Cluster-Based Sampling Approaches to Imbalanced Data Distributions

Show-Jane Yen and Yue-Shi Lee

Department of Computer Science and Information Engineering, Ming Chuan University
5 The-Ming Rd., Gwei Shan District, Taoyuan County 333, Taiwan
{sjyen, leeys}@mcu.edu.tw

Abstract. For classification problem, the training data will significantly influence the classification accuracy. When the data set is highly unbalanced, classification algorithms tend to degenerate by assigning all cases to the most common outcome. Hence, it is important to select the suitable training data for classification in the imbalanced class distribution problem. In this paper, we propose cluster-based under-sampling approaches for selecting the representative data as training data to improve the classification accuracy in the imbalanced class distribution environment. The basic classification algorithm of neural network model is considered. The experimental results show that our cluster-based under-sampling approaches outperform the other under-sampling techniques in the previous studies.

1 Introduction

The classification techniques usually assume that the training samples are uniformly-distributed between different classes. A classifier performs well when the classification technique is applied to a dataset evenly distributed among different classes. However, many datasets in real applications involve imbalanced class distribution problem [5, 7]. The imbalanced class distribution problem occurs while there are much more samples in one class than the other class in a training dataset. In an imbalanced dataset, the *majority class* has a large percent of all the samples, while the samples in *minority class* just occupy a small part of all the samples. In this case, a classifier usually tends to predict that samples have the majority class and completely ignore the minority class.

One simple method of under-sampling is to select a subset of MA randomly and then combine them with MI as a training set, which is called *random under-sampling approach*. Several advanced researches are proposed to make the selective samples more representative. The under-sampling approach based on distance [7] uses distinct modes: the nearest, the farthest, the average nearest, and the average farthest distances between MI and MA, as four standards to select the representative samples from MA. For every minority class sample in the dataset, the first method "nearest" calculates the distances between all majority class samples and the minority class samples, and selects k majority class samples which have the smallest distances to the minority

A Min Tjoa and J. Trujillo (Eds.): DaWaK 2006, LNCS 4081, pp. 427 – 436, 2006.

class sample. If there are n minority class samples in the dataset, the "nearest" approach would finally select $k \times n$ majority class samples ($k \geq 1$). However, some samples within the selected majority class samples might duplicate.

Similar to the "nearest" approach, the "farthest" approach selects the majority class samples which have the farthest distances to each minority class samples. For every majority class samples in the dataset, the third method "average nearest" calculates the average distance between one majority class sample and all minority class samples. This approach selects the majority class samples which have the smallest average distances. The last method "average farthest" is similar to the "average nearest" approach; it selects the majority class samples which have the farthest average distances with all the minority class samples. The above under-sampling approaches based on distance in [7] spend a lot of time selecting the majority class samples in the large dataset, and they are not efficient in real applications.

In 2003, J. Zhang and I. Mani [6] presented the compared results within four informed under-sampling approaches and random under-sampling approach. The first method "*NearMiss-1*" selects the majority class samples which are close to some minority class samples. In this method, majority class samples are selected while their average distances to three closest minority class samples are the smallest. The second method "*NearMiss-2*" selects the majority class samples while their average distances to three farthest minority class samples are the smallest. The third method "*NearMiss-3*" take out a given number of the closest majority class samples for each minority class sample. Finally, the fourth method "*Most distant*" selects the majority class samples whose average distances to the three closest minority class samples are the largest. The final experimental results in [6] showed that the *NearMiss-2* approach and random under-sampling approach perform the best.

In this paper, we study the effects of under-sampling [1, 3, 6] on the backpropagation neural network technique and propose some new under-sampling approaches based on clustering, such that the influence of imbalanced class distribution can be decreased and the accuracy of predicting the minority class can be increased.

2 Our Approaches

In this section, we present our approach *SBC* (under-Sampling Based on Clustering) which focuses on the under-sampling approach and uses clustering techniques to solve the imbalanced class distribution problem. Our approach first clusters all the training samples into some clusters. The main idea is that there are different clusters in a dataset, and each cluster seems to have distinct characteristics. If a cluster has more majority class samples and less minority class samples, it will behave like the majority class samples. On the opposite, if a cluster has more minority class samples and less majority class samples, it doesn't hold the characteristics of the majority class samples and behaves more like the minority class samples. Therefore, our approach *SBC* selects a suitable number of majority class samples from each cluster by considering the ratio of the number of majority class samples to the number of minority class samples in the cluster.

2.1 Under-Sampling Based on Clustering

Assume that the number of samples in the class-imbalanced dataset is N, which includes majority class samples (MA) and minority class samples (MI). The size of the dataset is the number of the samples in this dataset. The size of MA is represented as $Size_{MA}$, and $Size_{MI}$ is the number of samples in MI. In the class-imbalanced dataset, $Size_{MA}$ is far larger than $Size_{MI}$. For our under-sampling method *SBC*, we first cluster all samples in the dataset into K clusters. The number of majority class samples and the number of minority class samples in the ith cluster ($1 \leqq i \leqq K$) are $Size^i_{MA}$ and $Size^i_{MI}$, respectively. Therefore, the ratio of the number of majority class samples to the number of minority class samples in the ith cluster is $Size^i_{MA} / Size^i_{MI}$. If the ratio of $Size_{MA}$ to $Size_{MI}$ in the training dataset is set to be m:1, the number of selected majority class samples in the ith cluster is shown in expression (1):

$$SSize^i_{MA} = (m \times SizeMI) \times \frac{Size^i_{MA} \Big/ Size^i_{MI}}{\sum_{i=1}^{K} Size^i_{MA} \Big/ Size^i_{MI}} \tag{1}$$

In expression (1), $m \times SizeMI$ is the total number of selected majority class samples that we suppose to have in the final training dataset. $\sum_{i=1}^{K} Size^i_{MA} \Big/ Size^i_{MI}$ is the total ratio of the number of majority class samples to the number of minority class samples in all clusters. Expression (1) determines that more majority class samples would be selected in the cluster which behaves more like the majority class samples. In other words, $SSize^i_{MA}$ is larger while the ith cluster has more majority class samples and less minority class samples. After determining the number of majority class samples which are selected in the ith cluster, $1 \leqq i \leqq K$, by using expression (1), we randomly choose majority class samples in the ith cluster. The total number of selected majority class samples is $m \times Size_{MI}$ after merging all the selected majority class samples in each cluster. At last, we combine the whole minority class samples with the selected majority class samples to construct a new training dataset. Table 1 shows the steps for our under-sampling approach.

For example, assume that an imbalanced class distribution dataset has totally 1100 samples. The size of MA is 1000 and the size of MI is 100. In this example, we cluster this dataset into three clusters. Table 2 shows the number of majority class samples $Size^i_{MA}$, the number of minority class samples $Size^i_{MI}$, and the ratio of $Size^i_{MA}$ to $Size^i_{MI}$ for the ith cluster.

Assume that the ratio of $Size_{MA}$ to $Size_{MI}$ in the training data is set to be 1:1, in other words, there are 100 selected majority class samples and the whole 100 minority class samples in this training dataset. The number of selected majority class samples in each cluster can be calculated by expression (1). Table 3 shows the number of selected

Table 1. The structure of the under-sampling based on clustering approach *SBC*

Step1.	Determine the ratio of $Size_{MA}$ to $Size_{MI}$ in the training dataset.
Step2.	Cluster all the samples in the dataset into some clusters.
Step3.	Determine the number of selected majority class samples in each cluster by using expression (1), and then randomly select the majority class samples in each cluster.
Step4.	Combine the selected majority class samples and all the minority class samples to obtain the training dataset.

Table 2. Cluster descriptions

Cluster ID	Number of majority class samples	Number of minority class samples	$Size^i_{MA} / Size^i_{MI}$
1	500	10	500/10=50
2	300	50	300/50=6
3	200	40	200/40=5

Table 3. The number of selected majority class samples in each cluster

Cluster ID	The number of selected majority class samples
1	1×100× 50 / (50+6+5) =82
2	1×100× 6 / (50+6+5) = 10
3	1×100× 5 / (50+6+5)= 8

majority class samples in each cluster. We finally select the majority samples ran domly from each cluster and combine them with the minority samples to form the new dataset.

2.2 Under-Sampling Based on Clustering and Distances

In *SBC* method, all the samples are clustered into several clusters and the number of selected majority class samples is determined by expression (1). Finally, the majority class samples are randomly selected from each cluster. In this section, we propose other two under-sampling methods, which are based on *SBC* approach. The difference between the two proposed under-sampling methods and *SBC* method is the way to select the majority class samples from each cluster. For the two proposed methods, the majority class samples are selected according to the distances between the majority class samples and the minority class samples in each cluster. Hence, the distances between samples will be computed.

For a continuous attribute, the values of all samples for this attribute need to be normalized in order to avoid the effect of different scales for different attributes. For example, suppose A is a continuous attribute. In order to normalize the values of attribute A for all the samples, we first find the maximum value Max_A and the minimum value Min_A of A for all samples. To lie an attribute value a_i in between 0 to 1, a_i is normalized to $\frac{a_i - Min_A}{Max_A - Min_A}$. For a categorical or discrete attribute, the distance between

two attribute values x_1 and x_2 is 1 (i.e. x_1-x_2=1) while x_1 is not equal to x_2, and the distance is 0 (i.e. x_1-x_2=0) while they are the same.

Assume that there are N attributes in a dataset and V_i^X represents the value of attribute A_i in sample X, for $1 \leqq i \leqq N$. The Euclidean distance between two samples X and Y is shown in expression (2).

$$\text{distance}(X,Y) = \sqrt{\sum_{i=1}^{N}(V_i^X - V_i^Y)^2} \tag{2}$$

The two approaches we proposed in this section first cluster all samples into K ($K \geq 1$) clusters as well, and determine the number of selected majority class samples for each cluster by expression (1). For each cluster, the representative majority class samples are selected in different ways. The first method *SBCMD* (Sampling Based on Clustering with *Most Distant*) selects the majority class samples whose average distances to M closest minority class samples in the ith cluster are the farthest. The second method, which is called *SBCMF* (Sampling Based on Clustering with Most Far), selects the majority class samples whose average distances to all minority class samples in the cluster are the farthest.

3 Experimental Results

For our experiments, we use three criteria to evaluate the classification accuracy for minority class: the precision rate P, the recall rate R, and the F-measure for minority class. 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. We cannot use one of the two criteria to evaluate the performance of a classifier. Hence, the precision rate and recall rate are combined to form another criterion F-measure, which is shown in expression (3).

$$\text{MI's F-measure} = \frac{2 \times \text{P} \times \text{R}}{\text{P} + \text{R}} \tag{3}$$

In the following, we expression (3) to evaluate the performance of our approaches *SBC*, *SBCMD*, and *SBCMF* by comparing our methods with the other methods *AT*, *RT*, and *NearMiss-2* on synthetic datasets. The method *AT* uses all samples to train the classifiers and does not select samples. *RT* is the most common-used random under-sampling approach and it selects the majority class samples randomly. The last method *NearMiss-2* is proposed by J. Zhang and I. Mani [6], which has been discussed in section 1. The two methods *RT* and *NearMiss-2* have the better performance than the other proposed methods in [6]. In the following experiments, the classifiers are constructed by using the artificial neural network technique in *IBM Intelligent Miner for Data V8.1*.

For each generated synthetic dataset, the number of samples is set to 10000, the number of numerical attributes and categorical attributes are set to 5, respectively. The dataset DSi means that the dataset potentially can be separated into i clusters, and our methods also cluster the dataset DSi into i clusters. Figure 1 shows the

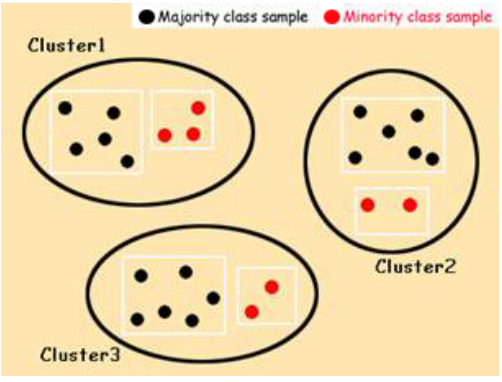

Fig. 1. The distribution of samples in a dataset

distribution of samples in a dataset which has three clusters inside. Moreover, in order to make the synthetic datasets more like real datasets, the noisy data are necessary.

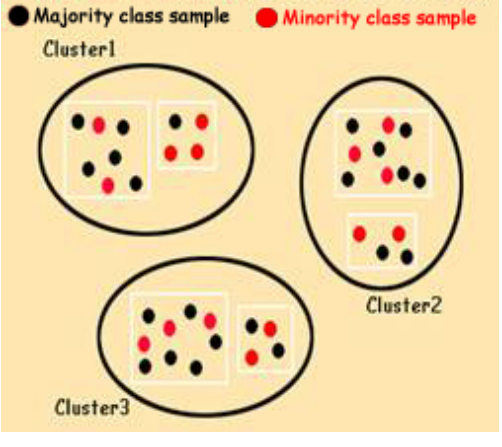

Fig. 2. Example for disordered samples

The synthetic datasets have two kinds of noisy data: disordered samples and exceptional samples. The disordered samples mean that some majority class samples (or minority class samples) lie to the area of minority class samples (or majority class samples). The disordered samples are illustrated with Figure 2. As for exceptional samples, they distribute irregularly in a dataset and outside the clusters. The samples outside the clusters in Figure 3 are exceptional samples. A dataset DSi with j% exceptional samples and k% disordered samples is represented as DSiEjDk. If there is no disordered sample in the synthetic dataset, the dataset is represented as DSiEjDN.

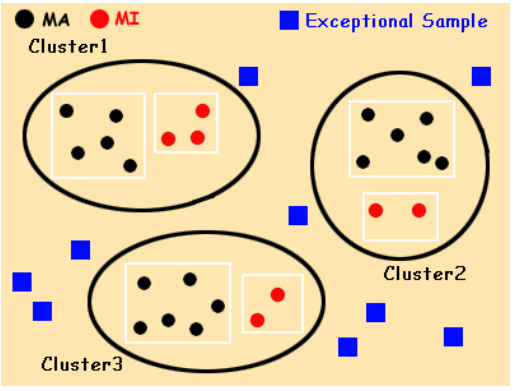

Fig. 3. Example for exceptional samples

Figure 4 shows the experimental results in the datasets in which the ratios of the number of majority class samples to the number of minority class samples are 2:1, 4:1, 9:1, 18:1, 36:1, and 72:1, respectively. For each specific ratio, we generate several synthetic datasets DSiE10D20 in which i is from 2 to 16. Hence, the average MI's F-measures are computed from all the datasets for each specific ratio. In Figure 4, we can see that the average MI's F-measure for *SBC* is higher than the other methods in most cases.

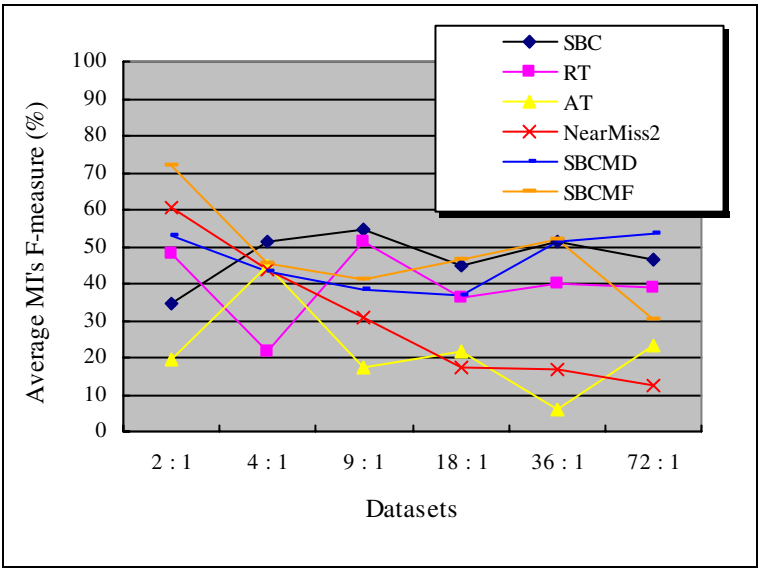

Fig. 4. Average MI's F-measure for datasets DSiE10D20

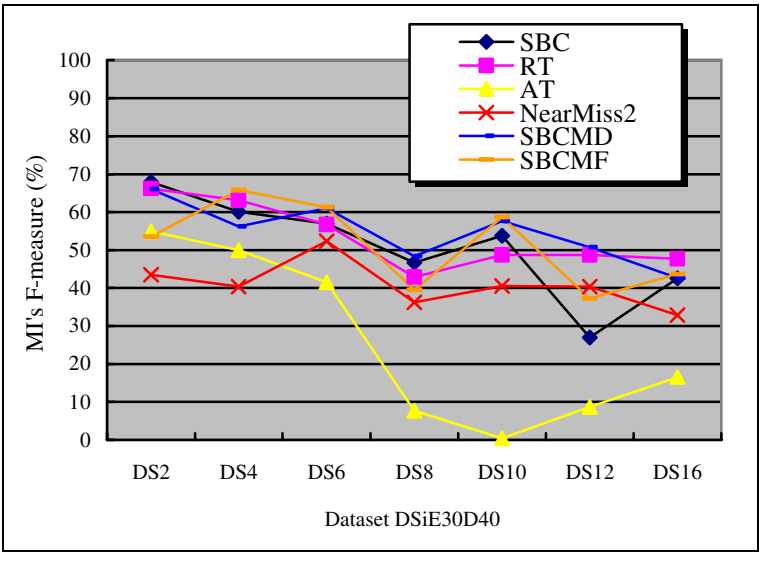

Fig. 5. MI's F-measure for each method on the datasets with 30% exceptional samples and 40% disordered samples

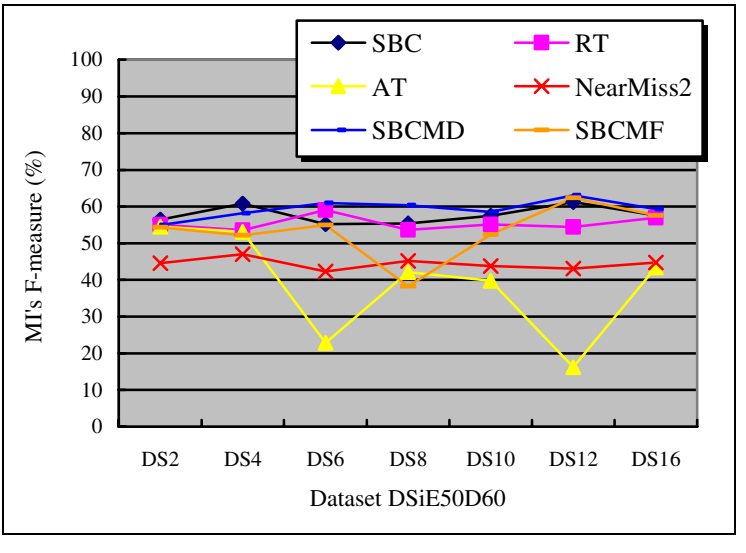

Fig. 6. MI's F-measure for each method on the datasets with 50% exceptional samples and 60% disordered samples

We raise the percentage of exceptional samples and disordered samples to 30% and 40%, respectively. And then we continue to raise the percentage of exceptional samples and disordered samples to 50% and 60%, respectively. Figure 5 and Figure 6

show the experimental results in DSiE30D40 and DSiE50D60, respectively, in which i is from 2 to 16. The experimental results show that *SBCMD* is the most stable method and has high MI's F-measure in each synthetic dataset. *RT* is also a stable method in the experiments, but the performance for *SBCMD* is better than *RT* in most cases. Although the MI's F-measure for *SBCMF* is higher than the other methods in some cases, the performance for *SBCMF* is not stable. Hence, the performance for *SBCMD* is the best in most of the cases when the datasets contain more exceptional samples and disordered samples, and *SBC* is stable and performs well in any case.

The average execution time for each method is shown in Figure 7. The execution time includes the time for executing the under-sampling method and the time for training the classifiers. According to the results in Figure 7, both *SBC* and *RT* are most efficient among all the methods, and *NearMiss-2* spends too much time for selecting the majority class samples.

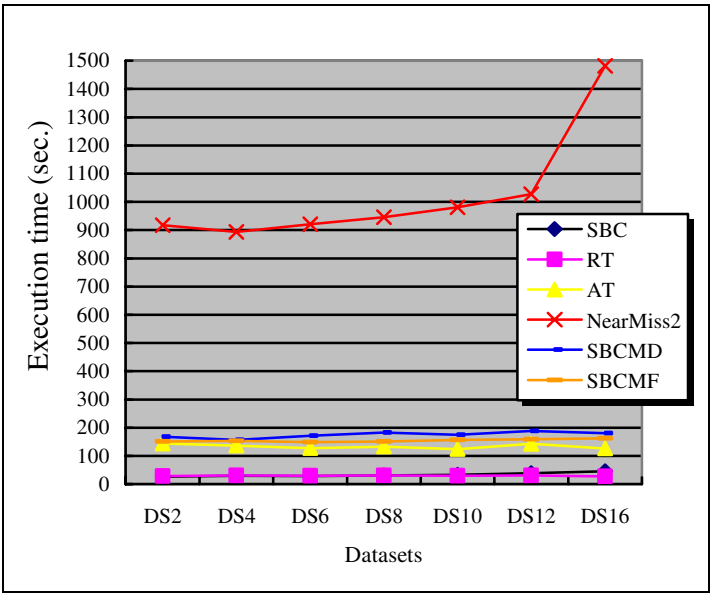

Fig. 7. Average execution time for each method

4 Conclusion

In a classification task, the effect of imbalanced class distribution problem is often ignored. Many studies [2, 4] focused on improving the classification accuracy but did not consider the imbalanced class distribution problem. Hence, the classifiers which are constructed by these studies lose the ability to correctly predict the correct decision class for the minority class samples in the datasets which the number of majority class samples are much greater than the number of minority class samples. Many real applications, like rarely-seen disease investigation, credit card fraud detection, and

internet intrusion detection always involve the imbalanced class distribution problem. It is hard to make right predictions on the customers or patients who that we are interested in.

In this study, we propose cluster-based under-sampling approaches to solve the imbalanced class distribution problem by using backpropagation neural network. The other two under-sampling methods, Random selection and *NearMiss-2*, are used to be compared with our approaches in our performance studies. In the experiments, our approach *SBC* has better prediction accuracy and stability than other methods. *SBC* not only has high classification accuracy on predicting the minority class samples but also has fast execution time. *SBCMD* has better prediction accuracy and stability when the datasets contain more exceptional samples and disordered samples. However, *SBCMF* does not have stable performances in our experiments. The two methods take more time than *SBC* on selecting the majority class samples as well.

References

1. Chawla, N. V.: C4.5 and Imbalanced Datasets: Investigating the Effect of Sampling Method, Probabilistic Estimate, and Decision Tree Structure. Proceedings of the ICML'03 Workshop on Class Imbalances (2003).
2. Caragea, D., Cook, D., Honavar, V.: Gaining Insights into Support Vector Machine Pattern Classifiers Using Projection-Based Tour Methods. Proceedings of the KDD Conference, San Francisco, CA (2001) 251-256.
3. Drummond, C., Holte, R. C.: C4.5, Class Imbalance, and Cost Sensitivity: Why Under-Sampling Beats Over-Sampling. Proceedings of the ICML'03 Workshop on Learning from Imbalanced Datasets (2003).
4. del-Hoyo, R., Buldain, D., Marco, A.: Supervised Classification with Associative SOM. Lecture Notes in Computer Science, Vol.2686 (2003) 334–341.
5. Japkowicz, N.: Concept-learning in the Presence of Between-class and Within-class Imbalances. Proceedings of the Fourteenth Conference of the Canadian Society for Computational Studies of Intelligence (2001) 67–77.
6. Zhang, J., Mani, I.: kNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Datasets (2003).
7. Chyi, Y.M.: Classification Analysis Techniques for Skewed Class Distribution Problems, Master Thesis, Department of Information Management, National Sun Yat-Sen University (2003).