International Journal of Scientific & Engineering Research, Volume 7, Issue 5, May-2016 56
ISSN 2229-5518

An Efficient Mechanism for Classification of Imbalanced Big Data

Krithika M V[1] ,Rajeev Bilagi[2], and Dr. Prashanth C M[3]

Abstract — In many real world applications, there is wide increment in data generation and storage. The classification algorithms are facing a problem in the classification of highly imbalanced datasets. All the classification algorithms are biased towards the majority classes ignoring most of the significant samples present in the minority class. To resolve this issue, a method called Hybrid Sampling technique is proposed to deal with multi class imbalanced data. This methods acts by balancing the data distribution of all the classes to some reference point called mean and addresses the problem of imbalanced data by eliminating insignificant samples that exists in the majority class.

Index Terms—*Classification, data mining, Imbalance Problems, Multi Class Imbalanced data, Sampling Techniques*

1. INTRODUCTION

Big data is a popular topic of research in today's world because of stupendous data generation and storage. As the volume, diversity and complexity of the data increases, there is a need for efficient algorithm, techniques and analysis to extract the value hidden information. Data mining techniques cannot analyze massive amount of data in a reasonable amount of time[1]. Decision making requires well defined methods for extracting knowledge or information from various domains. Data mining is the prediction of useful information from large datasets.

Classification is one of the important area of application in data mining. Classification involves assigning a class label to a set of undefined examples. Classification becomes a serious issue with highly skewed dataset. The classification algorithms proposed so far dealt with two class imbalanced problem. It is necessary to solve and negotiate the multi class imbalance problem that occurs in the real world. Class imbalance[2] problem is a major issue in the field of big data, data mining and machine learning techniques. All the classification algorithms are biased towards the majority classes, ignoring most of the minority class samples that occur very rarely but are found to be the most important.

[1]*Post Graduate Student, Dept of CS&E, SCE Bangalore, India.Email-id:kritsmv14@gmail.com*
[2]*Associate Professor , Dept of CS&E , SCE Bangalore, India. Email-id: rajeevbilagi@sapthagiri.edu.in*
[3]*Professor & HOD, Dept of CS&E, SCE Bangalore, India. Email-id: hodcse@sapthagiri.edu.in*

A. Class Imbalanced Data Problem

Class imbalance problem is said to occur when the number of instances in one class(majority class) is outnumbered by the number of instances in the other classes(minority classes). A class having large number of example instances is called as a majority class(negative class) and the one having relatively less number of example instances is called as a minority class or a positive class. As the majority class has large number of training instances, the classifiers have good accuracy on the negative class but show very poor classification rates on the minority classes. Classification algorithms[3] on imbalanced dataset show poor performances due to the following reasons:

1.The goal of any classification algorithm is to minimize the overall error rates.

2.They assume the class distribution of various class labels as equal.

3. Misclassification error rates of all the classes are considered to be equal[4].

4.Most of the data mining algorithms assume balanced distribution of classes and ignore all the minority classes when dealing with imbalanced dataset.

They blindly assume that all the costs associated with every misclassification is same as the ones that are correctly classified. This is not the case in many real world applications. Most of the real time applications contain dataset with skewed distribution[5]. A skewed dataset is the one, which has higher number of samples in one class than the other [6][7].

B. Effects of misclassification

In medical diagnosis application[5], prediction of the occurrence of rare disease is more important than treating the normal diseases that occur very frequently[7]. For example, consider a disastrous malignant disease such a cancer. As this disease occurs very rarely, the number of patients who are tested positive for this disease belong to the minority class label and the ones tested negative are categorized under majority class label. As the classifier is biased towards the majority class(class consisting of the patients who are tested negative), any patient who is tested positive for the cancer disease will also get classified as a cancer free patient. In this case, missing a cancer patient causes more threat than the false positive errors because he/she may even lose her life if proper medication is not given on time. Class imbalance problem is also observed in the areas such as fraud detection in banking operations, network intrusion detection[8], managing risk and predicting failures of technical equipments. When such situations are observed, the classifier shows poor classification rates on the minor classes because the classifiers are biased towards the majority classes.

C. Mitigation of misclassification rates

For the correct classification of minority classes, the classifier has to be trained with a balanced data so that it can evenly segregate and distinguish both the classes.

Techniques that can be used to solve the class imbalance problem[1][9] can be divided into three basic categories:

i] Data Level Approach[10] : This approach tries to rebalance the class distribution by employing preprocessing technique. Preprocessing technique involves the application of methods such as oversampling and undersampling.

ii] Algorithm Level Approach[11] : This approach modifies or adopts the existing algorithms over the imbalanced class distribution and achieves a balanced distribution of both the classes by biasing the classifier towards the minority class.

iii] Cost Sensitive approach[12] : Cost sensitive approach takes misclassification error costs into consideration. It does this by associating higher error costs to each of the misclassified example. In other words, no cost is associated for a correctly classified example. Its objective is to minimize the overall cost on the training examples.

D. Sampling

Sampling may be defined as the inference or judgment made on some part of the aggregate or totality that is considered. Sampling can be applied over a dataset either to create/add new samples or to remove few samples from the existing dataset. Sampling is a preprocessing technique.

Data sampling may be achieved in two different ways:

Adding a new sample to the existing dataset can be referred as oversampling and removing or eliminating the samples from the existing dataset can be referred as undersampling. As the class imbalance ratio is high, sampling method can be used with the application of an algorithm[13].

Undersampling: Random undersampling method is one of the most important method in undersampling. Random removal of samples from the majority class is the technique employed by this method to achieve a balanced distribution. Figure 1 shows the method of removing samples from the majority class by employing random undersampling method[14]. Training examples from the majority class are eliminated randomly to get a balanced ratio between the classes that are considered.

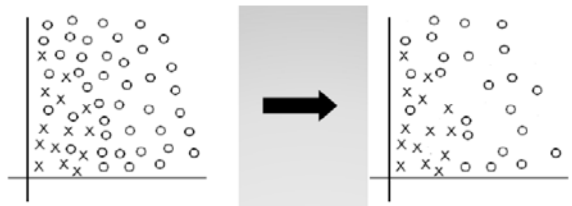

Fig.1 Random Undersampling

Oversampling: Random oversampling method acts by replicating the randomly chosen minority class samples to achieve a balanced distribution on both the classes[15]. Figure 2 shows random oversampling. It is a simple resampling effective approach.

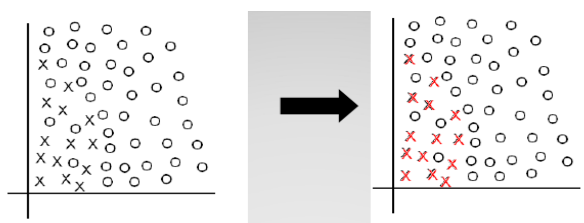

Fig.2 Random Oversampling

II. RELATED WORK

A. Parallel Selective Sampling Method:

Parallel selective sampling method[16] considered huge amount of imbalanced data and provided solutions for classifying them. Performances were assessed using Parallel Selective Sampling (PSS), a method that reduces the imbalance in large data sets by selecting the data from the majority class. PSS-SVM was used for improving the classification rates.

Disadvantage: As PSS is an undersampling method, it removes or eliminates the examples from the majority class. These randomly removed examples affect the class distribution because, the eliminated samples may be the significant samples that are considered to be important during classification. Eliminating such significant samples may degrade the classifiers performance.

B. Neighborhood Based Rough Set Boundary Synthetic Minority Oversampling Technique:

Hu, F., Li, H. [17] proposed an oversampling method, called Neighborhood Based Rough Set Boundary Synthetic Minority Oversampling Technique (NRSBoundarySMOTE), to achieve a balanced distribution. The minority class samples present in the boundary region are considered for oversampling.

Disadvantage: Though the proposed method is an effective method for oversampling, filtering the synthetic samples take more time and hence there is a difficulty in processing the large datasets that are considered. Also, oversampling method consists of the instances or the datasets that do not represent the universal sample. This is because, the oversampling method creates a superset of the original dataset by replicating some of the examples of the minority class.

C. Cost sensitive learning and Ensemble techniques:

Lopez, Fernandez, Garcia, Palade & Herrera [18] proposed solutions and presented specific metrics to evaluate the performance in class imbalanced learning by reviewing many issues in machine learning and its applications. They described preprocessing, cost sensitive learning and ensemble techniques.

Disadvantage: Their classification with imbalanced data was not able to provide good alternatives or define good solutions because they did not pay much attention on measuring and detecting the most significant data properties that is required for classification.

D. Cost sensitive learning methods:

Cost Sensitive Learning[19] for Imbalanced Bad Debt Datasets in Healthcare Industry provides an effective way of classifying imbalanced bad debt datasets for unknown cases by using cost sensitive learning methods and compares the results with undersampling and oversampling methods that is used for processing imbalanced datasets. They also analyzed how a semi supervised learning algorithm behaves under different circumstances.

Disadvantage: Their results showed that the minority classification accuracy rates were very poor. However, the overall and majority class classification accuracy rates improved when using oversampling and the cost sensitive learning methods with the semi supervised learning. In order to handle the imbalanced bad debt datasets very well, the semi supervised learning algorithms need to be further improvised.

E. Imbalanced big data classification using Random Forest Approach:

Rio, Lopez, Benitez, & Herrera [20] used Random Forest classifier to analyze the performance over the techniques such as oversampling, undersampling and cost sensitive learning approach to deal with imbalanced datasets. They evaluated the performance of diverse algorithms using Random Forest classifier and showed that their classifier outperforms all others with respect to the data that they have considered.

Disadvantage: There is a drop in the performance accuracy though there is a progress in time. They did not emphasize the need to analyze the intrinsic properties of data and also didn't find the necessity to design new techniques that generates synthetic data in a best way to represent the minority class samples when map reduce framework is considered.

F. Data and algorithmic level Approach:

Vaishali Ganganwar[3] proposed solutions to deal with class imbalance problem, both at the data and algorithmic levels. They artificially rebalanced the imbalanced dataset to increase the accuracy of the classifier using oversampling and undersampling through support vector machine, rough set based minority class oriented rule learning methods and cost sensitive classifier. They concluded that oversampling would be better for local classifiers than under sampling and found that undersampling strategies was well suited while employing global learning classifiers.

Disadvantage: Many other worthwhile research possibilities that could increase the performance of the classifier and enhance its accuracy rates were not considered as their area of interest. Better results for imbalanced datasets could be achieved if robust and skew insensitive classifiers are developed. Also, the classification methods focused only on two class imbalance problem.

G. Borderline SMOTE Technique:

Borderline SMOTE[21]: is a Synthetic minority oversampling technique (SMOTE) that addresses the problem of imbalanced classification of data sets. They presented two new oversampling technique based on SMOTE namely, borderline SMOTE1 and borderline SMOTE2.

Disadvantage: Borderline SMOTE suffer from curse of dimensionality because they rely heavily on Euclidean distance. They did not consider how to handle danger examples in different strategies. Also, they focused only on two class imbalance problem.

H. RUSBoost Approach:

RUSBoost: A Hybrid Approach to Alleviating Class Imbalance[22] evaluates the performances of RUSBoost and SMOTEBoost, for learning from training data set that is skewed.

Disadvantage: Though it is Simple, faster and less complex than SMOTE Boost algorithm, it is unable to solve Multiclass imbalance problem.

III. METHEDOLOGY

A. Architecture

Figure 3 summarizes the proposed technique for handling multi class imbalance problem. Datasets are obtained from UCI machine learning repositories. To balance the multi class imbalanced big data, hybrid sampling technique is applied over the minority and majority samples. Further, to gain fast, scalable and parallel implementations,

MapReduce framework is used for classification. Finally, the performance of the classifier is evaluated.

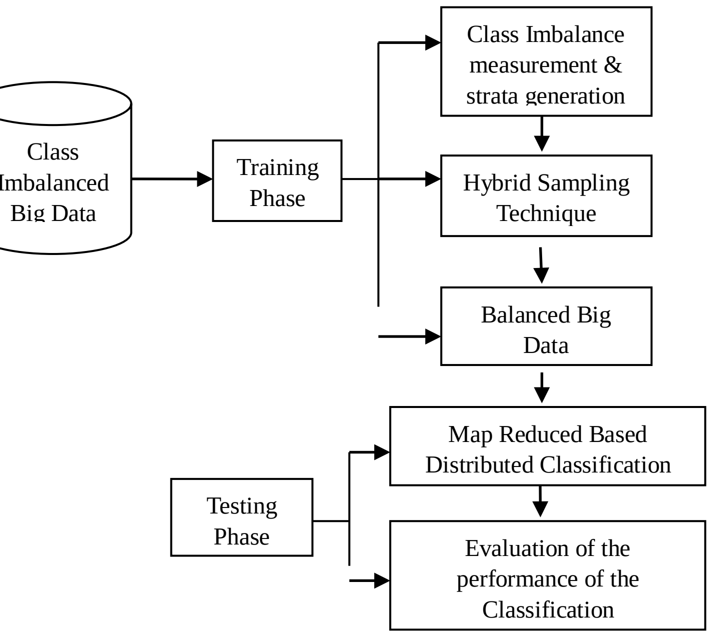

Fig. 3 Proposed system architecture

B. Modules

The methodology involved in balancing the imbalanced class distribution can be divided into 4 phases.

a. Class Imbalance measurement
b. Strata generation
c. Hybrid Sampling
d. Classifier Training

a. Class Imbalance Measurement: Given a dataset for each class, measure the class imbalance for each of these classes from this dataset. E.g. Consider the class distribution in a class imbalanced data, with 4 classes containing a total of 1000 training records as depicted in table 1 below:

Table 1 Class distribution

Class	No.of Records
Class #1	150
Class #2	600
Class #3	50
Class #4	200

Pseudo code for Class Imbalance Measure

```
Input: Class Imbalanced Dataset
Output: Class Distribution

Scanner = File.Open("Dataset File Path)
Map<Integer,Integer> classCounter = 0
WHILE(Scanner.hasNextDataPoint)
START
   dataPoint = Scanner.NextDataPoint()
   ClassIndex = dataPoint.getClassIndex()
   Count=   classCounter.get(ClassIndex).getPreviousCount(
)
   Count = Count + 1
   classCounter.replace(ClassIndex, Count)
END
Mean = 0.0
Sum = 0.0
FOR each classIndex in classCounter
   Sum = Sum + classCounter.getCount(classIndex)
END FOR
Mean = Sum / No.ofClasses
FOR each classIndex in classCounter
   Print(classCounter.get(classIndex).getCount( ))
END FOR
```

b. Strata Generation: The subpopulation of individual class records (table 2) separated for sample selection may be referred to as strata.

Table 2 Strata generation

Class	No.of Records	Strata #
Class #1	150	1
Class #2	600	2
Class #3	50	3
Class #4	200	4

Pseudo code for Stratification of Dataset

```
Input: Class Imbalanced Dataset
Output: Stratified Data

Scanner = File.Open("Dataset File Path)
Map<Integer,List<DataPoint>> classStratas
WHILE(Scanner.hasNextDataPoint)
START
   dataPoint = Scanner.NextDataPoint()
   ClassIndex = dataPoint.getClassIndex()
```

ISSN 2229-5518

```
   dataPointList = classStrata.get(ClassIndex)
   dataPointList.add(dataPoint)
END
```

c. Hybrid Sampling

i. *Simple Random Sampling with Replace (Oversampling):* One of the sampling technique used for oversampling is Simple Random Sampling with Replace. The strata's having records less than the mean value of a given dataset are selected and are sampled randomly in this module. A record which is selected as a sample for oversampling is again eligible for the process of resampling provided that, the record chosen should be same as the previous that belongs to the original training set and not the new training set. This condition is called "sampling with replacement".

Pseudo code for Random Oversampling

```
Input: Class Stratified Data having data points count less than Mean, Mean Value
Output: Oversampled Data
```

```
List<dataPoints> OverSampledList
SampleCount = DataPointSize
 WHILE SampleCount < Mean
START
   dataPointID = RandomNumberGenerator(0 to DataPointSize)
   dataPoint = DataPointsList.get(dataPointID)
   OverSampledList.add(dataPoint)
   SampleCount = SampleCount + 1
END
```

ii. *Stratified Random Sampling without Replace (Undersampling):* Stratified Random Sampling without Replace is one of the sampling technique used for the process of undersampling. This module clusters the data points from the majority class strata and picks the records randomly from different clusters, proportional to the size of the cluster.

E.g. For the given example, Mean value is 250 (Table 3), and class #2 is the majority class containing 600 records. Assume that the data points of this class are represented in 4 clusters of different sizes.

Cluster 1: 50 Data points
Cluster 2: 50 Data points
Cluster 3: 200 Data points
Cluster 4: 300 Data points

The data points of the majority class is distributed in 1:1:4:6 ratio in the clusters.

Required data points are 250.
Ratio Total = 1+1+4+6 = 12.
Minimum Records to be fetched from a cluster = 250/12 = 20
Now fetch,
20 records from Cluster 1
20 records from Cluster 2
80 records from Cluster 3
Remaining 250 - (20+20+80) = 250 – 120 = 130 records from cluster 4 (Larger Cluster)

Here Class #2 corresponds to the majority class and Class#3 corresponds to a minority class.

Table 3 Sampling Technique employed for each class

Class	No.of Records	No.of Records Inserted /Deleted	Sampling Technique used	No. of Records in each Class after sampling
Class #1	150	+100	Oversampling	250
Class #2	600	-350	Undersampling	250
Class #3	50	+200	Oversampling	250
Class #4	200	+50	Oversampling	250

Pseudo code for stratified Undersampling

```
Input: Class Stratified Data having data points count More than Mean, Mean Value
Output: Undersampled Data
```

```
List<dataPoints> UnderSampledList
ClusterList=KMeansClustering
(OriginalDataPointList , 4)
Map<ClusterID, Count> clusterSize
FOR each cluster in ClusterList
START
   ClusterSize.put (ClusterID, cluster.count( ) )
END
Ratio = CalculateRatio(ClusterList)
MinimumDataPoints = Mean / Sum(Ratios)
FOR each cluster in ClusterList
START
  No.ofDataPointsToFetch = MinimumDataPoints * Ratio[x]
UnderSampledList.add(Cluster[x].getRandomDataPoints
(No.ofDataPointsToFetch)
END
```

IV RESULTS

This section throws light on the results obtained from the analysis of proposed modules.

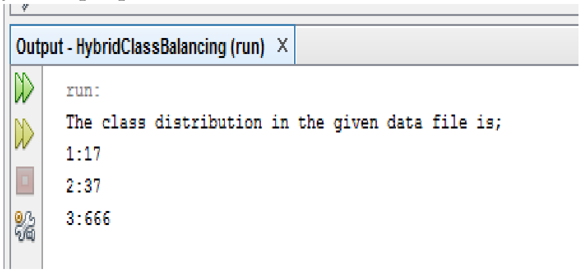

Fig 4 Class distribution in a given dataset

Fig 4 shows the class distribution of a Thyroid dataset obtained from UCI repository. The imbalance measure of each class is measured from the aggregated set of training records.

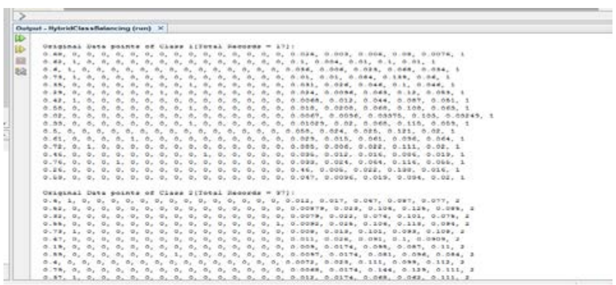

Fig 5 Strata containing records relevant to Class 1 and Class 2

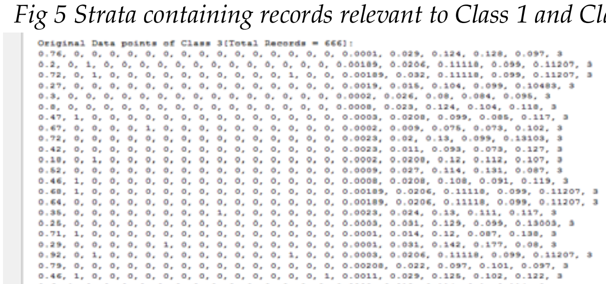

Fig 6 Strata containing records relevant to Class 3

Fig 5 and 6 represent the strata's with respect to class 1,2 and class 3 respectively

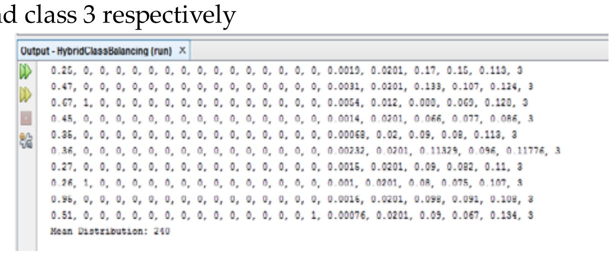

Fig 7 Mean distribution computed for all the 3 classes

Fig 7 shows the mean computation value for all the three classes. This mean distribution acts as a reference point for balancing the data distribution.

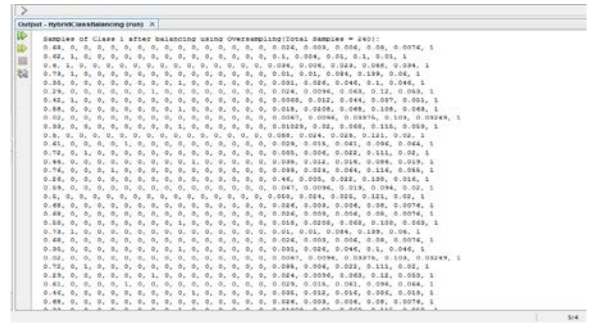

Fig 8 Application of oversampling technique over Class 1

Fig 8,9 shows the application of oversampling technique over class1 and class 2 respectively. Additional +223 records (240-17=223) are inserted to Class 1 to match the mean distribution. Similarly, the Distribution in Class 2 is altered by inserting additional +203 records (240-37=203).

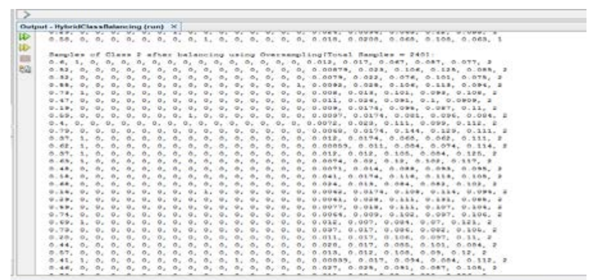

Fig 9 Application of oversampling technique over Class 2

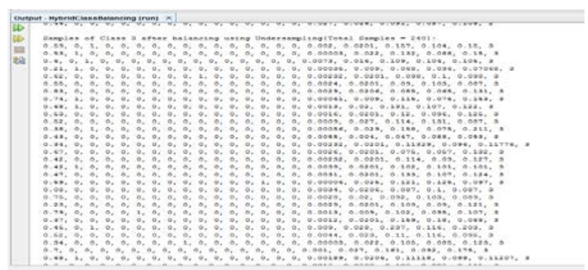

Fig 10 Application of undersampling technique over Class 3

As shown in fig 10, Class 3 consists of 666 records. K means Clustering method is used to alter the distribution of class 3 by deleting 426 records (666-240=426).

V. CONCLUSION

The data preprocessing technique called Hybrid Sampling technique has been proposed to generate balanced big data from multi class imbalanced big data. Sampling technique is applicable to the cases where one or more than one minority class is of interest. Mean is used as a reference point to sample all the class records without any cross check for balancing and multiple iterations. It reduces the processing time as all the records are balanced in a single stage of processing. Also, efficient sample selection strategy is proposed using K means Clustering method instead of random undersampling. A classifier that can eventually enhance the classification accuracy must be utilized on the proposed procedure to perform classification.

VI ACKNOWLEDGEMENT

I express my deep thanks to Dr. Prashanth C M, Professor &Head, Department of CS&E for warm hospitality and affection towards me. I thank the anonymous referees for their reviews that significantly improved the presentation of this paper. I am thankful to Mr. Rajeev Bilagi for his variable advice and support extended to me without whom I could not complete my paper. Words cannot express my gratitude for all those people who helped me directly or indirectly in my endeavor. I take this opportunity to express my sincere thanks to all the staff members of CS&E department of SCE for their valuable suggestion.

REFERENCES

[1] Reshma C.Bhagat, R.C., Sachin S. Patil " Enhanced SMOTE Algorithm for Classification of Imbalanced Big-Data using Random Forest". Advance Computing Conference (IACC), 2015 IEEE International, pp 403 - 408.

[2] Shaza M.Abd Elrahman and Ajith Abraham "A Review of Class Imbalance Problem". Journal of Network and Innovative Computing, Volume 1(2013) pp.332-340.

[3] Vaishali Ganganwar. " An overview of classification algorithms for imbalanced datasets" International Journal of Emerging Technology Vol.2, 4(2012).

[4] C.V. KrishnaVeni,T. Sobha Rani "On the Classification of Imbalanced Datasets", IJCST, Vol . 2, SP 1, December 2011.

[5] Mr. Rushi Longadge, Ms. Snehlata S. Dongre, Dr. Latesh Malik "Class Imbalance Problem in Data Mining: Review" International Journal of Computer Science and Network(IJCNSN), Volume 2, 2(2013).

[6] 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.

[7] Nitesh V. Chawla, Nathalie Japkowicz, Aleksander Ko lcz "Special Issue on Learning from Imbalanced Data Sets" Volume 6, Issue 1 - Page 1-6.

[8] 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.

[9] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). "SMOTE: synthetic minority over-sampling technique". Journal of artificial intelligence research, 16(1), 321-357.

[10] Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W., (2003). "SMOTEBoost: Improving prediction of the minority class in boosting". In Knowldge Discovery in Database: PKDD 2003 (pp. 107-119). Springer Belin Heidelberg.

[11] Park, B. J., Oh, S. K., & Pedrycz, W. (2013). "The design of polynomial function based neural network predicators for detection of software defects". Information Sciences, 229, 40-57.

[12] Takshak Desai., Udit Deshmukh, Prof. Kiran Bhowmick "Machine Learning for Classification of Imbalanced Big Data" International Journal on Recent and Innovation Trends in Computing and Communication(IJRITCC), October 2015, pp.6049 - 6053.

[13] Peng Liu, Lijun Cai, Yong Wang, Longbo Zhang "Classifying Skewed Data Streams Based on Reusing Data" International Conference on Computer Application and System Modeling (ICCASM 2010).

[14] Xinjian Guo, Yilong Yin1, Cailing Dong, Gongping Yang, Guangtong Zhou,"On the Class Imbalance Problem" Fourth International Conference on Natural Computation, 2008.

[15] Alexander Yun-chung Liu, B.S. " The Effect of Oversampling and Undersampling on Classifying Imbalanced Text Datasets" August 2004.

[16] Annarita D'Addabbo, Rosalia Maglietta. "Parallel selective sampling method for imbalanced and large data classification". Institute of Intelligent Systems for Automation - National Research Council, Volume 62. 5(2015)., pp 61-67.

[17] Hu, F., Li, H. (2013). "A novel boundary oversampling algorithm based on neighboured rough set model: NRSBoundary SMOTE". Mathematical Problems in Engineering, 2013.

[18] Lopez, V., Fernandez, A., Garcia, S., Palade, V., & Herrera, F. (2014). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences, 2013.,Vol. 250, pp 113-141.

[19] Donghui Shi., Hefei, Jian Guan., Zurada, J. (2015)" Cost Sensitive Learning for Imbalanced Bad Debt Datasets in Healthcare Industry". Computer Aided System Engineering (APCASE)., IEEE, 7(2015): pp 30 - 35 .

[20] Rio, S., Lopez, V., Benitez, J. M., & Herrera, F. On the use of MapReduce for imbalanced big data using Random Forest. Information Sciences, Vol. 285, pp 112-137, 11(2014).

[21] Hui Han, Wen Yuan Wang, Bing Huan Mao. "Borderline SMOTE: A New OverSampling Method in Imbalanced Data Sets Learning" Springer Berlin Heidelberg., Vol. 3644, pp 878-887 (2005)

[22] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano "RUSBoost: A Hybrid Approach to Alleviating Class Imbalance", IEEE Transactions On

ISSN 2229-5518
Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol. 40, No. 1, IEEE January 2010: 1083-4427,pp185-7.