ZKG INTERNATIONAL

ISSN: 2366-1313

CLASS IMBALANCE PROBLEM USING BOOSTING TECHNIQUE

THATIKONDA SOMASHEKAR[1], P. PREMCHAND [2]

[1] Research Scholar, University College of Engineering, Department of Computer science and Engineering, Osmania University, Hyderabad.
Email id:- soma.ts@gmail.com

[2] Professor, University College of Engineering, Department of Computer science and Engineering, Osmania University, Hyderabad.
Email id:- ppremchand@gmail.com

ABSTRACT

Significant modifications and advancements have been made in data classification in the past several years. The size of data expands along with the application area of technology. The infinite size and uneven nature of the data make classification challenging. The biggest concern in data mining is class inequality. When one of the two classes contains more samples than the other, there is an imbalance issue. Major sample classification is the primary emphasis of most algorithms, with minority sample classification ignored or misclassified. The extremely significant yet infrequently occurring minority samples are known as such. Three primary categories include the many techniques available for classifying unbalanced data sets: algorithmic approach, data pretreatment approach, and feature selection strategy. A strong ensemble learning technique called "boosting" enhanced the poor classifier's performance. A few examples of boosting algorithms include RUSBoost and SMOTEBoost. The feature selection approach may be applied to the categorization of data that is imbalanced.Every one of these methods has benefits and drawbacks. This work defines a comprehensive analysis of each technique, providing the proper path for future research on issues related to class imbalance.

Keywords: Class imbalance problem, RUSBoost and SMOTEBoost Skewed data, Imbalance data

1.INTRODUCTION

A lot of data with a skewed distribution is generated in several real-time applications. If there is a larger sample size from one class than the other, the data set is considered highly skewed. A class with a higher number of occurrences is referred to as a major class in an imbalanced data set, whereas a class with a substantially lower number of instances is referred to as a minor class. Applications like the ability to forecast uncommon but significant diseases for medical diagnosis are far more significant than standard care. Similar circumstances have been noted in other domains, including risk management, detecting network breaches, identifying fraud in financial operations, and anticipating technical equipment breakdowns.

Due to their bias towards the big classes, the majority of classifiers in this scenario exhibit extremely low classification rates for the smaller classes. It's also conceivable that the classifier overlooks the minor class and forecasts everything as a large class. The issues surrounding class imbalance have been addressed by several methods, which may be broadly categorized into three areas: algorithmic approach, data-preprocessing approach, and feature selection approach. Several aspects may influence the performance achieved by a classifier created by a Machine Learning system. One of these aspects is related to the difference between the numbers of examples belonging to each class. When this difference is large, the learning system may have difficulties to learn the concept related to the minority class. In this work1 , we discuss several issues related to learning with skewed class distributions, such as the

ISSN: 2366-1313

relationship between cost-sensitive learning and class distributions, and the limitations of accuracy and error rate to measure the performance of classifiers. Also, we survey some methods proposed by the Machine Learning community to solve the problem of learning with imbalanced data sets, and discuss some limitations of these methods.

Many traditional learning systems are not prepared to induce a classifier that accurately classifies the minority class under such situation. Frequently, the classifier has a good classification accuracy for the majority class, but its accuracy for the minority class is unacceptable. The problem arises when the misclassification cost for the minority class is much higher than the misclassification cost for the majority class. Unfortunately, that is the norm for most applications with imbalanced data sets, since these applications aim to profile a small set of valuable entities that are spread in a large group of "uninteresting" entities. In this work we discuss some of the most frequently used methods that aim to solve the problem of learning with imbalanced data sets. These methods can be divided into three groups: 1. Assign misclassification costs. In a general way, misclassify examples of the minority class is more costly than misclassify examples of the majority class. The use of cost-sensitive learning systems might aid to solve the problem of learning from imbalanced data sets; 2. Under-sampling. One very direct way to solve the problem of learning from imbalanced data sets is to artificially balance the class distributions. Under-sampling aim to balance a data set by eliminating examples of the majority class; 3. Over-sampling. This method is similar to under-sampling. But it aims to achieve a more balanced class distributions by replicating examples of the minority class. This work is organised as follows: Section 2 discusses why accuracy and error rate are inadequate metrics to measure the performance of learning systems when data have asymmetric misclassification costs and/or class imbalance; Section 3 explains the relationship between imbalanced class distributions and cost-sensitive learning; Section 4 discusses which class distributions are best for learning; Section 5 surveys some methods proposed by the Machine Learning community to balance the class distributions; Section 6 presents a brief discus sion about some evidences that balancing a class distributions has little effect in the final classifier; finally, Section 7 shows the conclusions of this work.

Sampling is applied to data in the data-preprocessing procedure, wherein either new samples are introduced or current samples are discarded. Over-sampling refers to the act of adding additional samples to an existing sample, and under-sampling refers to the process of eliminating a sample. The second strategy for resolving the issue of class imbalance is to develop or alter an algorithm. The algorithms comprise kernel-based learning techniques including support vector machine (SVM) and radial basis function, as well as recognition-based methods and the cost-sensitive method. Because of the large amount of data and the high-class imbalance ratio, using an algorithm alone is not a good idea. Instead, a novel methodology called the combination of sampling method and algorithm is utilized. Algorithms in classification often emphasise accurately classifying samples belonging to the majority class. Misclassifying a rare event can lead to more significant issues than a typical occurrence in many applications.

In the context of medical diagnosis, misclassifying non-cancerous cells may result in more clinical testing, but misclassifying dangerous cells has grave health hazards. Minority class examples are more likely than majority class examples to be incorrectly classified in classification problems with imbalanced data, as most machine learning algorithms are designed to maximize overall classification accuracy, which leads to minority class misclassification.

2.LITERATURE SURVEY

Data sampling has received much attention in datamining related to class imbalance

ISSN: 2366-1313

problems. Data samplingtries to overcome the imbalanced class distribution problemby adding samples to or removing sampling from the dataset [1]. This method improves the classification accuracyof minority class but, because of infinite data streams andcontinuous concept drifting, this method cannot be suitablefor skewed data stream classification. Most existingimbalance learning techniques are only designed for two-classproblems. Multiclass imbalance problemsare mostly solvedby using class decomposition. AdaBoost.NC [2] is anensemble learning algorithm that combines the strengthof negative correlation learning and boosting methods.This algorithm is mainly used in multiclass imbalanced datasets. The results suggest that AdaBoost.NC combined withrandom oversampling can improve the predictionaccuracy on the minority class without losing the overallperformance compared to other existing class imbalancelearning methods. Wang et al. proposed the classificationalgorithm for skewed data stream in [3], which shows thatclustering sampling outperforms traditional undersamplingsince clustering helps to reserve more usefulinformation. However, the method cannot detect conceptdrifting. CNNs are gaining popularity in a number of machine learning application domains and are currently contributing to the state of the art in the field of computer vision, which includes tasks such as object detection, image classification, and segmentation. They are also widely used in other areas, including natural language processing or speech recognition where they are replacing or improving classical machine learning models [4]. CNN is a special type of feed-forward artificial neural network. It is based on the idea of transforming input data with a set of differentiable operations. The main difference compared to the standard multilayer perceptron is that CNN integrates automatic feature extraction and a classifier itself in one model. This allows them to learn meaningful hierarchical representations [5]. It is achieved by having small convolutional kernels with a limited receptive field that widens with the number of layers. Shared parameters in convolutional layers result in two desirable properties [6]. Firstly, they make CNNs translation invariant. Intuitively, it means that patterns and objects can appear anywhere in the processed image or other type of input. Secondly, they significantly reduce model complexity in terms of the number of parameters. This property makes them much easier to train and less prone to overfitting. A standard CNN architecture comprises fully connected layers and a number of blocks consisting of convolution kernels, activation function layer and max pooling [7]. The most commonly used activation function is rectified linear unit (ReLU). It effectively diminishes the effect of vanishing gradient and enables efficient training of deep networks. Still, stacking too many layers results in lower performance even on a training set. This issue is overcome in residual networks [8]. Therefore, with the help of modern GPUs and large datasets we are able to train extremely deep networks in a reasonable time. The algorithm most widely used for training all kinds of neural networks is backpropagation [9] together with mini-batch stochastic gradient descent (SGD) optimization [10]. The main difficulty in training large neural networks is their sensitivity to changes in hyperparameters [11]. They can be associated with SGD, e.g. the value of initial learning rate, momentum, weight decay or mini-batch size. Another important set of hyperparameters to tune is model-related [12], e.g. the number of layers, size and number of kernels, or finally initialization strategy [13]. All of them can have a huge impact on performance of a trained network.Chris [14] proposed that both sampling andensemble techniques are effective for improving theclassification accuracy of skewed data streams. SVM-basedone-class skewed data streams learning method wasproposed in [15], which cannot work with concept drift.Liu et al. [16] proposed one class data streams algorithm,which follows the single classifier approach and can beused to classify text streams. One of the most commondata sampling techniques is Random Under-sampling.RUS simply removes examples from the majority class atrandom until a desired class distribution is achieved.RUSBoost is a new hybrid sampling and boostingalgorithm for learning from skewed training data.RUSBoost provides a simpler and faster alternative toSMOTEBoost which is another algorithm that combinesboosting and data

ISSN: 2366-1313

sampling [17]. RUS decreases the timerequired to construct a model, which is beneficial whencreating an ensemble of models that is used in boosting.

3.PROPOSED SYSTEM

Numerous strategies and methods are proposed by the literature review to address the issue of the unbalanced sample distribution. These strategies may be broadly categorized into three techniques: feature selection, algorithms, and sampling.

3.1 Sampling

To address issues with a dataset's distribution, sampling techniques—also referred to as data preparation methods—involve intentionally resampling the data set. Sampling can be accomplished in one of two ways: either by oversampling the minority class by undersampling the majority class or by combining the two methods.

Under-sampling: The random undersampling approach, which aims to balance the class distribution by randomly deleting the majority class sample, is the most significant undersampling technique. The random undersampling approach [4] is seen in Figure 1. The method's drawback is the loss of important data.

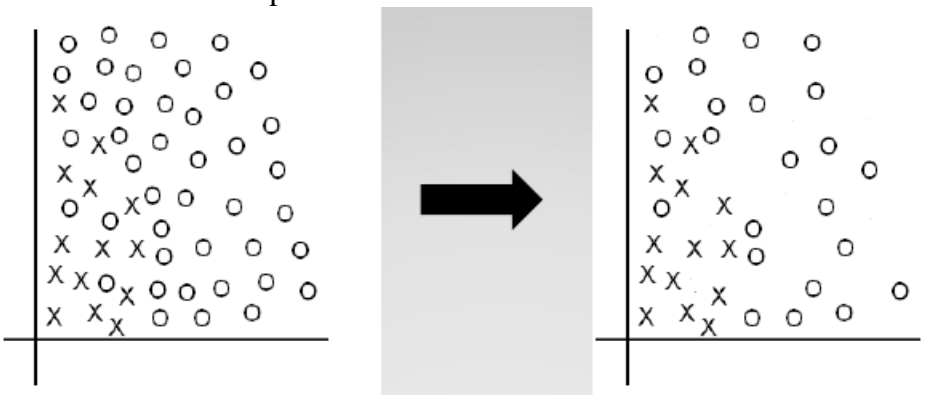


Fig.1.Randomly removes the majority sample.

Over-sampling: Random Over-sampling methods alsohelp to achieve balanced class distribution by replicationminority class sample.

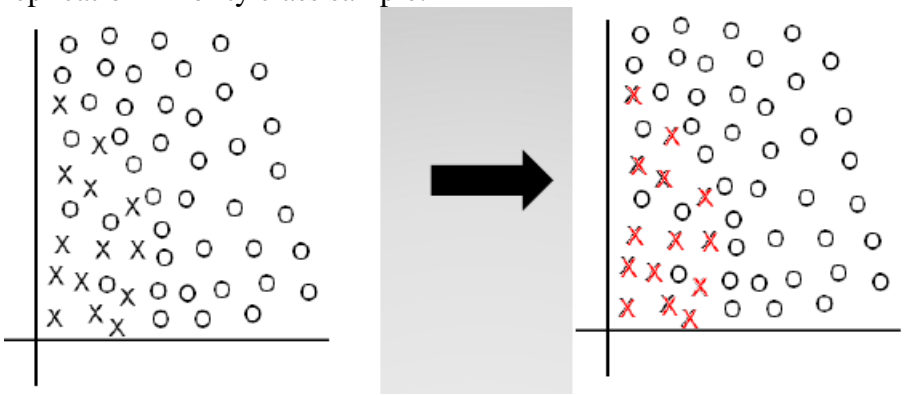


Fig.2. Replicate the minority class samples

Additional information is not required because the data is reused [12]. By creating fresh synthetic data from a minority sample, this issue can be resolved. Synthetic minority cases are produced by SMOTE to oversample the minority class. Due to the extremely tiny number of minority samples in the initial data set, this method's learning process takes longer.

3.2 Algorithms

ISSN: 2366-1313

To address the issue of class imbalance, several innovative methods have been developed. This strategy aims to maximize the learning algorithm's performance on unobserved data. Methods of one-class learning acknowledged that the sample belonged to that class and rejected others. When dealing with multi-dimensional data sets, for example, one-class learning performs better than other methods [5]. Changing the class distribution and using cost in decision-making is another method to enhance classifier performance. Learning strategies that are cost-sensitive aim to optimize a loss function related to a collection of data. The fact that the majority of real-world applications lack uniform consequences for misclassifications serves as the driving force for these learning techniques. Since it is usually uncertain how much each type of error would cost, these approaches must calculate the cost matrix from the data and apply it to the learning stage. A similar concept to cost-sensitive learners is manipulating a machine's bias to benefit the minority class [8].

By selecting the class with the lowest conditional risk, cost-sensitive classification aims to reduce the cost of incorrect categorization. Table 1 displays the cost matrix with two classes, i and j. The misclassification cost, λij, is shown. Indicating that the cost of accurate categorization is zero, and the diagonal elements are zero. Changing the classifier is an additional algorithmic strategy for skewed data distribution [8]. To transfer the unbalanced dataset onto a higher dimension space, the kernel-based technique uses the concept of support vector machines. It is thus expected that the classifier would perform significantly better than learning from the original dataset when combined with an ensemble approach or oversampling methodology.

Table 1: Cost matrix

		Prediction	
		Class i	**Class j**
True	**Class i**	**0**	λ_{ij}
	Class j	λ_{ji}	**0**

Several modifications have been made to SVMs to increase their accuracy in class prediction, and the results indicate that SVMs are capable of handling skewed vector problems without adding noise [9]. In the event of unbalanced data, boosting techniques may be coupled with SVMs rather well [16].

3.3 Feature Selection

Generally speaking, the objective of feature selection is to choose a subset of j features—where j is a user-defined parameter—that enables a classifier to achieve optimal performance. It makes use of filters that assign a rule-based score to each feature separately for high-dimensional data sets. A crucial stage in many machine learning algorithms is feature selection, particularly in cases when the input is high-dimensional. Utilizing feature selection techniques is crucial as the problem of class imbalance is frequently coupled with the problem of the data set's excessive dimensionality. High dimensional class imbalance issues can require more than sampling strategies and computational approaches to resolve [5]. Though feature selection as a general component of machine learning and data mining techniques has been extensively studied, its significance in addressing the issue of class imbalance is relatively new, with the majority of research on the topic emerging in the last few years [18]. During this time,

ISSN: 2366-1313

several scholars have studied the use of feature selection in addressing the issue of class imbalance. To categorize text data taken from the Yahoo Web hierarchy, Ertekin [17] investigated how well feature selection metrics performed in this regard. Using the naïve Bayes classifier, they assessed the strength of the top features by applying nine distinct metrics to the data set.

4.RESULTS

Analysis drawn from the comparative study of each of the algorithms is shown in the following table.

Table 2: Comparative Study

Sr. No	Algorithm	Advantages	Disadvantages
1	AdaBoost.NC[1]	Improve prediction accuracy of minority	Ignore overall performance of classifier
2	RUSBoost [2]	Simple, faster and less complex than SMOTE Boost algorithm	Unable to solve Multiclass imbalance problem
3	Infinitely imbalanced logistic regression [6]	Mostly used for binary classification	Performance is depends on number of outlier in data.
4	Linear Proximal support vector machines[7]	Handle dynamic class imbalance problem	No consideration for distribution of sample
5	BoostingSVM[20]	Improved the performance of SVM classifier for prediction minority sample	Ignore imbalance class distribution.

Class imbalance issues are prevalent in many locations. Numerous data mining approaches offer a useful but insufficient answer. The type of data utilized for the experiment has a significant impact on determining which strategy is optimal for solving a data distribution problem.

ISSN: 2366-1313

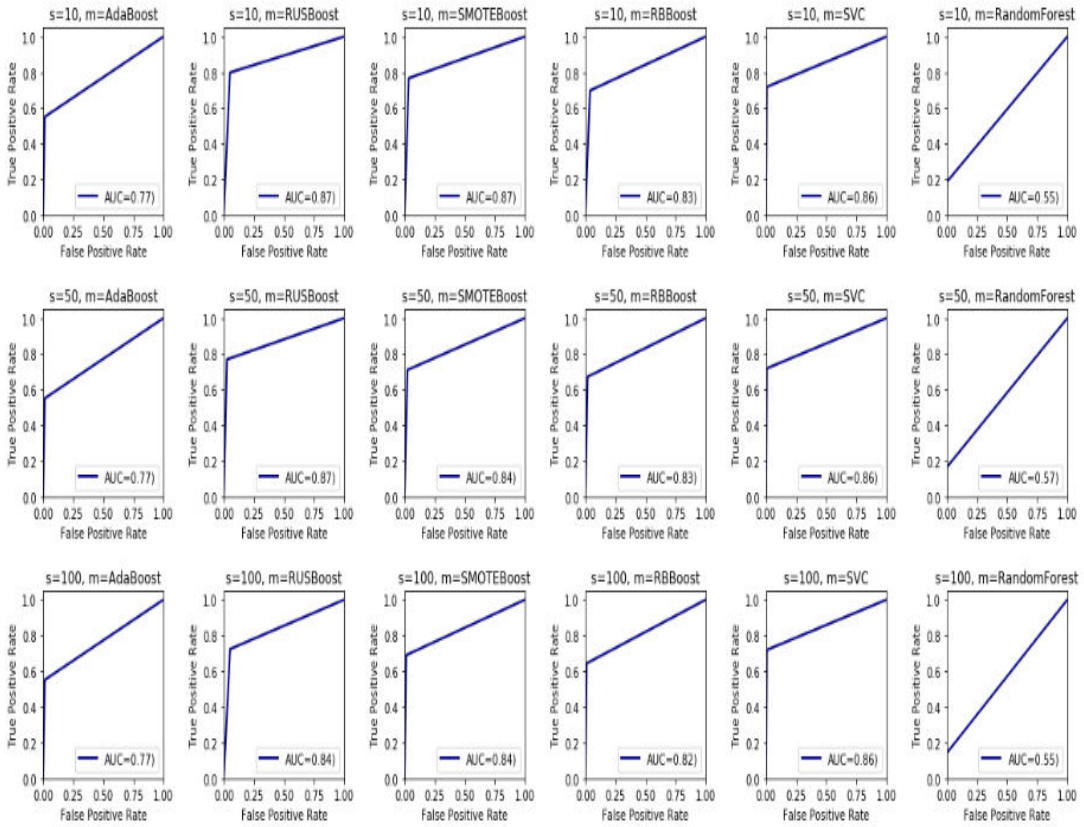


Fig.3.ROC Curve Plot of different Algorithms

ROC curve are constructed of two factors:

1. Y: Recall=TP/(TP+FN)

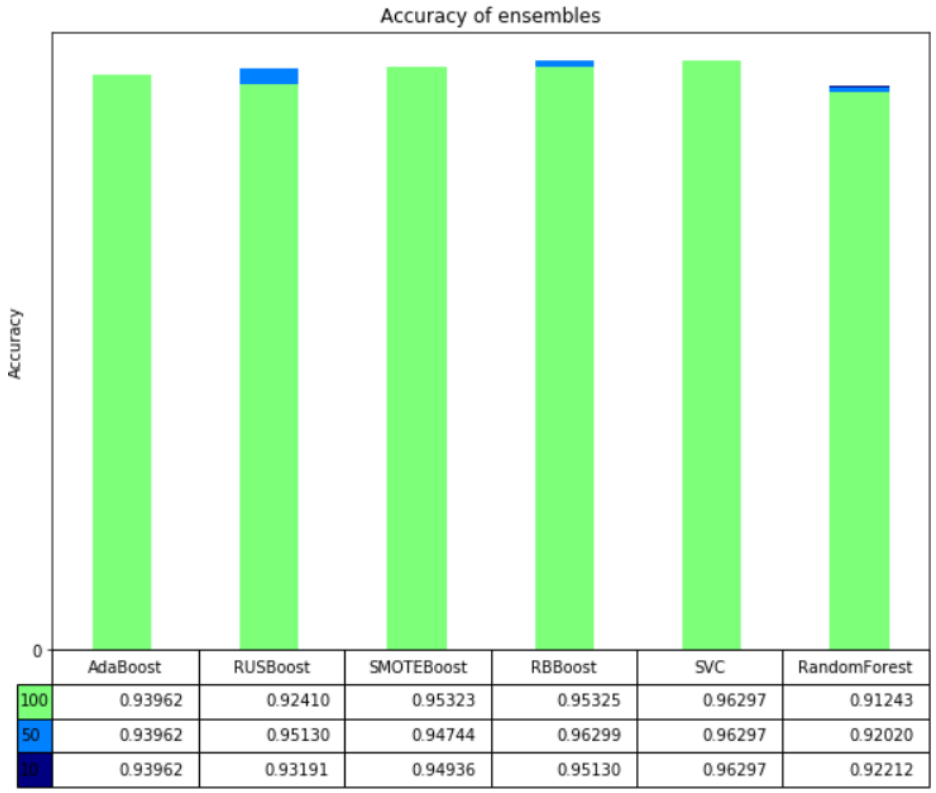
2. X: FPR=FP/(TN+FP)

As we explaned before in AUC section, increasing ensemble size decreases recall so AUC will decrease which means that ROC curve will tend towards random model curve which is y=x.

1. RandomForest has almost same curve as random model y=x
2. SMOTEBoost and RUSBoost outperformed all other approaches but about %1 increase from SVM and %4 from their nearest opponent which is RBBoost while outperformed base AdaBoost by approximately %10 percent increase in score.
3. SVM model has the lowest FPR rate while SMOTE, RUS and RB achieved same rate by increasing in ensemble size.
4. Recall of each ensemble decreases by increasing ensemble size as we can see in SMOTE, RUS and RB the curves tend toward y=x which means lower AUC score.

ISSN: 2366-1313

5. AdaBoost has low FPR rate but not as good as SVM meanwhile its very low recall failed it to have high AUC score.



	AdaBoost	RUSBoost	SMOTEBoost	RBBoost	SVC	RandomForest
100	0.93962	0.92410	0.95323	0.95325	0.96297	0.91243
50	0.93962	0.95130	0.94744	0.96299	0.96297	0.92020
10	0.93962	0.93191	0.94936	0.95130	0.96297	0.92212

Fig.4. Accuracy Comparision

Based on this graph, we can understand that SMOTE, RUS and RB algrithms all achived higher accuracy score versus base AdaBoost classifier. Also we can see that RandomForest which does not boosting, has lowest score. Although SVM model has highest accuracy, in the following graphs we can see that SVM's high accuracy is due to focusing on majority class.

5.CONCLUSION

Data preparation is said to offer a more practical solution than other techniques as it makes it possible to add new information or remove duplicate information, which contributes to data balance. Boosting is another useful strategy for resolving the class imbalance issue. A strong ensemble learning technique called "boosting" enhanced the poor classifier's performance. A few examples of boosting algorithms include RUSBoost and SMOTEBoost. The feature selection approach may be applied to the categorization of data that is imbalanced. An algorithm for feature selection will perform differently depending on the type of task. Ultimately, the hybrid approach—applying two or more techniques—offers a superior solution to the issue of class imbalance.

REFERENCES

ISSN: 2366-1313

[1] Shuo Wang, Member, and Xin Yao, “Multiclass ImbalanceProblems: Analysis and Potential Solutions”, IEEETransactions On Systems, Man, And Cybernetics—Part B:Cybernetics, Vol. 42, No. 4, August 2012.
[2] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, andAmri Napolitano “RUSBoost: A Hybrid Approach toAlleviating Class Imbalance”IEEE Transactions On Systems,Man, And Cybernetics-Part A: Systems And Humans, Vol.40, No. 1, January 2010.
[3] Björn Waske, Sebastian van der Linden, Jón AtliBenediktsson, Andreas Rabe, and Patrick Hostert“Sensitivity of Support Vector Machines to Random FeatureSelection in Classification of Hyper-spectral Data”, IEEETransactions On Geosciences And Remote Sensing, Vol. 48,No. 7, July 2010
[4] Xinjian Guo, Yilong Yin1, Cailing Dong, Gongping Yang,Guangtong Zhou,“On the Class Imbalance Problem” FourthInternational Conference on Natural Computation, 2008.
[5] Mike Wasikowski, Member and Xue-wen Chen, “Combatingthe Small Sample Class Imbalance Problem Using FeatureSelection”, IEEE Transactions on Knowledge and DataEngineering, Vol. 22, No. 10, October 2010.
[6] Rukshan Batuwita and Vasile Palade,“Fuzzy Support VectorMachines for Class imbalance Learning” IEEE TransactionsOn Fuzzy Systems, Vol. 18, No. 3, June 2010.
[7] Lei Zhu, Shaoning Pang, Gang Chen, and AbdolhosseinSarrafzadeh, “Class Imbalance Robust IncrementalLPSVM for Data Streams Learning” WCCI 2012 IEEEWorld Congress on Computational Intelligence June, 10-15,2012 - Australia.
[8] David P. Williams, Member, Vincent Myers, and MirandaSchatten Silvious, ”Mine Classification With ImbalancedData”, IEEE Geosciences And Remote Sensing Letters, Vol.6, No. 3, July 2009.
[9] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse,Amri Napolitano “A Comparative Study of Data Samplingand Cost Sensitive Learning”, IEEE InternationalConference on Data Mining Workshops. 15-19 Dec. 2008.
[10] Mikel Galar,Fransico, “A review on Ensembles for theclass 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
[11] Yuchun Tang, Yan-Qing Zhang, Nitesh V. Chawla, andSven Krasser “Correspondence SVMs Modeling for HighlyImbalanced Classification” IEEE Transactions On Systems,Man, And Cybernetics—Part B: Cybernetics, Vol. 39, No. 1,February 2009
[12] Peng Liu, Lijun Cai, Yong Wang, Longbo Zhang“Classifying Skewed Data Streams Based on Reusing Data”International Conference on Computer Application andSystem Modeling (ICCASM 2010).
[13] Zhi-Hua Zhou, Senior Member, and Xu-Ying Liu “TrainingCost-Sensitive Neural Networks with Methods Addressingthe Class imbalance Problem” IEEE Transactions OnKnowledge And Data Engineering, Vol. 18, No. 1, January2006
[14] Qun Song Jun Zhang Qian Chi “ Assistant Detection ofSkewed Data Streams Classification in Cloud Security”,IEEE Transaction 2010.
[15] Nadeem Qazi,Kamran Raza, “Effect Of Feature Selection,Synthetic Minority Over-sampling (SMOTE) And UndersamplingOn Class imbalance Classification”, 14th International Conference on Modeling and Simulation- 2012.
[16] Nitesh V. Chawla, Nathalie Japkowicz, Aleksander Ko lcz“Special Issue on Learning from Imbalanced Data Sets”Volume 6, Issue 1 - Page 1-6.
[17] S¸ eyda Ertekin1, Jian Huang, LeonBottou, C. Lee Giles“Active Learning in Imbalanced Data Classification”
[18] Saumil Hukerikar, Ashwin Tumma, Akshay Nikam, VahidaAttar “SkewBoost: An Algorithm for Classifying ImbalancedDatasets” International Conference on Computer &Communication Technology (ICCCT)-2011.

ISSN: 2366-1313

[19] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse,“Improving Learner Performance with Data Sampling andBoosting” 2008 20th IEEE International Conference onTools with Artificial Intelligence.
[20] Benjamin X. Wang and Nathalie Japkowicz “BoostingSupport Vector Machines for Imbalanced Data Sets”Proceedings of the 20th International Conference onMachine Learning-2009.