Cluster-based Majority Under-Sampling Approaches for Class Imbalance Learning

Yan-Ping Zhang
School of Computer Science and Technology, Anhui University, Hefei, China
zhangyp@mail.hfah.cn

Li-Na Zhang
School of Computer Science and Technology, Anhui University, Hefei, China
zlnmsy@163.com

Yong-Cheng Wang
School of Computer Science and Technology, Anhui University, Hefei, China
luoxiying@foxmail.com

***Abstract*—The class imbalance problem usually occurs in real applications. The class imbalance is that the amount of one class may be much less than that of another in training set. Under-sampling is a very popular approach to deal with this problem. Under-sampling approach is very efficient, it only using a subset of the majority class. The drawback of under-sampling is that it throws away many potentially useful majority class examples. To overcome this drawback, we adopt an unsupervised learning technique for supervised learning.We proposes cluster-based majority under-sampling approaches for selecting a representative subset from the majority class. Compared to under-sampling, cluster-based under-sampling can effectively avoid the important information loss of majority class. We adopt two methods to select representative subset from k clusters with certain proportions, and then use the representative subset and the all minority class samples as training data to improve accuracy over minority and majority classes. In the paper, we compared the behaviors of our approaches with the traditional random under-sampling approach on ten UCI repository datasets using the following classifiers: k-nearest neighbor and Naïve Bayes classifier. Recall, Precision, F-measure, G-mean and BACC (balance accuracy) are used for evaluating performance of classifiers. Experimental results show that our cluster-based majority under-sampling approaches outperform the random under-sampling approach. Our approaches attain better overall performance on k-nearest neighbor classifier compared to Naïve Bayes classifier.**

Keywords-classification; clustering; under-sampling; class imbalance learning

I. INTRODUCTION

When one class samples severely outnumbers the other class samples in a dataset, the dataset is defined as an unbalanced dataset. The class imbalance problem usually occurs in real applications such as fraud detection [1, 2], customer requirements management and rare disease prediction in medical diagnosis [3, 4].Imbalance has a serious impact on the performance of classifiers.

Conventional learning algorithms do not take into account the imbalance of class. They give the same attention to the majority class and the minority class. When the imbalance level is huge, it is hard to build a good classifier for conventional learning algorithms. Conventional classification algorithms like neural networks, decision tree, Native Bayes and K-nearest neighbor assume that all classes have the same number of records in training data and the cost derived from all the classes is equal. Actually, the cost in miss-predicting minority classes is higher than that of the majority class for many class imbalance datasets. Therefore, if a classifier can make correct prediction on the minority class efficiently, it will be useful to solving many real applications and save a lot of cost. Predicting minority class data from unbalanced binary class data sets is an important problem in data mining and machine learning. Therefore, prediction of minority class samples and improvement of overall performance are very critical and are addressed in this paper.

Since the class imbalance problem exists in many real applications, researchers in data mining, machine learning and other related field have proposed several methods to solve this problem. These methods can fall into two broad categories: at the algorithmic level and at the data level. At the algorithmic level, developed methods mainly include cost-sensitive learning [5, 6, 7] and ensemble learning. At the data level, methods include multi-classifier committee and re-sampling approaches. Cost-sensitive learning gives different costs for the two classes and is considered as an important class of methods to handle class-imbalance [8]. A cost-sensitive classifier improves the accuracy of minority class by setting a high cost to the misclassification of a minority class sample. However, misclassification costs are often unknown and may result in over fitting training. For Ensemble learning, variants of AdaBoost are the most popular ones. Many cost-sensitive boosting algorithms have been proposed [9].These variants increase the weights of examples which have higher misclassification cost in the boosting process. Other algorithms raised the weight of high cost samples in every iteration of the boosting process [10, 11, 12].

Re-sampling is a class of approaches which aim to attain a balance dataset. Under-sampling and over-sampling are two popular approaches. One simple method of under-sampling is to select a subset of majority class samples randomly. Over-sampling replicates or generates the minority class samples to attain a balanced dataset. Over-sampling may causes longer training time and over-fitting. Drummond and Holte showed that random under-sampling yields better minority prediction than random over-sampling [13].However, they suggested random under-sampling leads to loss of loss of majority class information.

To deal with the problem of minority prediction and avoid data lose of majority class information in under-sampling in class imbalance problem, we adopts an

978-1-4244-6928-4/10/$26.00 ©2010 IEEE

unsupervised learning technique for supervised learning: K-means clustering. We propose cluster-based majority under-sampling approaches for selecting a representative subset from the majority class. Then we use the representative subset and the all samples of minority class as training data to improve prediction rates over minority and majority classes. Compared to under-sampling, cluster-based majority under-sampling can effectively avoid the important information loss of majority class.

Experimentally we compared the behaviors of our approaches with the traditional random under-sampling approach on ten UCI repository datasets using the following classifiers: k-nearest neighbor and Naïve Bayes classifier. Recall, Precision, F-measure, G-mean and BACC (balance accuracy) are used for evaluating performance of classifiers. Experimental results show that our cluster-based majority under-sampling approaches outperform the random under-sampling approach. Our approaches attain better overall performance on k-nearest neighbor classifier compared to Naïve Bayes classifier.

The rest of this paper is organized as follows. In section Ⅱ we present the related work. Section Ⅲ describes cluster-based majority under-sampling approaches. We conducted the experiments on 10 UCI repository [14] data sets. The experimental study used two different classifiers like k nearest neighbor and Naive Bayes. Section Ⅳ provides the results and discussion based on an experimental study. In section Ⅴ we conclude the paper.

II. RELATED WORK

Kubat et al [15] proposed one-sided selection approach. The approach divides majority class into four categories: noise, borderline, redundant, safe samples. They applied Condensed Nearest Neighbor (CNN) Rule for identifying the safe and redundant majority points in the dataset and eliminate the noise and borderline points. This method brings clear separation boundary between majority and minority samples, but it still suffers from the lack of data problem from the minority class samples.

Chawla et al [16] proposed SMOTE. SMOTE produces synthetic minority class samples by selecting some of the nearest minority neighbors of a minority sample which is named S, and generates new minority class samples along the lines between S and each nearest minority neighbor. SMOTE beats the random oversampling approaches by its informed properties, and reduce the imbalanced class distribution without causing over fitting. However, SMOTE blindly generate synthetic minority class samples without considering majority class samples, so it may cause over-generalization.

In 2003, Zhang and Mani [17] discussed four ways of selecting majority samples from the class distribution. They proposed three Near-miss methods and one distinct method. The Near-miss methods select the majority points from those that are close to all or some of the minority points. The distinct method selects the majority points whose average distance is farthest to three closest minority points. Finally, they compared their method with random under-sampling, and discussed the results of NearMiss_2 method by comparing with random under-sampling over nearest neighbor classifiers and C4.5. The experimental results showed that the NearMiss-2 method and random under-sampling method perform the best.

However, as the percentage of number of samples increases, near-miss method suffers from considerable dropping in recall and the distinct method suffers from over-generalization of the minority class.

H. Altmcay et al [18] proposed cluster based synthetic sample creation techniques to under-sample the majority class. They divide the majority class samples into N number of clusters, where N is the number of minority class samples in the dataset. These cluster centroids are used for balancing the training class distribution. Then they modified the weights of the Adaboost algorithm according to the average samples in each majority clusters for speaker verification dataset. Since they are using N centroids take place of original majority data, misclassification of centroids leads to much necessary information loss from the majority class.

Yue-Shi Lee et al [19] proposed cluster-based under-sampling technique for imbalanced class distribution. First, they partitioned the whole data into k number of clusters rather than the majority data. Based on imbalance ratio in each cluster, they selected some of the samples from each cluster for balancing the training class distribution by adopting random under-sampling technique [20]. Their experiments on synthetic and real world data sets proved that cluster-based random under-sampling perform better than the other under-sampling techniques. This method improves the prediction accuracy of majority samples but still there is some majority samples included important information loss.

III. CLUSTER-BASED MAJORITY UNDER-SAMPLING PREDICTION (CBMP) APPROACHES

As was shown by [20], under-sampling is an efficient strategy to deal with class-imbalance. However, the drawback of under-sampling is that it losts many potentially useful data. In this section, we propose two strategies to explore the majority class examples ignored by under-sampling. In order to achieve good prediction over minority class and avoid necessary information loss from the majority class, we use both K-means [21] algorithm and random sampling approach.

Assume the size of the class-imbalanced data set is N, which includes majority class samples (MA) and minority class samples (MI). The number of the samples is the size of the data set. The size of MA is represented as NMA. NMI is the number of minority class samples (MI). In the class-imbalanced data set, NMA is far larger than NMI. For our cluster-based majority under-sampling prediction (CBMP) approaches, we first divide all the majority class samples in the data set into k clusters. In the experiments, we will study the performances for the under-sampling methods on different number of clusters.

Let the number of majority class samples in the ith cluster ($1 \le i \le k$) be MAi. Therefore, the ratio of the

number of majority class samples to all the number of majority class samples in the ith cluster is $ri = {}^{MAi}/_{NMA}$, $1 \le i \le k$,

The number of selected majority class samples in the ith cluster is computed use (1)

$$si = NMI * ri \quad 1 \le i \le k \tag{1}$$

Eq. (1) determines that more majority class samples would be selected in ith cluster which has more majority class samples.

Algorithm:

Input: majority class samples, minority class samples.

Output: two balanced training set D1 and D2.

step1: Cluster all the majority class samples into k clusters using k-means clustering.

step2: Compute the number of selected majority class samples in each cluster by using (1), and then we adopt two methods to select si majority samples in ith cluster. The first method is select si majority class samples in ith cluster randomly. The second method is select si majority class samples which nearest to the ith centroid in ith cluster. Finally, we get two majority sample subsets C1 and C2.

step3: Combine C1, C2 with all the minority class samples respectively to obtain the training set D1, D2.

IV. EXPERIMENTS

A. *Evaluation Criteria of Classification Performance*

It is now well-known that accuracy or error rate is an appropriate evaluation criterion for conventional classification. However, it is not an appropriate evaluation criterion when there are class-imbalance or unequal costs. In this paper, we use Recall, Precision, F-measure, G-mean and BACC (balanced accuracy) as performance's evaluation measures. F-measure and G-mean are functions of the confusion matrix as shown in Table I. Recall, Precision, F-measure, G-mean and BACC are then defined as follows. Here, we take minority class as positive class.

$$\text{Re } call = {}^{TP}/_{(TP + FN)}$$

$$\text{Pr } ecision = TP/(TP + FP)$$

$$G - mean = \sqrt{\frac{TP}{TP + FN} * \frac{TN}{TN + FP}}$$

$$F - measure = \frac{2 * \text{Re } call * \text{Pr } ecision}{\text{Re } call + \text{Pr } ecision}$$

$$BACC = \frac{\text{Re } call + {}^{TP}/_{(TN + FP)}}{2}$$

TABLE I. CONFUSION MATRIX

Class Name	***Predicted Positive Class***	***Predicted Negative Class***
Actual Positive Class	TP(True Positives)	FN(True Positive)
ActualNegative Class	FP(False Positives)	TN(True Negatives)

In our experiments, we use these five criteria to evaluate the classification performance of the approaches.Recall measures the predicted accuracy of the positive samples (minority samples). Precision gives the trade-off between the predicted accuracy of the negative samples ((majority samples) and TP. Generally, for a classifier, if the precision rate is high, then the recall rate will be low, that is, the two criteria are trade-off. If both precision, recall are larger then F-Measure is also larger. For unbalanced data sets, generally, the higher the recall is, the lower the precision is. So increasing recall rates without decreasing the precision of the minority class is a challenging problem. F-Measure is a popular measure for unbalanced data classification problems [19]. F-Measure depicts the trade-off between precision and recall. Barandela et al. introduced the metric called the geometric mean (GM) [22].This measure allows Barandela et al. to simultaneously maximize the accuracy in positive and negative examples with a favorable trade-off. So we use G-mean to maximize accuracy of majority samples and Recall with a favorable trade-off.

B. *Experimental setting*

We tested our proposed approaches on 10 UCI data sets [14]. Information about these data sets is summarized in Table II .Here, for each data set, Number of example, Number of attributes, Name of minority class and Ratio are depicted.

TABLE II. DATASET DESCRIPTIONS

Dataset name	***Number of example***	***Number of attributes***	***Name of minority class***	***Ratio***
abalone	4177	8	Ring=7	9.7
balance	625	4	Balance	11.8
breast-cancer-w	699	9	Malignant	1.9
haberman	306	3	Class 2	2.8
hepatitis	155	19	DIE	3.8
housing	506	13	[20,23]	3.8
ionosphere	351	33	bad	1.8
wdbc	569	30	malignant	1.7
wpbc	198	33	recur	3.2
spambase	4601	58	1	1.5

In our experiment, we use two kinds of classifiers: k-nearest neighbor (k=1) and Naive Bayes. We compare our approaches with random under-sampling and the approach of adopting all original sets.

CBMP1 approach is a classifier learned by random under-sampling on each cluster of the majority class. CBMP2 approach is a classifier learned by select some samples which nearest to the centroid of each cluster. ALL

approach is a classifier learned on original data set. RUS is a classifier learned by random under-sampling of the majority class. For every data set, we perform a 5-fold cross validation. Within each fold, the classification method is repeated 10 times considering the randomness of the sampling in each cluster. The whole cross validation process is repeated for 10 times, the final values of Recall, Precision, F-measure, G-mean and BACC are the averages of these 10 times.

C. *Results and Analyses*

Table Ⅲ shows the Recall of the four approaches using Naive Bayes classifier on ten different data sets respectively.

As shown in Fig.1, CBMP1 and CBMP2 approaches attain high recall, good precision, considerable F-measure, high G-mean and high BACC. For Naive Bayes classifier, the CBMP1 approach attains 69.18% recall, 43% precision, 52% F-measure, 68% G-mean and 0.69 % BACC rates on ten data sets. CBMP2 approach attains 69.84% recall, 43% precision, 50% F-measure, 66% G-mean and 0.67% BACC rates on ten data sets. However, RUS approach attains 65.01% recall, 44% precision, 48% F-measure, 64% G-mean and 66 % BACC rates on ten data sets. So the CBMP1 approach is better than other approaches on Naive Bayes classifier.

Table Ⅳ shows the average behavior of the four approaches (CBMP1, CBMP2, ALL and RUS) using k-nearest neighbor classifier (k=1) on ten different data sets. Fig.2 depicts the average values of Recall, Precision, F-measure, G-mean and BACC of four approaches on ten imbalance data sets using k-nearest neighbor (k=1) classifier. From Fig. 2 and table 4, the CBMP1and CBMP2 approaches have got better performance (5%~7% Recall, 2% ~ 9% G-mean, 2% ~ 5% F1-measure and 1% ~ 3% BACC, respectively) than ALL and RUS. Comparing Naive Bayes classifier with k-nearest neighbor classifier (k=1), the overall performance of k-nearest neighbor classifier (k=1) is better. Compared CBMP1 with CBMP2 approach, we can see that CBMP1 is better and more stable than CBMP2 on ten data sets.

TABLE III. RECALL OF COMPARED APPROACHES ON TEN IMBALANCE DATA SETS USING NAIVE BAYES CLASSIFIER. THE ROW AVG SHOW THE AVERAGE RECALL OF EACH APPROACH ON TEN DATA SETS.

Recall	***ALL***	***RUS***	***CBMP1***	***CBMP2***
abalone	0.8289	0.8447	0.8553	0.8289
balance	0	0.55	0.625	0.6625
breast-cancer-w	0.9795	0.9795	0.9816	1
haberman	0.2222	0.2667	0.3889	0.3889
hepatitis	0.6129	0.5387	0.5161	0.5161
housing	0.8182	0.8000	0.8136	0.8181
ionosphere	0.5789	0.4737	0.4842	0.5211
wdbc	0.9211	0.9237	0.9473	0.9474
wpbc	0.1111	0.1556	0.3333	0.3333
spambase	0.9671	0.9682	0.9726	0.9671
AVG	0.6040	0.6501	0.6918	0.6984

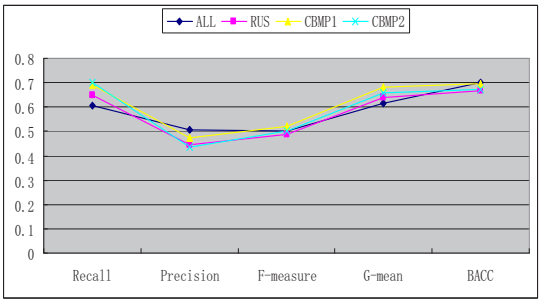

Figure 1. The average values of Recall, Precision, F-measure, G-mean and BACC of four approaches on ten imbalance data sets using Naive Bayes classifier.

TABLE IV. THE AVERAGE VALUES OF RECALL, PRECISION, F-MEASURE, G-MEAN AND BACC OF FOUR APPROACHES ON TEN IMBALANCE DATA SETS USING K-NEAREST NEIGHBOR (K=1) CLASSIFIER.

Evaluation criteria	***ALL***	***RUS***	***CBMP1***	***CBMP2***
Recall	0.5734	0.6934	0.7449	0.7417
Precision	0.580647	0.55268	0.5553	0.5416
F-measure	0.5688	0.5919	0.6123	0.5946
G-mean	0.6687	0.7256	0.7545	0.7410
BACC	0.7342	0.736	0.7604	0.7445

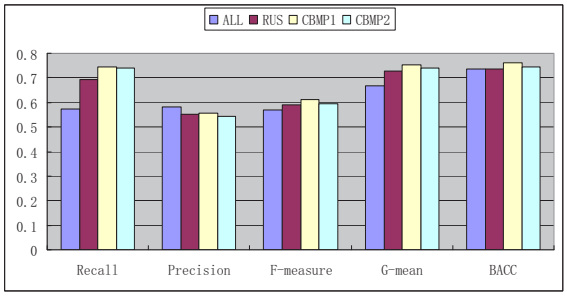

Figure 2. The average values of Recall, Precision, F-measure, G-mean and BACC of four approaches on ten imbalance data sets using k-nearest neighbor (k=1) classifier.

V. CONCLUTION

Many class imbalance problems exist in real world applications such as rarely-seen disease investigation, credit card fraud detection, and internet intrusion detection. Resampling are popular and simple approaches compared with other approaches to solve the problem. Under-sampling approach is a conventional and simple approach in resampling approaches, but it suffers from some important information loss of majority samples. In order to solve this problem, we propose cluster based majority under-sampling approaches for selecting the representative majority class samples. K-means clustering algorithm is adopted for majority class samples. First, we cluster all majority class

samples into k and adopt two methods to select the representative samples in each cluster. Then we use the representative subset and the all samples of minority class as training set to improve prediction rates over minority and majority classes. Compared to under-sampling, cluster-based under-sampling can effectively avoid the important information loss of majority class. We tested our approaches on k-nearest neighbor classifier, Naïve Bayes classifier on 10 UCI imbalanced data sets. Compared with under-sampling approach, experimental results show that the proposed approaches attain good Precision, high Recall, good F-measure, good G-mean and BACC rates for the ten imbalanced datasets. K-nearest neighbor classifier gave better prediction compared to Naïve Bayes classifier.

ACKNOWLEDGEMENT

This research was supported by the National Natural Science Foundation of China (Grant Numbers 60675031), National 973 Key Project of China (2007CB311003).

REFERENCES

[1] C. Phua, A. Damminda, and V. Lee, "Minority Report in Fraud Detection: Classification of Skewed Data," ACM Sigkdd Explorations: Special Issue on Imbalanced Data Sets, 6(1), 2004, pp. 50-59.

[2] P. Chan, W. Fan, A. Prodromidis, and S. Stolfo, "Distributed Data Mining in Credit Card Fraud Detection," IEEE Intelligent Systems, Vol. 14, 1999, pp. 67-74

[3] K. Yoon, S. Kwek, "A data reduction approach for resolving the imbalanced data issue in functional genomics," Neural Computing and Applications, 2007, pp. 295-306.

[4] J. X. Chen, T. H. Cheng, A. L. F. Chan, and H. Y. Wang, "An application of classification analysis for skewed class distribution in therapeutic drug monitoring-the case of vancomycin," Workshop on Medical Information Systems (IDEAS-DH'04).IEEE. Beijing, China, 2004, pp.35-39.

[5] C. Drummond, and R. C. Holte, "C4.5, Class Imbalance, and Cost Sensitivity: Why Under-sampling beats Over-sampling," In Works hop on Learning from Imbalanced Data Sets II, 2003.

[6] Elkan. C, "The foundations of cost-sensitive learning," In Proceedings of the 17th international joint conference on artificial intelligence, 2001, pp. 973–978.

[7] Turney. P, "Types of cost in inductive concept learning," In Proceedings of the ICML' 2000 workshop on cost-sensitive learning, 2000, pp. 15–21.

[8] G. M. Weiss, "Mining with rarity: A unifying framework," ACM SIGKDD Explorations, vol. 6, no.1,2004, pp. 7–19.

[9] K. M. Ting, "An empirical study of MetaCost using boosting algorithms," in Proceedings of the 11th European Conference on Machine Learning, Barcelona, Spain, 2000, pp. 413–425.

[10] P. Viola, and M. Jones, "Fast and robust classification using asymmetric AdaBoost and a detector cascade," in Advances in Neural Information Processing Systems 14, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. Cambridge, MA: MIT Press, 2002, pp. 1311–1318.

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

[12] G. J. Karakoulas, and J. Shawe-Taylor, "Optimizing classifiers for imbalanced training sets," in Advances in Neural Information Processing Systems11.Cambridge, MA: MIT Press, 1999, pp. 253–259

[13] C. Drummond, and R. C. Holte, "C4.5, Class Imbalance, and Cost Sensitivity: Why Under-sampling beats Over-sampling," In Workshop on Learning from Imbalanced Data Sets II, 2003.

[14] http://archive.ics.uci.edu/ml/

[15] M. Kubat, and S. Matwin, "Addressing the Curse of Imbalanced Data Sets:One-Sided Sampling," Proceedings of the Fourteenth International Conference on Machine Learning, 1997, pp. 79-186.

[16] Chawla. N. V, Bowyer. K. W, Hall. L. O, and Kegelmeyer.W. P, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, 2002, pp. 321–357.

[17] J. Zhang, and I. Mani, "knn approach to unbalanced data distributions: A case study involving information extraction," In Proc. of the ICML-2003 Workshop: Learning with Imbalanced Data Sets II, pp. 42–48.

[18] H. Altinçay, and C. Ergün, "Clustering based under-sampling for improving speaker verification decisions using AdaBoost," Joint IAPR International Workshops on Structural, Syntactic, and Statistical Pattern Recognition. Lecture Notes in Computer Science 2004, pp. 698-706.

[19] Show-Jane. Yen, and Yue-Shi. Lee, "Under-Sampling Approaches for Improving Prediction of the Minority Class in an Imbalanced Dataset," In Proceedings of the Intelligent Control and Automation, Lecture Notes in Control and Information Sciences (LNCIS), Vol. 344, August 2006, pp. 731-740.

[20] C. Ling, and C. Li, "Data Mining for Direct Marketing Problems and Solutions," In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), AAAI Press, New York, 1998, pp. 73-79.

[21] P. S. Bradley, and U. M. Fayyad, "Refining initial points for K-means clustering," Proceedings of the Fifteenth International Conference on Machine Learning (ICML98), 1998, pp. 91-99.

[22] Barandela. R, Sanchez. J, S. Garc´ıa. V, and Rangel. E, "Strategies for learning in class imbalance problems," Pattern Recognition, 2003, pp. 849–851.