

Multi-Cluster Based Approach for skewed Data in Data Mining

¹Mr.Rushi Longadge, ² Ms. Snehlata S. Dongre, ³Dr. Latesh Malik

1,2,3 Department of Computer Science and Engineering G. H. Rasoni College of Engineering

Abstract: In data mining difficulties are encountered when applying machine learning techniques to real-world data, which frequently show skewness properties. A typical example from industry where skewed data is an intrinsic problem is fraud detection in finance data, medical diagnosis on rare disease, finding network intrusion in network. This problem is also known as class imbalance problem. The class imbalance problem defines as the sample of one class may be much less number than another class in data set. There are many technologies developed for handling class imbalance. Basically designed approaches are divided into two types. First is designed a new algorithm which improves the minority class prediction, second modify the number samples in existing class, it is also known as data pre-processing. Under-sampling is a very popular data pre-processing approach to deal with class imbalance problem. Under-sampling approach is very efficient, it only uses the subset of the majority class. The drawback of under-sampling is that it removes away many useful majority class samples. To solve this problem we propose multi cluster-based majority under-sampling and random minority oversampling approach. Compared to under-sampling, cluster-based random under-sampling can effectively avoid the important information loss of majority class.

Keyword: Skewed data, Random under-sampling, class Imbalance problem, clustering, imbalance dataset.

I. Introduction

Some real-life data mining problems involve learning classifiers from imbalanced data, which means that one of the classes called a minority class includes much smaller number of examples than the other classes called as majority classes. Typical such problems are medical diagnosing dangerous illness, analyzing financial risk, detecting oil spills in satellite images, predicting technical equipment failures or information filtering [1], [2]. Class imbalance constitutes a difficulty for most learning algorithms, which are biased toward learning and prediction of the majority classes. As a result, minority examples tend to be misclassified.

The information in the minority class is very less as compared to the majority class samples. It is easy to be overlapped by the information of the majority and lead to misclassification [3]. As a result, the performance of the classifier based on balanced data sets is far better than that based on the imbalanced ones.

Therefore, the traditional classification approaches and their evaluation criteria are not suitable for the imbalanced data set. The achievements in classifiers based on imbalanced data sets have been presented with different approaches. Sampling method based on the preprocessing of data, which reconstructs the data set artificially to reduce the degree of imbalance. Over-sampling is to increase the number of the minority, but it may lead to over-fitting because of the duplication of data. While, under-sampling [4] is to cut down the number of the majority class samples. But it may lose information of the majority and decrease the performance of classification. The other method focuses on the algorithm based approach, which introduces certain mechanisms to handle the imbalance and make it suitable for the classification on imbalanced data sets. Examples of such techniques are: cost sensitive, support vector machines algorithm (SVM), and some ensemble methods.

There are many mechanisms in revising algorithms for imbalanced data mining. For example, the use of adjustment of cost function, the use of different values of weight, the change of probability density. Cost sensitive study algorithm uses the cost of each class to make classification decision. Its target is to cut down the overall cost instead of reducing the error rate as much as possible [5]. Support vector machine has been modified to process imbalanced data sets. One simple modification is to make a moderate skewing to the majority boundary. Thus, there will be fewer samples in the minority class to be misclassified [6].

To solve the problem of minority samples classification and avoid data loss of majority class in under-sampling in class imbalance problem we propose multi cluster-based majority random under-sampling and over-sampling techniques. Here we create multiple clusters of input dataset. Then we use the subset and all the 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 information loss of majority class and oversampling will help to balance data. Performance of classifier is evaluated by using Recall, Precision, F-measure and G-mean. Experimental results show that this hybrid approach improves the minority class prediction.

The rest of this paper is organized as follows. Section II represents the related work. Section III describes the multi-cluster based under-sampling approach and random over-sampling approach. Experiment

and results are conducted in section IV. The data set used for experiment is collect from UCI and KEEL repository. In section V we conclude the paper.

II. Related work

The most of research has been performed with respect to the class imbalance problem. The study found many methods for solving the problem of class imbalance, including data sampling and boosting. This methods are further categories into two parts first is data level and second is algorithmic level [7]. The first approach is also called as data preprocessing. It preprocess the data by under-sampling the majority instances i.e. remove the majority class sample or oversampling the minority instances i.e. adding or replicating minority class sample.

The simplest form of under-sampling is RUS. RUS randomly removes examples from the majority class until a desired class distribution is found. Over-sampling [8] is the opposite of under-sampling approach. It duplicates or generates new minority samples in the hope of reducing class imbalance. With over-sampling, the neighborhood of positive samples is assumed to be also positive as are the samples between two positive samples. Experimental results show that under-sampling produces better results than over-sampling in many cases. The belief is that although over-sampling does not lose any information about the majority class, it introduces an unnatural bias in favour of the minority class.

SMOTE (Synthetic Minority Oversampling Technique) was proposed by Chawla et al. to overcome the problem of imbalance by a special approach to generate new synthetic examples [10]. As the authors said, this method generates artificial examples based on the feature space similarities between original examples of the minority class. Its main idea is to take each example of the minority class and to introduce synthetic examples along the lines between it and it's selected nearest neighbors also from the minority class. Although SMOTE proved to be successful in the experiments it also has some shortcomings, which we further discuss. Firstly the way of identifying minority samples for over-sampling could be problematic [10].

Data cleaning techniques, such as Tomek links, have been effectively applied to remove the overlapping that is introduced from sampling methods. Generally speaking, Tomek links [11] can be defined as a pair of minimally distanced nearest neighbors of opposite classes. By removing overlapping examples, one can establish well-defined class clusters in the training set, which can, in turn, lead to well defined classification rules for improved classification performance. While sampling methods attempt to balance distributions by considering the representative proportions of class examples in the distribution, cost-sensitive learning methods consider the costs associated with misclassifying samples [12]. Instead of creating balanced data distributions through different sampling strategies, cost sensitive learning targets the imbalanced learning problem by using different cost matrices that describe the costs for misclassifying any particular data sample.

Fundamental to the cost-sensitive learning methodology is the concept of the cost matrix. The cost matrix can be considered as a numerical representation of the penalty of classifying examples from one class to another. For example, in a binary classification scenario, we define $C(Min, Maj)$ as the cost of misclassifying a majority class example as a minority class example and let $C(Maj, Min)$ represents the cost of the contrary case.

Typically, there is no cost for correct classification of either class and the cost of misclassifying minority examples is higher than the contrary case, i.e., $C(Maj, Min) > C(Min, Maj)$. The objective of cost sensitive learning then is to develop a hypothesis that minimizes the overall cost on the training data set. In regards to decision trees, cost-sensitive fitting can take three forms: first, cost-sensitive adjustments can be applied to the decision threshold; second, cost-sensitive considerations can be given to the split criteria at each node; and lastly, cost-sensitive pruning schemes can be applied to the tree. Pruning is beneficial for decision trees because it improves generalization by removing leaves with class probability estimates below a specified threshold. However, in the presence of imbalanced data, pruning procedures tend to remove leaves describing the minority concept. It has been shown that though pruning trees induced from imbalanced data can hinder performance, using un-pruned trees in such cases does not improve performance [13].

III. Multi cluster-based majority under-sampling

Under-sampling is an efficient strategy to deal with class imbalance. However, the drawback of under-sampling is that it lost many potentially useful data. In this section, we propose two strategies to explore the majority class examples ignored by under-sampling and generating balance training dataset. In order to achieve good prediction over minority class and avoid necessary information loss from the majority class, we use both K-means algorithm and random sampling approach.

Assume the size of the class imbalanced data set is N , which includes majority class samples \mathbf{X}_{Maj} and minority class samples \mathbf{X}_{Min} . For our multi-cluster-based majority under-sampling prediction algorithm, 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 i th cluster ($k_i \leq 1$) be $X_{Maj.i}$. Therefore, the ratio of the number of majority class samples to all the number of majority class samples in the i th cluster is

$$R_i = X_{Maj.i} / X_{Maj} \quad 1 \leq i \leq k$$

The number of selected majority class samples in the i th cluster for under-sampling is computed as

$$S_i = X_{min} * R_i \quad 1 \leq i \leq k \quad (1)$$

The above equation determines that more majority class samples would be selected in i th cluster which has more majority class samples. When apply this method to highly imbalance data then the size training data that created after this process will contain less number of data of sample. To improve prediction rate of classifier over majority and minority class we have multiply the selected samples number (S_i) by total number of cluster.

Algorithm:

Input: majority class samples, minority class samples.

Output: Balanced training set

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 select S_i majority samples in i th cluster. This is achieved by selecting S_i majority class samples in i th cluster randomly. Finally, we get majority sample subsets C

Step3: Combine C with all the minority class samples respectively to obtain the training set B .

IV. Experiments and result

It is now well-known that accuracy or error rate is the evaluation criterion for conventional classification. However, it is not an appropriate evaluation criterion when there is class-imbalance in dataset. In this paper, we use Recall, Precision, F-measure and G-mean 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 and G-mean are then defined as follows. Here, we take minority class as positive class.

$$\text{Sensitivity (Recall)} = TP / (TP+FN)$$

$$\text{Specificity} = TN / (TN+FP)$$

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

$$\text{F-measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

$$\text{G-mean} = \sqrt{\text{Sensitivity} * \text{Specificity}}$$

Table 1 confusion matrix

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

In our experiments, we use these four criteria to evaluate the classification performance of the approaches. Recall measures the predicted accuracy of the positive samples (minority samples). Precision refers to the proportion of actual positive samples among all samples that are predicted as being positive while Recall is the proportion of actual positive samples that are correctly identified by the classifier, which is the same as Sensitivity. 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, higher the recall lowers the precision. 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 [16]. F-Measure depicts the trade-off between precision and recall. Barandela et al. introduced the

metric called the geometric mean (GM) [17]. 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.

We tested our proposed approaches on UCI data sets. Information about these data sets are summarized in Table 2. Here, for each data set, Number of attributes, Number of positive and negative sample and Ratio are depicted.

Table 2. The table below lists the UCI datasets

	Data Set	No of Attribute	Majority Class samples	Minority Class samples	Imbalance Ratio
1	Glass	9	333	53	6.28
2	Heart	13	127	10	12.7
3	led	7	348	58	6
4	Yeast	8	1146	41	27.96
5	ecoli	7	208	61	3.4

4.1 Result Analysis

For evaluation purpose we test our technique on two classifier, SVM and KNN classifier. Both classifiers implanted in MATLAB and tested on five different dataset, performance of Classifier measure as follow.

Sensitivity and specificity are statistical measures of the performance of classifier. Sensitivity also called the *true positive rate* or the recall rate in some field's measures the proportion of actual positives which are correctly identified as such (e.g. the percentage of sick people who are correctly identified as having the condition). Specificity measures the proportion of negatives which are correctly identified as such (e.g. the percentage of healthy people who are correctly identified as not having the condition, sometimes called the *true negative rate*). G-mean is tradeoff between sensitivity and specificity. If G-mean is nearly equal to one, it means that TP and TN rate are well balance. If G-mean if zero it means that all the positive samples are misclassified. From this we can conclude that if any classifier having high G-mean, it means this classifier is good than another one. In this paper we are compare two classifier. Following tables give the result values.

Table 3 Performance of SVM classifier on original dataset

Data Set	Sensitivity	Precision	Specificity	G-mean	F-measure
Glass	0.41	0.88	0.99	0.64	0.56
Heart	0.33	1	1	0.57	0.5
led	0	0	1	0	0
Yeast	0.46	0.15	0.90	0.64	0.22
Ecoli	0	0	1	0	0

Table 4 Performance of SVM classifier After MCMUS algorithm

Data Set	Sensitivity	Precision	Specificity	G-mean	F-measure
Glass	0.96	0.76	0.96	0.94	0.85
Heart	0.66	0.67	0.96	0.80	0.66
led	0.91	0.53	0.86	0.88	0.67
Yeast	0.68	0.16	0.87	0.77	0.26
Ecoli	0.95	0.71	0.88	0.91	0.81

Table 5 Performance of KNN classifier on original dataset

Data Set	Sensitivity	Precision	Specificity	G-mean	F-measure
Glass	0	0	1	0	0
Heart	0	0	0.99	0	0
led	0.2	1	1	0.45	0.34
Yeast	0.09	0.15	0.14	0.97	0.11
Ecoli	0.81	0.87	0.96	0.88	0.84

Table 6 Performance of KNN classifier After MCMUS algorithm

Data Set	Sensitivity	Precision	Specificity	G-mean	F-measure
Glass	0.90	0.09	0.57	0.71	0.17
Heart	0.6	0.13	0.68	0.64	0.21
led	0.98	0.59	0.88	0.93	0.74
Yeast	0.75	0.16	0.85	0.80	0.26
Ecoli	0.93	0.73	0.89	0.91	0.82

As we know sensitivity and specificity are the true positive rate and true negative rate respectively. Classifier is better than another if its sensitivity and specificity are both not less than of the other classifier's. Graph

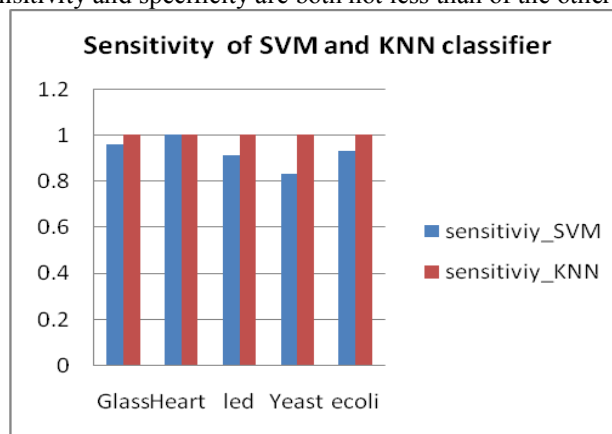


Fig1. Compare Sensitivity of SVM and KNN classifier

Form figure1 it is identified that sensitivity of KNN classifier is more than the SVM classifier.

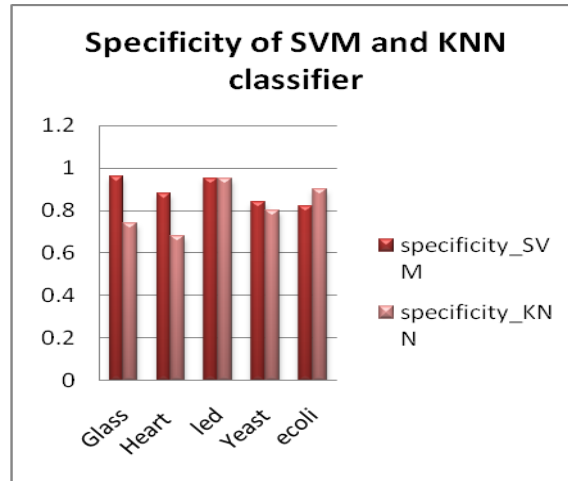


Fig2. Compare Specificity of SVM and KNN classifier

Form figure2 we can say that specificity of classifier is depends upon the size of dataset. Specificity of KNN for led, yeast and ecoli dataset is higher than other because size of this dataset is more than other while for Glass; heart Specificity of SVM is high.

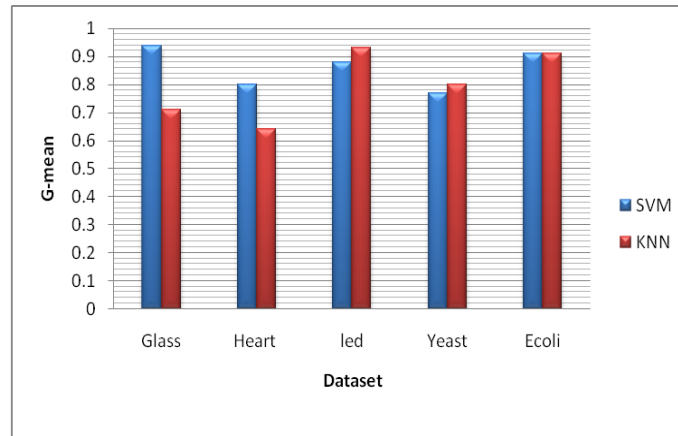


Fig3. G-mean for the datasets

G-means of classifier is also depends upon the size of samples. Figure3 indicate that G-mean of KNN classifier is higher than the SVM.

Table 7 the values of recall, precision, f- measure, g-mean on five imbalance using SVM classifier. The last rows AVG show the Average value.

Data Set	Sensitivity	Precision	Specificity	G-mean	F-measure
Glass	0.96	0.76	0.96	0.94	0.85
Heart	0.66	0.67	0.96	0.8	0.66
led	0.91	0.53	0.86	0.88	0.67
Yeast	0.68	0.16	0.87	0.77	0.26
Ecoli	0.95	0.71	0.88	0.91	0.81
AVG	0.832	0.566	0.906	0.86	0.65

Figure 4 show the performance of KNN classifier over the dataset after applying MCMUS algorithm.

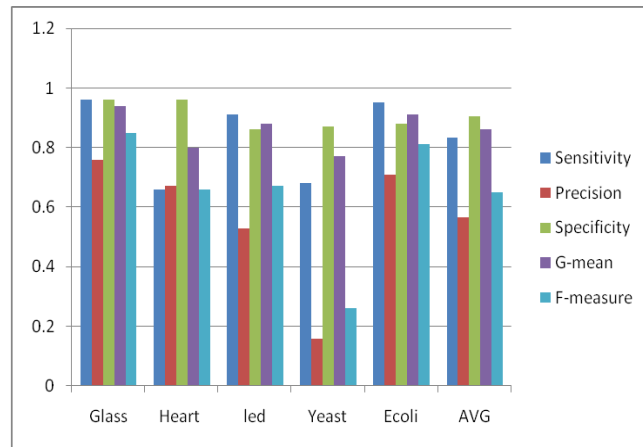


Fig.4. Sensitivity (Recall), Precision, F-measure, G-mean four approaches on five imbalance data sets using SVM classifier.

V. Conclusion

The result shows that Multi-clustered based majority under-sampling algorithm (MCMUS), it can improve the performance of classifiers for imbalanced datasets. In three out of five datasets MCMUS algorithm with K-NN classifier has the highest g-means as compared to SVM classifier. Comparison between F-measure of SVM classifier before and after applying MCMUS algorithm on dataset in chapter 5 conclude that, performance of classifier increase when preprocess data by MCMUS algorithm before applying classifier. K-means clustering algorithm used for clustering majority class samples in to k clusters. In this project k=3 used. Once data is clustered two methods are used to select the data samples from each clustered. These selected samples then combine with minority class sample and new training dataset will generate. The size of new training samples is small but helpful to classify the imbalance dataset. Comparison between SVM and KNN classifier demonstrates that performance of KNN classifier with MCMUS algorithm is better than SVM classifier. Although SVM classifier has good theoretical foundation in classification, performance will degrades as the class imbalance ratio increased.

VI. Acknowledgement

To discover, analyze and to present something new is to venture on an untrodden path towards an unexplored destination is an arduous adventure unless one gets a true torchbearer to show the way. I would have never succeeded in completing my task without the cooperation, encouragement and help provided to me by various people. Words are often too less to reveals one's deep regards. I take this opportunity to express my profound sense of gratitude and respect to all those who helped me through the duration of this thesis. I acknowledge with gratitude and humility my indebtedness to Ms. S. S. Dongre, Assistant professor Department of Computer Science Engineering, GHRCE, Nagpur, under whose guidance I had the privilege to complete this thesis. I wish to express my deep gratitude towards her for providing individual guidance and support throughout the thesis work.

I convey my sincere thanks to Dr. L. G. Malik, Head of Department of CSE, GHRCE Nagpur for her encouragement and cooperation.

I express my heartfelt gratitude towards Dr. P. R. Bajaj, Director, GHRCE, Nagpur, for their valuable guidance, encouragement, inspiration and the enthusiasm.

I would also like to thank all staff members and my co-students who were always there at the need of the hour and provided with all the help and facilities, which I required for the completion of my thesis. My greatest thanks are to all who wished me success especially my parents. Above all I render my gratitude to the Almighty who bestowed self-confidence, ability and strength in me to complete this work for not letting me down at the time of crisis and showing me the silver lining in the dark clouds.

Reference

- [1] Shuo Wang, Member, and Xin Yao, "Multiclass Imbalance Problems: Analysis and Potential Solutions", IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 42, No. 4, August 2012.
- [2] Tomasz Maciejewski and Jerzy Stefanowski, "Local Neighbourhood Extension of SMOTE for Mining Imbalanced Data" 2011 IEEE
- [3] Gang Wu and Edward Y. Chang, "KBA: Kernel Boundary Alignment Considering Imbalanced Data Distribution" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 17, NO. 6, JUNE 2005
- [4] Xinjian Guo, Yilong Yin, Cailing Dong, Gongping Yang, Guangtong Zhou, "On the Class Imbalance Problem" Fourth International Conference on Natural Computation
- [5] Jia Li, Hui Li *, Jun-Ling Yu, "Application of Random-SMOTE on Imbalanced Data Mining", 2011 Fourth International Conference on Business Intelligence and Financial Engineering
- [6] Yuchun Tang, Yan-Qing Zhang, Nitesh V. Chawla, and Sven Krasser, "SVMs Modeling for Highly Imbalanced Classification", Journal of LATEX class files, vol. 1, no. 11, November 2002
- [7] Mikel Galar, Fransico, "A review on Ensembles for the class Imbalance Problem: Bagging, Boosting and Hybrid- Based Approaches" IEEE Transactions On Systems, Man, And Cybernetics—Part C: Application And Reviews, Vol.42, No.4 July 2012
- [8] Chris Seiffert, Taghi M. Khoshgoftaar, "Mining Data with Rare Events: A Case Study" 19th IEEE International Conference on Tools with Artificial Intelligence 2007
- [9] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano "RUSBoost: A Hybrid Approach to Alleviating Class Imbalance" IEEE Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol. 40, No. 1, January 2010.
- [10] Tomasz Maciejewski and Jerzy Stefanowski "Local Neighborhood Extension of SMOTE for Mining Imbalanced Data" 2011 IEEE.
- [11] N.V. Chawla, A. Lazarevic, L.O. Hall, and K.W. Bowyer, "SMOTEBoost: Improving Prediction of the Minority Class in Boosting," Proc. Seventh European Conf. Principles and Practice of Knowledge Discovery in Databases, pp. 107-119, 2003
- [12] N.V. Chawla, N. Japkowicz, and A. Kolcz, "Editorial: Special Issue on Learning from Imbalanced Data Sets," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 1-6, 2004.
- [13] N. Japkowicz and S. Stephen, "The Class Imbalance Problem: A Systematic Study," Intelligent Data Analysis, vol. 6, no. 5, pp. 429-449, 2002.
- [14] Li Zhang WenXian Wang, "A Re-sampling Method for Class Imbalance Learning with Credit Data" 2011 International Conference of Information Technology, Computer Engineering and Management Sciences.
- [15] Haibo He and Edwardo A. Garcia "Learning from Imbalanced Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 9, SEPTEMBER 2009.
- [16] 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.
- [17] Barandela. R, Sanchez. J, S. Garc'a. V, and Rangel. E, "Strategies for learning in class imbalance problems," Pattern Recognition, 2003, pp. 849-851.
- [18] Barendela, R., Sanchez, J.S., Garcia, V., Rangel, E. (2003) Strategies for learning in class imbalance problems, Pattern Recognition vol.36, pp.849-851.
- [19] Jingrui He, May 2010, "Rare Category Analysis", Ph. D. Thesis, Department of Machine Learning, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.
- [20] Li,J., Li,H. and Yu,J.L. (2011) "Application of Random-SMOTE on Imbalanced Data Mining", bife, 2011 Fourth International Conference on Business Intelligence and Financial Engineering, pp.130-133.