A Cost Adjusting Method for Increasing Customers' Sentiment Classification Performance

Long-Sheng Chen and Sheng-Jhe Cai

***Abstract*—The internet could be a perfect platform for spreading the electronic word of mouth (e-WOM). Consumers not only heavily depended comments regarding the products or services in social media to make their purchase decisions. The negative product reviews could cause a negative impact on business products. When online reviews increase, inevitably there will produce imbalanced class data, in which the amount of positive comments (negative comments) is far larger than the number of negative comments (positive comments). When training a classifier using this kind of imbalanced data, it'll lead to a higher accuracy for determining the majority example, but an unacceptable error for classifying the minority examples. However, in the domain of sentiment classification, the available works didn't discuss this issue to solve the imbalanced comments. Therefore, this study aims to find the best combination from the cost adjustment, under-sampling, and over-sampling methods based on support vector machines (support vector machines, SVM) to improve the classification performance of imbalanced semantic comments. A comparative analysis of the experimental results will be provided to evaluation these methods. In addition, we use a real online travel site reviews as the case study to verify the effectiveness of the methods.**

***Index Terms*—Sentiment classification, class imbalance problems, social media, text mining.**

I. Introduction

Recently, with the rise of social media, people can communicate from face-to-face to on the virtual communities. The internet could be a perfect platform for spreading the electronic word of mouth (e-WOM). Enterprises also can receive diverse customer needs and messages. Consumers not only heavily depended comments regarding the products or services in social media to make their purchase decisions, but also considered these comments as evaluation of product and service satisfaction levels. But, these online comments are inevitable to be emotional. The positive comments have a positive impact on the reputation of consumer purchase intentions, brand image, and loyalty. They have considerable influences. However, the negative product reviews could cause a negative impact on business products, are more likely to affect the real sales figures [1].

When online reviews increase, inevitably there will produce imbalanced class (category) data, in which the amount of positive comments (negative comments) is far larger than the number of negative comments (positive comments) [2]. When training a classifier using this kind of imbalanced data, it'll lead to a higher accuracy for determining the majority example, but an unacceptable error for classifying the minority examples [3]. This is called "class imbalance problems". A growing number of scholars attempted to propose solutions, such as constructing the framework of sampling model to enhance the performance of SVM (Support Vector Machines) classifier [3]. García *et al.* [4] proposed new classification indicators by integrating the two methods, performance indicators and resampling, to solve the problem of unbalanced. Yang *et al*. [5] is adopted dimension reduction and proposed a new classification method CMFS (Comprehensively measure feature selection) to improve the classification performance and accuracy of SVM.

However, in the domain of sentiment classification, the available works didn't discuss this issue to solve the imbalanced comments. Therefore, this study aims to find the best combination from the cost adjustment, under-sampling, and over-sampling methods based on support vector machines (Support Vector Machines, SVM) [6] to improve the classification performance of imbalanced semantic comments. A comparative analysis of the experimental results will be provided to evaluation these methods. In addition, we use a real online travel site reviews as the case study to verify the effectiveness of the methods.

II. Related Works

A. Electronic Word of Mouth

With the change of time and advancement of technology, and conventional word of mouth gradually becomes a concrete expression, such as text description or pictures to convey messages. So, the original word of mouth message will become a so-called electronic word of mouth (e-WOM), also known as online word of mouth (online word-of-mouth). Cheung *et al.* [1] said electronic word of mouth will affect the credibility of the product. Their study confirmed the negative electronic word of mouth, and on the willingness of consumers to influence the purchase of goods, much larger than the positive comments.

Gupta *et al*. [7] considered the presence of electronic word of mouth will make consumers spend more time with reference to the relevant product reviews and recommendations before buying goods. When consumers have an incentive to buy commodities, their interests will be higher to accept the recommendations of its information and comments. Yoo *et al*. [8] considered the electronic word of

Manuscript received March 5, 2014; revised May 29, 2014. This work was supported in part by by National Science Council of Taiwan under Grant No. NSC 101-2628-E-324-004-MY3.

L.-S. Chen and S.-J. Cai are with the Department of Information Management, Chaoyang University of Technology, Taichung 41349, Taiwan (e-mail: lschen@cyut.edu.tw).

DOI: 10.7763/IJIEE.2014.V4.460

mouth on the Internet represents a "visibility". They will affect a company's potential customer trust and loyalty. Their findings indicated that the platform between the electronic word of mouth will help to improve the delivery and consumption communication. Verhagen *et al*. [9] found that negative comments can discover the true consumer needs, and can help companies respond to the consumer for improving products and services. Therefore, we need an effective way to help us to effectively detect the negative comments.

B. Class Imbalance Problems

In recent years, more and more scholars proposed solutions for class imbalance problems, such as feature selection, machine learning methods, and others. For examples, Liu *et al*. [5] used SVM and presented a sampling approach which combined under-sampling and over-sampling. Their results showed that their sampling model can effectively improve the classification performance of SVM. García *et al*. [4] compared the performances of several sampling methods such as performance indicators, resampling, etc. Then, they proposed a new evaluation index IBA (Index of balanced accuracy). Their experimental results shown that this indicator can effectively deal with the class imbalance problems.

Zhao *et al*. [10] proposed a weighted maximum margin criterion to optimize the data set, which makes SVM accurately determine the minority class. Zhang and Li [11] proposed RWO-Sampling (Random Walk Over-Sampling) approach to deal with the imbalance data. Yang *et al*. [5] proposed CMFS (Comprehensively measure feature selection) for class imbalance problems, and made a comparison with other feature selection methods. Therefore, this study has taken three kinds of experimental methods to increase its multi-faceted aspects have increased the performance of SVM classifier comparison.

In the work of Moraes *et al*. [12], their study showed that both in the feature selection methods or changing the weights, SVM can effectively handle with imbalanced data compared to neural networks considering the computational cost. Sun *et al*. [13] found that the SVM classifier is the best method for dealing with the imbalanced data from their experiments. In addition, because SVM has a complete theory of modules and it's easy to use, suitable for high-dimensional and nonlinear classification problems. Therefore, this study uses SVM to be our basic classifier. In addition, this study employs three methods, under-sampling, over-sampling, cost adjustment method, to explore what combination methods is the best classifier for effectively identify the minor negative comments.

III. METHODOLOGY

This section describes the employed approaches in this study. Actually, there are 5 major steps shown in Fig. 1. The detailed implemental procedure listed as below.

Step 1: Data collection

We collect online reviews or product comments in social media to be our corpus. We focus on positive and negative comments.

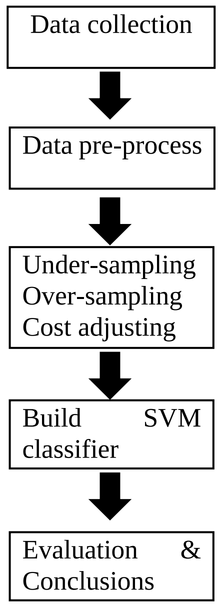

Fig. 1. The implemental procedure of this study.

Step 2: Data pre-process

For the collected sentiment data, we used uni-gram to segment sentences. Then, we will remove some unnecessary words and stop words. Next, we can build the Term-Document Matrix (TDM). In this study, TF-IDF weights will be used to describe the collect text data.

Step 3: Implement under-sampling, over-sampling, and cost-adjusting methods

Step 3.1: Implement under-sampling

The majority examples (positive comments) will be randomly removed till the amount of minority examples is equal to the amount of majority examples (positive comments).

Step 3.2: Implement over-sampling

The minority examples (negative comments) will be duplicated till the amount of minority examples is equal to the amount of majority examples (positive comments).

Step 3.3: Implement cost-adjusting

We adjust the misclassification cost till the classification performance has been improved. For example, if the cost of misclassifying the majority examples (positive comments) into minority examples (negative comments) is equal to "1", we can set the cost of misclassifying the minority examples (negative comments) into majority examples (positive comments) larger then 1. It can force the classifier tend to identify the minority negative comments.

Step 4: Build SVM classifier

Step 4.1: Construct training and test sets

Step 4.2: Select a kernel function and find optimal settings of parameters

Step 4.3: Train SVM

Step 5: Evaluation and Conclusions

In this work, we used geometric mean (GM) of positive accuracy (the ability of identifying positive comments) and negative accuracy (the ability of identifying the minority negative comments) to evaluate the classification performance. And thus, we can draw conclusions based on the experimental results.

IV. Experimental Results

A. Employed Data

This research collected online comments as experimental data. We collected data from the world's largest travel (tripadvisor) website (http://www.tripadvisor.com). This site covers all over the world tourist attractions comments. Each comment has been considered as an example. We focus on the topic, "subtropical island area– Boracay", related reviews as our corpus from March 2012 to September 2013. In tripadvisor review sites, it uses five-star ranking system for the evaluating the comments. Four-stars and above comments are marked as positive reviews. 2-stars and below comments were marked as negative comments. Those comments with 3 stars are viewed as neutral and then will be ignored. After removing 143 minor comments including non-English, icons, symbols and so on, a total of 566 examples left for further analysis. 526 examples are positive reviews 526 and 40 examples are negative comments. The background of the employed data is listed in Table I.

In addition, this work utilized 5-fold cross-validation experiments to verify the effectiveness of the used method. The collected data will be divided into five equal portions. One of them will be the test data set, and other four ones are the training data set.

TABLE I: Background of the Collected Data

Data size	Retrieval time	Class distribution	Source
566	Mar. 2012~Sep. 2013	Positive comments: 526 Negative comments 40	http://www.tripadvisor.com

B. Measurements

Traditionally, the easiest way to evaluate the classification performance is based on the confusion matrix shown as Table II.

TABLE II: Conclusion Matrix for Binary Class Classification 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)

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

$$PA = {TP}/{TP+FN} \quad (1)$$

$$NA = {TN}/{FP+TN} \quad (2)$$

We use an integrated index, G-mean which is defined as equation (3) to measure the performance of imbalance classification. 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 (GM).

$$G-mean = \sqrt{\mathrm{P}A \times NA} \quad (3)$$

C. Results

Table III shows the results of original SVM, over-sampling, and under-sampling. In this table, we list the mean value (Mean) and standard deviation (SD). We look at the performance of original SVM. We can find that the original SVM has a very high PA (97.8%), but a very low NA (34.4%). It means that SVM has poor ability of identify minority negative comments in this imbalanced data. Considering GM, under-sampling technique can improve the performance to 76.8%. Compared to original SVM (55.2%) and over-sampling (34.1%), the under-sampling can has better performance.

TABLE III: Background of the Collected Data

	Original SVM		Over-sampling		Under-sampling	
	Mean	SD	Mean	SD	Mean	SD
PA	97.8%	2%	100%	0%	73.6%	13%
NP	34.4%	28%	15.0%	10%	82.5%	24%
GM	55.2%	33%	34.1%	20%	76.8%	14%
OA	94.4%	2%	93.9%	1%	74.2%	12%
F1	97.0%	1%	96.9%	0%	83.6%	9%

Fig. 2 summarizes the results of implementing cost adjusting method. We increase the misclassification cost from 1:1 to 1:11. "1:11" means we set the cost of misclassifying the minority examples (negative comments) into majority examples (positive comments) to be "11", when the cost of misclassifying the majority examples (positive comments) into minority examples (negative comments) is "1". The best GM is 52%. Fig. 3 summarizes the results of original SVM, under-sampling, over-sampling, and cost adjusting. We can find under-sampling can have the better performance.

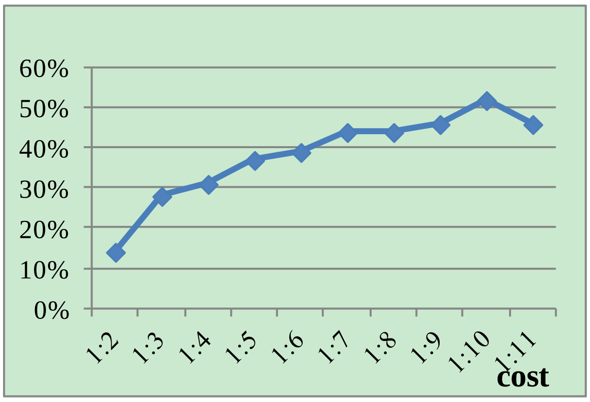

Fig. 2. The results of cost adjusting.

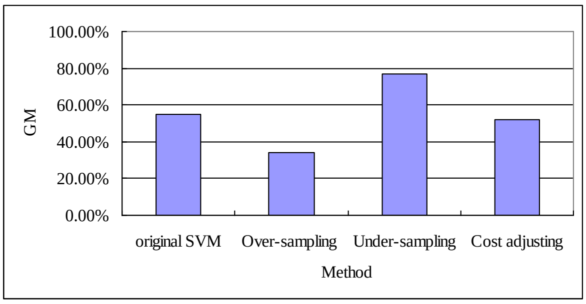

Fig. 3. Summary of all methods.

V. CONCLUSIONS

This study aims to improve the ability of classifying the minority negative comments, since these few negative comments will bring a great damage to enterprises. However, in available sentiment classification works, they didn't discuss this class imbalance problem. Therefore, we collect real comments regarding travel and hotel services. And, we attempted to compare these techniques including under-sampling, over-sampling, and cost adjusting. Experimental results indicated under-sampling has better performance. But, we merely used one data set retrieved from a real world travel websites. We might not get a general conclusion. Therefore, more data sets focus on different issues should be considered in the future works. In addition, even the best method, under-sampling, its performance is not good enough. Hence, other techniques for dealing with imbalanced data should be employed for further studies.

REFERENCES

[1] C. M. L. Cheung, M. K. O. Lee, and D. R. Thadani, "The impact of electronic word-of-mouth: The adoption of online opinions in online customer communities," *Internet Research*, vol. 18, no. 3, pp. 229-247, 2008.

[2] S. Li, G. Zhou, Z. Wang, S. Y. M. Lee, and R. Wang, "Imbalanced sentiment classification," in *Proc. the 20th ACM International Conference on Information and Knowledge Management*, pp. 2469-2472, 2011.

[3] Y. Liu, X. Yu, J. X. Huang, and A. An, "Combining integrated sampling with SVM ensembles for learning from imbalanced datasets," *Information Processing and Management*, vol. 47, pp. 617-631, 2011.

[4] V. García, J. S. Sánchez, and R. A. Mollineda, "On the effectiveness of preprocessing methods when dealing with different levels of class imbalance," *Knowledge-Based Systems*, vol. 25, no. 1, pp. 13-21, 2011.

[5] J. Yang, Y. Liu, X. Zhu, Z. Liu, and X. Zhang, "A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization," *Information Processing and Management*, vol. 48, pp. 741–754, 2012.

[6] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verla, 1995.

[7] P. Gupta and J. Harris, "How e-WOM recommendations influence product consideration and quality of choice: A motivation to process information perspective," *Journal of Business Research*, vol. 63, pp.1041–1049, 2010.

[8] C. W. Yoo, G. L. Sanders, and J. Moon, "Exploring the effect of e-WOM participation on e-Loyalty in e-commerce," *Decision Support Systems*, vol. 55, pp. 669–678, 2013.

[9] T. Verhagen, A. Nauta, and F. Feldberg, "Negative online word-of-mouth: Behavioral indicator or emotional release?" *Computers in Human Behavior*, vol. 29, pp.1430–1440, 2013.

[10] Z. Zhao, P. Zhong, and Y. Zhao, "Learning SVM with weighted maximum margin criterion for classification of imbalanced data," *Mathematical and Computer Modelling*, vol. 54, pp. 1093–1099, 2011.

[11] H. X. Zhang and M. Li, "RWO-Sampling: A random walk over-sampling approach to imbalanced data classification," *Information Fusion*, 2014.

[12] R. Moraes, J. F. Valiati, and W. P. G. Neto, "Docnment-level sentiment classification: An empirical comparison between SVM and ANN," *Expert Systems with Applications*, vol. 40, pp. 621–633, 2013.

[13] A. Sun, E.-P. Lim, and Y. Liu, "On strategies for imbalanced text classification using SVM: A comparative study," *Decision Support Systems*, vol. 48, pp. 191–201, 2009.

Long-Sheng Chen is an associate professor of the Department of Information Management, and dean of student affairs office, Chaoyang University of Technology, Taiwan. He received his Ph.D. in the Department of Industrial Engineering and Management, National Chiao Tung University, Taiwan in 2006, and his BS and MS degrees both in industrial management from National Cheng Kung University, Tainan, Taiwan in 1998 and 2000, respectively. His teaching and research interests include data mining, quality management, new service creation, and customer relationship management.

Sheng-Jhe Cai is a graduate student of the Department of Information Management, Chaoyang University of Technology, Taiwan. He received his BS degree in the Department of Information Management, Chienkuo Technology University, Changhua, in 2012. His research interests include text mining and sentiment classification.