DT algorithm which doesn't implement any technique for imbalanced data has the highest PA (97.9%) and the lowest NA (2.78%). It means DT has a serious class imbalance problem for handling the collected data. Therefore, it's necessary to run re-sampling techniques.

If we think the cost of misclassifying the minority examples into the majority class is equal to the misclassification cost of identifying majority example, OA could be used the measurement metric. From these results, DT has a high OA (85.71%) greater than OS (81.84%), CBS (41.53%), SWAI (69.02%), MD-DT (83.27%), MCBS (66.84%), and VS (84.36%). But, as we know, the cost of misclassifying the negative bloggers' comments (the minority) can not be equal to the cost of misclassifying the positive comments (the majority). Generally speaking, enterprisers know the damage of negative comments is larger than the positive comments. They need to detect negative comments immediately and effectively to avoid they spreads in the cyber space. Therefore, the performance index OA is not suitable for imbalanced classifiers.

G-mean which considers both PA and NA in the same time is very suitable to be employed to imbalance classification tasks. From table 2, the performances of our proposed VS (Mean: 66.62%, S.D.:3.12%) and MCBS (Mean: 66.47%, S.D.:4.84%) are better than OS (Mean: 61.15%, S.D.: 10.59%), CBS (Mean: 55.74%, S.D.: 8.71%), SWAI (Mean: 8.33%, S.D.: 16.67%), MD-DT (Mean: 47.77%, S.D.: 6.10%). But, VS has a smaller standard deviation than MCBS. Besides, the OA of VS (84.36%) is also larger MCBS's OA (66.84%).

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Table 2. Computational Results

Positive accuracy (%)							
Method / Experiment	DT	OS	CBS	SWAI	MD-DT	MCBS	VS
Fold1	100	86.67	46.67	0.00	96.67	56.67	88.33
Fold2	100	95.00	33.33	100	96.67	73.33	98.33
Fold3	93.3	86.67	30.00	100	86.67	73.33	88.33
Fold4	98.4	81.25	26.56	100	87.50	64.06	82.81
Mean	97.9	87.40	34.14	75.0	91.88	66.85	89.45
S. D.	3.16	5.68	8.80	50.0	5.54	8.07	6.47
Negative accuracy (%)							
Method / Experiment	DT	OS	CBS	SWAI	MD-DT	MCBS	VS
Fold1	0.00	55.56	100.00	100.0	33.33	77.78	55.56
Fold2	0.00	22.22	88.89	0.00	22.22	66.67	44.44
Fold3	11.11	44.44	88.89	11.11	22.22	66.67	44.44
Fold4	0.00	55.56	88.89	0.00	22.22	55.56	55.56
Mean	2.78	44.44	91.67	27.78	25.00	66.67	50.00
S. D.	5.56	15.71	5.56	48.43	5.56	9.07	6.42
Overall accuracy (%)							
Method / Experiment	DT	OS	CBS	SWAI	MD-DT	MCBS	VS
Fold1	86.96	82.61	53.62	13.04	88.41	59.42	84.06
Fold2	86.96	85.51	40.58	86.96	86.96	72.46	91.30
Fold3	82.61	81.16	37.68	88.41	78.26	72.46	82.61
Fold4	86.30	78.08	34.25	87.67	79.45	63.01	79.45
Mean	85.71	81.84	41.53	69.02	83.27	66.84	84.36
S. D.	2.09	3.09	8.46	37.32	5.15	6.66	5.02
G-Mean (%)							
Method / Experiment	DT	OS	CBS	SWAI	MD-DT	MCBS	VS
Fold1	0.00	69.39	68.31	0.00	56.76	66.39	70.05
Fold2	0.00	45.95	54.43	0.00	46.35	69.92	66.11
Fold3	32.20	62.06	51.64	33.33	43.89	69.92	62.66
Fold4	0.00	67.19	48.59	0.00	44.10	59.66	67.83
Mean	8.05	61.15	55.74	8.33	47.77	66.47	66.66
S. D.	16.10	10.59	8.71	16.67	6.10	4.84	3.12

V. Conclusions

In this study, we consider class imbalance problems in bloggers' sentiment classification and proposed two novel methods, MCBS and VS, to attempt to provide possible solutions. We also use a real case of bloggers' sentiment classification to evaluate MCBS and VS. Experimental results indeed indicated that both of our methods outperform traditional techniques for imbalanced data, including over-sampling, cluster based sampling, SWAI, MD based two phase learning (MD-DT). The results also told us that high dimensional data (one of characteristics in textual data) could decrease the performance of traditional techniques for imbalanced data. To confirm the mentioned above results, additional experiments of higher dimensional data should be implemented in the future.

References

[1] E. Cohen, and B. Krishnamurthy, "A short walk in the Blogistan," *Computer Networks*, vol. 50, no. 5, April 2006, pp. 615-630.

[2] T. Singh, L. Veron-Jackson, J. Cullinane, "Blogging: A new play in your marketing game plan," *Business Horizons*, vol. 51, no.4, July-August 2008, pp.281-292.

[3] C. H. Wu, Z. J. Chuang, and Y. C. Lin, "Emotion recognition from text using semantic labels and separable mixture models," *ACM Transactions on Asian Language Information Processing*, vol. 5, no. 2, 2006, pp. 165-183.

[4] P. Chaovalit and L. Zhou, "Movie review mining: a comparison between supervised and unsupervised Classification Approaches," *Proceedings of the 38th Annual HICSS*, 2005.

[5] A. Abbasi, H. Chen, S. Thoms, T. Fu, "Affect analysis of web forums and blogs using correlation ensembles," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 9, September 2008, pp. 1168-1180.

[6] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," *EMNLP*, 2002, pp. 79-86.

[7] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: opinion extraction and semantic classification of product reviews," *The 12th WWW*, 2003, pp. 519-528.

[8] L. M. Manevitz and M. Yousef, "One class SVMs for document classification," *Journal of Machine Learning Research*, vol. 2, March 2002, pp.139-154.

[9] G. Wu and E. Y. Chang, "KBA kernel boundary alignment considering imbalance data distribution," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, 2005, pp. 786-795.

[10] C. T. Su, L. S. Chen, and T. L. Chiang, "A neural network based information granulation approach to shorten the cellular phone test process," *Computers In Industry*, vol. 57, no. 5, 2006a, pp. 412-423.

[11] C. T. Su, L. S. Chen, and Y. Yih, "Knowledge acquisition through information granulation for imbalanced data," *Expert System with Applications*, vol. 31, no. 3, 2006b, pp. 531-541.

[12] M -C. Chen, L. S. Chen, C. C. Hsu, and W. R. Zeng, "An information granulation based data mining approach for classifying imbalanced data," *Information Sciences*, vol. 178, no. 16, 2008, pp. 3214-3227.

[13] H. Altincay, and C. Ergun, "Clustering based undersampling for improving speaker verification decisions using AdaBoost," *Lecture Notes in Computer Science*, vol. 3138, 2004, pp. 698- 706.

[14] G. M. Weiss, and F. Provost, "The effect of class distribution on classifier learning," *Technical Report*, *MLTR43*, *Department of Computer Science*, *Rutgers University*, 2001.

[15] L. S. Chen, C. C. Hsu, Y. S. Chang, "MDS: A Novel Method for Class Imbalance Learning," *The 3rd International Conference on Ubiquitous Information Management and Communication*, January 15-16, 2009, SKKU, Suwon, Korea.

[16] L. S. Chen, Y. S. Chang, L. W. Lin, M. C. Chen, "A SOM based approach for learning from imbalanced data sets," *The 19th Intelligent System Symposium (FAN 2009)*, Sep. 17~18 2009, Aizu- Wakamatsu, Japan.

[17] N. Japkowicz, and S. Stephen, "The class imbalance problem: a systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, 2002, pp. 429-449.

[18] K. Huang, H. Yang, I. King and M. Lyu, "Learning classifiers from imbalanced data based on biased minimax probability machine," Proceedings of the 04' *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)*, 2004, pp. 558-563.

[19] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behaviour of several methods for balancing machine learning training data," *SIGKDD Explorations*, vol. 6, no. 1, 2004, pp. 20–29.

[20] A. Estabrooks, T. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," *Computational Intelligence*, vol. 20, no. 1, 2004, pp. 18–36.

[21] H. Guo, and H. L. Viktor, "Learning from imbalanced data sets with boosting and data generation: the DataBoost- IM approach," *SIGKDD Explorations*, vol. 6, no. 1, 2004, pp. 30-39.

[22] T. Fawcett, F.J. Provost, "Adaptive fraud detection," *Data Mining Knowledge Discovery*, vol. 1, no. 3, 1997, pp. 291-316.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)