Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 7 – No. 36, May 2021 – www.caeaccess.org

Classification of Imbalanced Data of Medical Diagnosis using Sampling Techniques

Varsha Babar
Assistant Professor
Department of Computer Engineering
Dr. D. Y. Patil School of Engineering and Technology, Pune

ABSTRACT
When there is gigantic difference between the ratio of two classes in the classification algorithms, then the classifier may tend to favor the instances of majority class whereas, it becomes difficult for the classifier to learn the minority class samples. Either, undersampling is used or oversampling is used for this imbalance but, most of the undersampling techniques does not consider distribution of information among the classes while the oversampling technique leads overfitting of the trained model. So, to resolve this issue integration of undersampling as well as oversampling technique can be done. Majority class samples can be undersampled using a new approach, namely, MLP-based undersampling technique (MLPUS). Majority Weighted Minority Oversampling Technique (MWMOTE) can be used for generating the synthetic samples for minority class. The main objective is to handle the imbalance classification problem occurring in the medical diagnosis of rare diseases and combines the benefits of both undersampling and oversampling Experiments are performed on 7 real world data sets for the evaluation of proposed framework's performance.

Keywords
Sampling technique, Imbalance Data, MLPUS, MWMOTE, Ensemble Technique

1. INTRODUCTION
NOWADAYS, in many real-world applications, data sets are imbalanced in nature. Imbalanced data sets are nothing but one class containing much more samples and another contains very little. The class having more (negative) instances is called as a majority class while class containing little (positive) instances is called as minority class. Most of the machine learning algorithms performs better when data sets are almost balanced. But problem arises when given data sets are very much imbalanced in nature. Classification of these imbalanced data sets is a very critical and a challenging task for the classifier as classifier may tend to favor the instances of majority class.

Due to unequal distribution of data, majority class significantly dominates the minority class. The ratio of imbalance can be in the order of 100:1, 1000:1 or 10,000:1. This form of imbalance is known as between- class imbalance. When there is more than one concept in classes and some of them are rarer than others, then such imbalance is called as within class imbalance. In this paper, the proposed system is dealing with between class imbalances only. In to-day's era of machine learning, many data mining applications such as medical diagnosis [1], detection of oil spills in radar im-ages [2], information retrieval systems [3], helicopter fault monitoring [4], data mining from direct marketing [5] facing this imbalanced learning problem. It is necessary to overcome this imbalanced learning problem as it affects the performance of a classifier very greatly. In the medical diagnosis of rare diseases, the data samples for heart diseases are very much less than data samples of non-heart diseases. After classification classifier may misclassify some of the heart disease samples as a non-heart disease sample. The consequence of this misclassification can be very much costlier than misclassifying non heart disease sample as a heart disease sample.

The imbalanced data can be approximately balanced by either undersampling i.e., reducing samples from majority class, or by oversampling i.e., adding samples to the minority class. There are various undersampling techniques which try to resolve imbalance learning problem. Along with undersampling, various oversampling methods such as SMOTE, Borderline SMOTE, ADASYN, RAMOBoost uses synthetic sample generation for balancing the distribution among classes. But in many scenarios, these methods may generate the wrong synthetic minority samples and make learning tasks harder. To this end, a method can be proposed which will integrate undersampling and over-sampling. Majority class samples can be undersampled using a new approach, namely, MLPUS [6]. Majority Weighted Minority Oversampling Technique (MWMOTE) [7] can be used for generating the synthetic samples for minority class. The main objective is to handle the imbalance classification problem occurring in the medical diagnosis of rare diseases and combines the benefits of both undersampling and oversampling.

The remainder of this paper is divided into five sections. In section 2, a brief review of existing works on imbalanced learning domain is provided. Proposed method and its components are described in section 3. The experimental study and simulation results are presented in Section 4. Finally, section 5 concludes the paper with some future research directions.

2. RELATED WORK
To deal with the imbalanced learning problem significant works have been done which can be categorized as a: sampling-based methods [8], cost-based methods [9], active learning-based methods [10] and kernel-based methods [11]. Though there is no single method which handles the imbalance optimally, sampling-based methods have been shown to be very successful nowadays. This section provides a brief review of the works performed in this category only. Details of works performed in other categories can be found in [12]. Sampling based methods focus on altering the size of either majority or minority class in order to balance between

them. Various sampling techniques exist in the literature, they are under-sampling and oversampling. Reduction of samples from majority class is done in under sampling. These methods again categorized into random and informed undersampling. Random under-sampling technique randomly removes the samples from majority class, but it may lead to loss of important samples also. To overcome this issue, researchers proposed some informed undersampling techniques such as EasyEnsemble, Balance cascade [13], and KNN based methods, namely Near miss 1, Near miss 2, Near miss 3 and most distant method [14]. One sided selection method also performs well to deal with imbalanced data [15]. To further refine this method ClusterOSS has been proposed [16].

In oversampling techniques, artificial samples are added to the minority class to balance between classes. Oversampling can be random or synthetic sample generation. In random oversampling, samples are randomly replicated, but which can lead to over fitting [17]. On the other hand, in synthetic oversampling method, it generates the synthetic samples to minority class. These generated samples add essential information to the minority class, resulting in improved performance of the classifier. In [18] proposed a powerful method, namely synthetic minority oversampling technique (SMOTE) which has been shown a great success in many applications. Initially for each minority sample k-nearest neighbors are determined. Then synthetic sample is generated along the line segment joining minority sample and its nearest neighbor. Firstly, SMOTE takes the difference between minority sample and its nearest neighbor. This difference is then multiplied by a random number between 0 and 1, and adds this to original minority sample. In this way synthetic sample is generated. SMOTE generates an equal number of synthetic samples for each minority sample. Several modifications are made to the SMOTE, e.g., borderline-SMOTE [19], safe-level SMOTE [20], local neighborhood-based SMOTE [21], and rough sets theory based SMOTE [22]. To handle the imbalanced learning problem in big data a novel approach, namely, the Enhanced SMOTE algorithm has been proposed in [23]. Haibo He, E.A. Garcia, pro-pose a novel approach adaptive synthetic sampling (ADASYN) to handle imbalanced data set. In synthetic sample generation process, there is no need to consider all minority samples as there may be problem of overlapping [24]. In [25] a novel method namely RAMOBoost has been proposed which systematically generates synthetic samples depending on sampling weights. It adjusts these weights of minority samples according to their distribution.

Oversampling methods improve performance of the classifier and a lot more useful than undersampling. However, both under-sampling and oversampling can work efficiently. Remainder categories of imbalanced learning methods also perform well, but there is no single best method for all scenarios. While comparing undersampling and oversampling, one observation favoring oversampling is that undersampling may lose essential information while oversampling does not.

3. PROPOSED FRAMEWORK

The goal of the proposed system is to handle the imbalanced learning problem arises in the medical diagnosis of rare diseases. As shown in Fig.1 it will take imbalanced data of rare diseases as an input and produces balanced data. This system will contain two major subsystems. One is used for reducing the samples from majority class and another is used to increase the number of samples in the minority class. Both these subsystems produce different balanced data sets. The Number of classifiers is then applied to classify the data, and these classifiers again ensembled to improve the accuracy of classification. In this way data sets can be balanced.

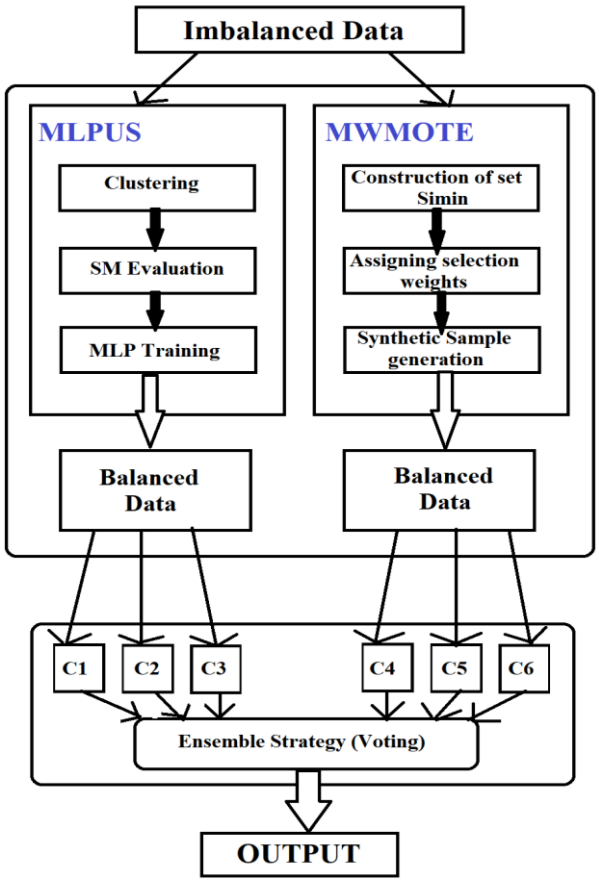

Figure 1: Overall flow of a System

3.1 MLPUS

Figure. 2 shows the workflow of the MLPUS.The MLPUS consists of three major components:

1)Clustering of samples in the majority class using k-means clustering algorithm.

2)Finding most important samples(undersampling) using the stochastic sensitivity measure (SM).

3)A MLP trained by using training samples selected by the SM.

This method selects a sample which is located closest to the center of each of these clusters as representative samples and then their SM values are calculated. The k samples having the largest SM values will be selected from the majority class.

Similarly, k samples having the largest SM values will also be selected from the minority class. These samples with largest SM value are then added to the initial training set to form a balanced training data set for the MLP. In each turn of iteration, there are 2tk samples in the training data set, where t is the number of iterations. The value of t is at most equal to k. Selected samples are removed from the candidate set and these steps repeat until the number of samples in the minority class is less than k.

Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 7 – No. 36, May 2021 – www.caeaccess.org

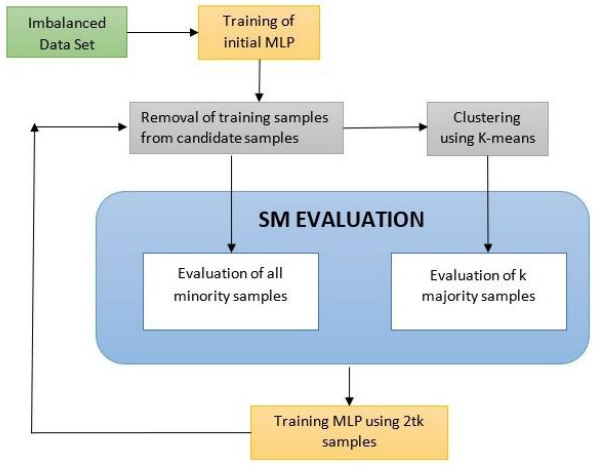

Figure 2: Workflow of MLPUS

3.2 MWMOTE

The objective of MWMOTE is twofold: to improve the sample selection process and to improve the synthetic sample generation process. MWMOTE involves three key phases. In the first phase, MWMOTE identifies hard-to-learn and the most important minority class samples from the original minority set, S_{min} and construct a set, S_{imin} by the identified samples. In the second phase, each member of S_{imin} is given a selection weight, S_w, according to its importance in the data. In the third phase, using the clustering approach, MWMOTE generates the synthetic samples from S_{imin} using Sws and produces the output set, by adding the synthetic samples to S_{min}.

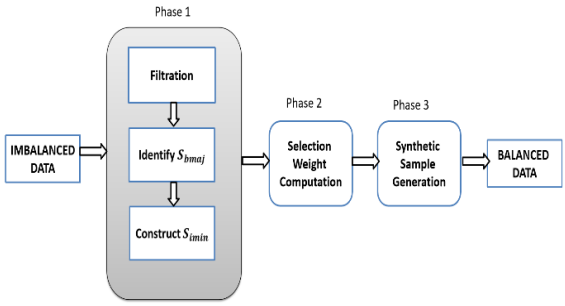

Figure 3: Workflow of MWMOTE

3.3 Ensemble Technique

There are many ensemble techniques available, but here the system will use majority voting *method* to ensemble the classifiers Majority (plurality) voting:

$$\sum_{t=1}^{T} d_{t,J(x)} = max_{j=1,\ldots c} \sum_{t=1}^{T} d_{t,j}$$

Under the condition that the classifier outputs are independent, it can be shown the majority voting combination will always lead to a performance improvement. If there are a total of T classifiers for a two-class problem, the ensemble decision will be correct if at least [T/2+1] classifier choose the correct class

3.4 Mathematics Relevant to the System

Notations:

S_{maj} : Set of majority samples

S_{min} : Set of minority samples

N_p: Candidate samples in minority class

S_{bmaj} : Filtered minority set

S_{bmaj} : Borderline majority set

S_{imin}: Identified minority set

S_w: Selection weight

Problem Description:

Let S be the system,

S= {S_{maj}, S_{min}}

Where, S_{maj}>S_{min}

Activity 1: Undersampling (MLPUS)

Step 1: Training the initial MLP

a. Cluster both S_{maj} and S_{min} into $k = \lfloor \sqrt{N_p} \rfloor$ clusters each.

b. Set both P_0 and R_0 to be empty sets.

c. For each of k clusters of the minority class, add the sample located closest to its center to P_0

d. For each of k clusters of the majority class, add the sample located closest to its center to R_0

e. $S_{min} = S_{min} - P_0$, $S_{maj} = S_{maj} - R_0$, S = $P_0 \cup R_0$ and b = 0

Step 2: Find most important samples from S_{maj} and add them to set C.

Step 3: Compute the value of SM for each sample of C and S_{min}.

Step 4: Add k samples from C and S_{min} having largest SM value to set P_b and R_b respectively.

Step 5: S = S ∪ P_b ∪ R_b

Step 6: Train a MLP using S.

Activity 2: Oversampling (MWMOTE)

Step 1: Construction of set S_{imin}

1. Construct S_{minf}, as

   $S_{minf} = S_{min}$ - { $x_i \in S_{min}$: NN (x_i) contains no minority example}

2. For each $x_i \in S_{minf}$, compute $N_{maj}(x_i)$

3. $S_{bmaj} = \bigcup_{x_i \in S_{minf}} N_{maj}(x_i)$

4. For each $y_i \in S_{bmaj}$, compute $N_{min}(y_i)$

5. $S_{imin} = \bigcup_{y_i \in S_{bmaj}} N_{min}(y_i)$

Step 2: Finding Selection weights

6. For each $y_i \in S_{bmaj}$ and for each $x_i \in S_{imin}$, compute $I_w(y_i, x_i)$.

7. For each $x_i \in S_{imin}$, compute $S_w(x_i) = \sum_{y_i \in S_{bmaj}} I_w(y_i, x_i)$

8. Convert each $S_w(x_i)$ into selection probability $S_p(x_i)$ such as

$$S_p(x_i) = S_w(x_i) / \sum_{z_i \in S_{imin}} S_w(z_i)$$

Step 3: Generating Synthetic Samples

Let
$L_1, L_2, L_3, \ldots.. L_M$ be the clusters of S_{min}
Initialize $S_{omin} = S_{min}$

9. Generate Synthetic sample using

$$s = x + \alpha \times (y - x)$$
$$S_{omin} = S_{omin} \cup \{s\}$$

Activity 3: Classification

- Classify the balanced data using CART
- Ensemble of classifiers using Majority voting

$$\sum_{t=1}^{T} d_{t,J(x)} = max_{j=1,\ldots.c} \sum_{t=1}^{T} d_{t,j}$$

4. EXPERIMENTS AND RESULTS

This section presents the performance of proposed framework on 7 real-world data sets collected from the UCI repository. Here, for this system, imbalanced data sets of rare diseases are taken as an input. Table 1. shows characteristics of these data sets in the form of a number of attributes, number of minority and majority samples and imbalance ratio. All data sets are in a binary form. These data sets are selected in such a way that they have a different number of samples, attributes and imbalance ratio.

Table 1: Characteristics of data sets

Data set	No. of Attributes	Majority Samples	Minority Samples	Imbalance Ratio
Pima Diabetes	9	499	120	0.65:0.35
Breast Cancer	10	201	85	0.65:0.35
Hepatitis	20	123	32	0.79:0.21
Ionosphere	35	225	126	0.64:0.36
Mammo-Graphic	6	516	445	0.54:0.46
Liver Disorder	7	200	145	0.57:043
Spect Heart	23	3541	231	0.939:0.061

To integrate MLPUS and MWMOTE, the balance data sets derived from these 2 algorithms are used. As MLPUS processes majority samples to undersample the instances, in combine algorithm majority class is taken from MLPUS. Likewise, minority class is taken from MWMOTE as this technique generates synthetic samples to the minority class.

Table 2: Comparison in terms of accuracy

Dataset	MLPUS	MWMOTE	Combined Algorithm
Pima Diabetes	90.83	82.54	98.54
Breast Cancer	98.26	96.73	90.72
Hepatitis	97.14	98.91	98.76
Ionosphere	98.54	99.14	97.47
Mammographic	81.17	78.61	88.54
Liver Disorder	87.34	74.88	96.62
Spect Heart	97.35	94.36	84.97

Finally, this new data set is classified using voting technique. For voting technique 3 classifiers namely, Multi-layer Perceptron (MLP), Classification and Regression Tree (CART) and Support Vector Machine (SVM) are used. MLPUS, MWMOTE and combined algorithm is compared with each other in terms of accuracy, time required and number of instances. Table 2 shows this comparison in terms of accuracy.

As MLPUS does undersampling, it does not consider all samples in data set. Likewise, MWMOTE adds extra samples to the original data set. Figure 5 shows comparison of these techniques in terms of number of samples.

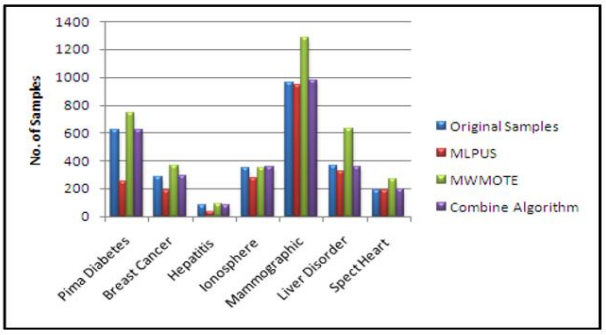

Figure 4: Comparison in terms of no of samples

From these observations, it is clear that MLPUS always take less samples than original. Similarly, MWMOTE uses more samples than original, but in case of combine algorithm, number of samples used for this technique are almost equals to the original number of samples.

There is another parameter on which performance of techniques can be evaluated which is nothing but Receiver Operating Characteristics (ROC) curve. ROC graph can be obtained by plotting the false positive rate on the X-axis and the true positive rate on the Y-axis, where,

$$FPR = \frac{FP}{n},\ TPR = \frac{TP}{n},$$

Figure 5 depicts the ROC curve for MLPUS, MWMOTE and COMBINE algorithm. If the curve is closer to the left border and top of the border, then more accurate results of classification. If the curve is closer to 45-degree diagonal then less accurate results. From the figure 6 it is clear that ROC of combine algorithm is closer to left and upper border as compared to MLPUS and MWMOTE. Hence, if ROC is considered then also combine algorithm performs well.

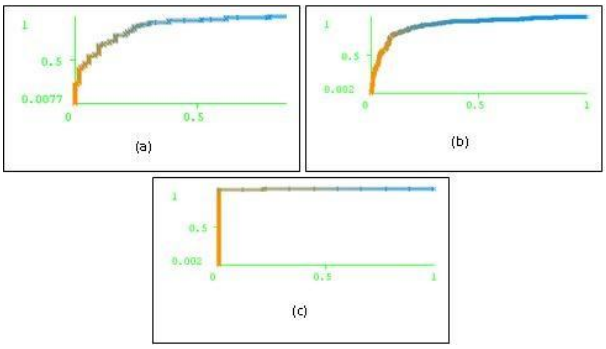

Figure 5: (a) ROC for MLPUS (b) ROC for MWMOTE (c) ROC for Combined Algorithm

5. CONCLUSION AND FUTURE SCOPE

In the proposed system, in order to handle the imbalanced data, undersampling as well as oversampling is used together. This system tries to combine the benefits of both sampling techniques. MLPUS is used for undersampling, which preserves the distribution information of the majority class, and selects informative samples from both classes. The oversampling method not selects the most important minority class samples effectively and as-signs them appropriate weights. Furthermore, it is able to generate correct synthetic samples. These both techniques perform well separately, but if they are combined using ensembling of classifier, then it will improve the classification results greatly. Several other research issues are left to be considered. This system can be generalized to handle imbalanced problem occurring in multiclass data sets. In MLPUS, k-means algorithm is used for clustering, which could be replaced by other clustering methods. Moreover, when the data changes across time, incremental learning needed.

6. REFERENCES

[1] P.M. Murphy and D.W. Aha, UCI Repository of Machine Learning Databases, Dept. of Information and Computer Science, Univ. of California, Irvine, CA, 1994.

[2] M. Kubat, R.C. Holte, and S. Matwin, Machine Learning for the Detection of Oil Spills in Satellite Radar Images, Machine Learning, vol. 30, no. 2/3, pp. 195-215, 1998.

[3] D. Lewis and J. Catlett, Heterogeneous Uncertainty Sampling for Supervised Learning, Proc. Intl Conf. Machine Learning, pp. 148- 156, 1994.

[4] N. Japkowicz, C. Myers, and M. Gluck, A Novelty Detec-tion Approach to Classification, Proc. 14th Joint Conf. Ar-tificial Intelligence, pp. 518-523, 1995.

[5] C.X. Ling and C. Li, Data Mining for Direct Marketing: Problems and Solutions, Proc. Intl Conf. Knowledge Discovery and Data Mining, pp. 73-79, 1998.

[6] Wing W. Y. Ng, Junjie Hu, Daniel S. Yeung, Shaohua Yin, and Fabio Roli, Diversified Sensitivity-Based Undersam-pling for Imbalance Classification Problems, IEEE Trans. Cybernetics vol. 45, no. 11, Nov. 2015.

[7] Sukarna Barua, Md. Monirul Islam,Xin Yao, MWMOTEMajority Weighted Minority Oversampling Technique for imbalanced data set learning, IEEE Trans. Knowledge and data engineering, vol. 26, no. 2, February 2014

[8] Varsha S. Babar, Roshani Ade, A Review on Imbalanced Learning Methods, International Journal of Computer Ap-plications (IJCA) 0975 8887 , Dec 2015.

[9] Xingyi LIUi, -sensitive Decision Tree with Missing Values and Multiple Cost Scales,Intl Joint Conf. on Artificial In-telligence, 2009.

[10] Jing Zhang, XindongWu and Victor S. Sheng, Active Learning with Imbalanced Multiple Noisy Labeling, IEEE Trans. on Cybernetics, vol. 45, no. 5, May 2015.

[11] ZhiQiang ZENG and ShunZhi ZHU, A Kernel-based Sam-pling to Train SVM with Imbalanced Data Set, Conference Anthology, IEEE, January 2013.

[12] H. He and E.A. Garcia, Learning from Imbalanced Data, IEEE Trans. Knowledge Data Eng., vol. 21, no. 9, pp. 1263-1284, Sept. 2009.

[13] X.Y. Liu, J.Wu, and Z.H. Zhou, Exploratory Under Sampling for Class Imbalance Learning, Proc. Intl Conf. Data Mining, pp. 965- 969, 2006.

[14] J. Zhang and I. Mani, KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extrac-tion, Proc. Intl Conf. Machine Learning,Workshop Learn-ing from Imbalanced Data Sets, 2003.

[15] M. Kubat and S. Matwin, Addressing the Curse of Imbalanced Training Sets: One-Sided Selection, Proc. Intl Conf. Machine Learning, pp. 179-186, 1997.

[16] Victor H. Barella, Eduardo p. Costa, and Andre C P L F Carvalho, ClusterOSS: a new undersampling method for imbalanced learning.

[17] H.He, Self-Adaptive Systems for Machine Intelligence,Wiley, Aug 2011

[18] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyr, SMOTE: Synthetic Minority oversampling Technique,J. Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[19] H. Han, W.Y. Wang, and B.H. Mao, Borderline-SMOTE: A New Oversampling Method in Imbalanced Data Sets Learning, Proc. Intl Conf. Intelligent Computing, pp. 878-887, 2005.

[20] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, Safe-level-SMOTE: Safe level-synthetic minority over-sampling technique for handling the class imbalanced problem, in Advances in Knowledge Discovery and Data Mining. Berlin, Germany: Springer, 2009, pp. 475 482, 2009.

[21] T. Maciejewski and J. Stefanowski, Local neighbourhood extension of SMOTE for mining imbalanced data, in Proc. IEEE Symp. Comput. Intell. Data Min. (CIDM), Paris, France, pp. 104111, 2011.

[22] E. Ramentol, Y. Caballero, R. Bello, and F. Herrera, SMOTE-RSB*: A hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory, Knowl. Inf. Syst., vol. 33, no. 2, pp. 245265, 2012.

[23] Reshma C. Bhagat and Sachin S. Patil, Enhanced SMOTE Algorithm or Classification of Imbalanced Big-

Data using Random Forest, IEEE International Advance Comput-ing Conference (IACC), 2015.

[24] H. He, Y. Bai, E.A. Garcia, and S. Li, ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning, Proc. Intl Joint Conf. Neural Networks, pp. 1322-1328, 2008.

[25] S. Chen, H. He, and E.A. Garcia, RAMOBoost: Ranked Minority Oversampling in Boosting, IEEE Trans. Neural Networks, vol. 21, no. 20, pp. 1624-1642.