Mathematical Foundations of Computing
©American Institute of Mathematical Sciences
Volume **3**, Number **1**, February **2020**
doi:10.3934/mfc.2020001
pp. **1–9**

A NOVEL APPROACH FOR SOLVING SKEWED CLASSIFICATION PROBLEM USING CLUSTER BASED ENSEMBLE METHOD

Gillala Rekha*
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Guntur
Andhra Pradesh, India – 522502

V Krishna Reddy and Amit Kumar Tyagi
Koneru Lakshmaiah Education Foundation
Vaddeswaram, Guntur
Andhra Pradesh, India – 522502
Vellore Institute of Technology
Chennai Campus, Chennai, 600127
Tamilnadu, India

(Communicated by Song Wang)

Abstract. In numerous real-world applications, the class imbalance problem is prevalent. When training samples of one class immensely outnumber samples of the other classes, the traditional machine learning algorithms show bias towards the majority class (a class with more number of samples) lead to significant losses of model performance. Several techniques have been proposed to handle the problem of class imbalance, including data sampling and boosting. In this paper, we present a cluster-based oversampling with boosting algorithm (Cluster+Boost) for learning from imbalanced data. We evaluate the performance of the proposed approach with state-of-the-art methods based on ensemble learning like AdaBoost, RUSBoost and SMOTEBoost. We conducted experiments on 22 data sets with various imbalance ratios. The experimental results are promising and provide an alternative approach for improving the performance of the classifier when learned on highly imbalanced data sets.

1. **Introduction.** In machine learning, creating an effective learning model can be challenging, if the training data set used to train the model is highly imbalanced. When samples of one class greatly outnumber the samples of the other classes, traditional data mining algorithms trained on such data will result in fall-off the classification accuracy. These models will be unsuccessful in identifying samples of the minority class. The real-world application scenario such as fault diagnosis, medical diagnosis [8, 14] and recommendation systems[13] suffer from imbalanced class distribution. Hence over the years, researchers proposed new techniques to address the problem of class imbalance [17] [18] [3] [15] [9]. Broadly, the solutions

2010 *Mathematics Subject Classification.* Primary: 58F15, 58F17; Secondary: 53C35.
Key words and phrases. Class Imbalance, Boosting, Binary classification, Sampling, Ensemble methods.
The first author is supported by KL University.
* Corresponding author: Gillala Rekha.

for addressing class imbalance problems are classified into three categories, namely data level, algorithm level, and hybrid methods. In data level techniques, the problem of class imbalance is handled by pre-processing the data either by using data sampling or synthetic generation method to establish an equal distribution between the classes. In algorithm level techniques, the existing algorithms are modified, or new algorithms are proposed to handle the class imbalance problems. Hybrid methods are also known as ensemble methods, which combine the data-level technique with algorithm-level techniques. Most recent advancement combines traditional ensemble methods such as bagging and boosting with data sampling techniques to improve the performance of the classifier when trained on class imbalance problems.

Data sampling techniques balance the skewed distribution in the training data set by either adding samples to the minority class called an oversampling technique or discarding samples from the majority class called an undersampling technique. Several techniques are proposed for performing undersampling and oversampling. The simplest of it is random over-sampling and random undersampling methods. In Random OverSampling (ROS) duplicate samples of minority class are generated randomly until an expected class ratio is achieved. Likewise, Random UnderSampling (RUS) discard the samples from the majority class randomly until the balanced dataset is achieved. Oversampling and undersampling have their advantages and disadvantages. The advantage of undersampling is decreasing the time needed to train the classification model since the size of the training data is reduced. However, the drawback is the loss of information due to deletion of samples from the training data. On the other hand, oversampling techniques do not result in loss of information, as very original training data appears in the resampled training data. But the main drawback of using oversampling is it may lead to overfitting [8] and also increases in the time of the learning algorithm when trained on the oversampled data sets. To overcome the limitation of ROS, a variant of data level technique have been proposed such as Synthetic Minority Over-sampling Technique (SMOTE) [6] and Adaptive Synthetic Sampling Approach (ADASYN) [10].

To improve the performance of the learning models, the boosting technique is used irrespective of whether the data are imbalanced. AdaBoost [8] is the most commonly used boosting algorithm, which builds an ensemble method iteratively. Initially, each instance of a training data set is assigned with equal weight. The weights of the instance are modified for each iteration with the goal of classifying the samples correctly in the next iteration. The weights for each instance is adjusted based on how they were classified. If the instance is misclassified, then its weight is increased else its weight is decreased. Upon completion, all classifier model takes part in a weighted vote to classify the test or unlabelled instances. Such a technique works effectively for the imbalanced dataset as most of the minority samples likely to be misclassified and therefore assigned higher weights in the next iterations. However, imbalanced data distribution does not hamper the learning task by itself [21] but a series of difficulties turn up related to this problem such as small sample size, class overlapping, and small disjuncts. Data-level methods face difficulty to handle the different characteristics of imbalanced class distribution such as overlapping, small disjunct, and small sample size. The difficulties of the classifier learning task are not directly caused by skewed distributions [21] but usually due to the existence of small sample size, small disjuncts or class overlapping, exist in the skewed class distribution [4]–[7]. To overcome such issues, newer data generation methods

based on clustering approach have been proposed. Therefore, in this paper we utilize clustering-based oversampling with boosting, to oversample the training data using clustering technique and train the data set using boosting. We present a novel hybrid approach based on cluster oversampling with the ensemble approach, to improve the performance of the classification models trained on imbalanced dataset. We compare the performance of the proposed model to that of AdabBoost, RUSBoost, and SMOTEBoost algorithms which combines data sampling with boosting technique. The main motivation for introducing cluster oversampling with boosting (Cluster+Boost) is to handle small disjunct exist in imbalanced data sets, which degrade the performance of the standard classification algorithm.

The remainder of the paper is organized as follows. Section 2 reviews current literature work. Section 3 introduces the proposed approach. Then, Section 4 presents the details of our experiments and Section 5 provides the experimental results. Finally, we concluded this paper in Section 6.

2. **Literature review.** In the last decade, much research has been performed to address the class imbalance problem. Ali [1] provides a review on issues that come with learning from imbalanced data-sets. The study identifies various existing approaches for handling class imbalance problems, including data sampling and ensemble techniques, which are considered in this paper. Data level approaches also called pre-processing approaches to handle class imbalance problems can be divided into re-sampling methods and synthetic data generation methods. Re-sampling techniques to balance the class imbalance data includes undersampling the majority class, oversampling the minority class or both. On the other hand, the synthetic data generation method generates new data instances. One of the most popular and commonly used synthetic data generation methods is SMOTE [6]. This method creates new data instances by randomly selecting minority data instances and finding one of its nearest neighbors. To improve the selection process of data instances, a variation of this method has been proposed in the literature like MSMOTE [11], Borderline-SMOTE [12]. Modified SMOTE(MSMOTE) generates synthetic instances by applying a different strategy for selecting its near neighbors according to the type of samples. In Borderline-SMOTE only the minority instances near the borderline are over-sampled to generate the synthetic data instances. The main drawback of SMOTE is it randomly synthesizes the minority instances along a line joining a minority instance and its selected nearest neighbors, ignoring nearby majority instances. Bunkhumpornpat et.al [3] proposed a Safe Level-SMOTE method, samples minority instances along the same line with a different degree of weight called safe level. It computes by using nearest neighbor minority instances.

Barua et al. [4] proposed a majority weighted minority oversampling technique (MWMOTE) to efficiently handle class imbalance problems. Initially, MWMOTE identifies the hard-to-learn minority class instance and assigns weights based on their Euclidean distance from the nearest majority class instance. Afterward, it generates the synthetic samples using a clustering approach in a way that all the generated samples lie inside some minority class clusters. Rayhan et al. [19] proposed clustering-based under-sampling with boosting called CUSBoost. The method separates the majority and minority class instances and then clusters the majority class samples into k-clusters using the k-means clustering technique. After that, they applied random undersampling to each cluster to discard the samples in each cluster to be of same size that of minority samples. Then, AdaBoost algorithms have been

applied to train on the balanced dataset. Lin et al. [16] proposed a cluster-based undersampling technique to address the class imbalance problems.

In machine learning, the use of ensemble methods is known to increase the accuracy over a single classifier. The majority of ensemble learning algorithms have been designed specifically for handling class imbalance problems. The modification of the ensemble learning algorithm to handle class imbalance problems usually includes data level approaches to preprocess the data before learning each classifier. However, certain algorithms consider the inserting of cost-sensitive approach in the ensemble learning process. In general, algorithm level and cost-sensitive approaches are more dependent on the problem, whereas data level and ensemble learning methods are more versatile since they can be used independently of the base classifier. However, due to the focus of this paper, only data-level ensemble methods will be reviewed. Ensemble methods at data-level are classified based on bagging or boosting techniques such as SMOTEBagging [24] , SMOTEBoosting [7], and RUSBoost [22]. These techniques incorporated undersampling and SMOTE into bagging and boosting ensemble methods.

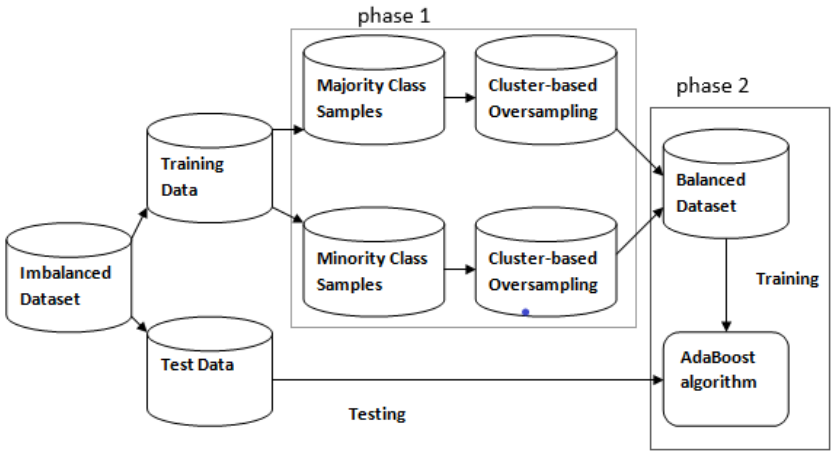

FIGURE 1. Framework of Cluster-based Oversampling with Boosting

3. **Cluster-based oversampling with boosting method (cluster+boost).** Figure 1 presents the proposed approach (Cluster+Boost) combines synthetic data generation with an ensemble algorithm. It is similar to SMOTEBoost with the variation occurring in the sampling approach. SMOTEBoost algorithm uses the SMOTE method to generate the synthetic data for minority class, while RUSBoost uses a random undersampling method to discard the samples in the majority class. In contrast, the proposed method works by applying a clustering technique with ensemble learning. The proposed method consists of 2 phases: (i) Synthetic data generation phase and (ii) Classification phase.

3.1. **Synthetic data generation phase.** In this phase, the skewed class distribution is divided into majority class and minority class instances. Afterward, we cluster the training data of each class separately and then perform oversampling using SMOTE cluster by cluster to generate the synthetic data. This method was

previously applied to several domains like text classification and letter recognition [21].

The main idea is to consider not only the imbalance data occurring between two classes, i.e., between-class imbalance but also the imbalance occurring between the subclusters of each class. i.e., within-class imbalance. The main motivation behind our approach is to remove these two types of imbalances at the same time. Before generating the synthetic data using SMOTE, the training samples in a majority and minority classes must be grouped into clusters. In this study, the K-means algorithms were adopted as a clustering algorithm. The K-means algorithm works by selecting K- training samples randomly as representative for each cluster. The input vector of these representative samples represents the mean of each cluster. The other training samples are processed one by one to calculate the distance between it and the cluster centers. The sample is attributed to the cluster closest to it. The clusters then update it mean by averaging the input vectors of all its corresponding samples. A study was conducted to fixed the value of K with different numbers varying from 1 to 10. Finally, K was assigned with 5 based on the sensitivity study.

Once the training samples of each class have been clustered, the oversampling starts using the SMOTE technique. All the clusters of majority classes, except the cluster with the largest samples, are oversampled using SMOTE to get the same number of samples as the largest majority class cluster. In the minority class, each cluster is oversampled using SMOTE until each cluster contains a large class size that of majority samples. The number of synthetic samples to be generated for each minority class will depend on the size of the large cluster divided by the number of minority clusters.

This process of generating synthetic data for both majority and minority clusters are bounded by the largest majority cluster.

- Let *majsize* be the overall size of the largest cluster.
- In each of the majority clusters, except *majsize* are oversampled using SMOTE until each cluster contains *majsize*
- Each cluster of minority classes are oversampled using SMOTE until each cluster contains $majsize/Nminsize$ where $Nminsize$ represents the number of subclusters in the minority class.

Our proposed method strength lies in overriding between class and with-in class imbalance, by oversampling both the classes.

3.2. **Classification phase.** In the classification phase, the ensemble learning approach has been proposed to address the class imbalance problem. Adaptive boosting algorithm (AdaBoost) is an iterative boosting algorithm constructing a strong classifier as a linear combination of weak classifiers [8]. It considers the whole dataset to train each classifier in sequence, but after each round, it gives more focus to difficult instances, with the goal of correctly classifying samples in the next iteration that were incorrectly classified during the current iteration. Hence, it gives more focus to samples that are harder to classify, the quantity of focus is measured by a weight, which initially is equal for all instances. After each iteration, the weights of misclassified instances are increased. On the contrary, the weights of correctly classified instances are decreased. Furthermore, another weight is assigned to each classifier depending on its overall accuracy which is then used in the test phase; more confidence is given to more accurate classifiers. Finally, when a new instance

is submitted, each classifier gives a weighted vote, and the class label is selected by the majority.

4. **Experimental results, analysis and discussion.** In this section, we present the experimental analysis to examine the performance of our proposed approach. We conducted experiments on 22 imbalance data sets from KEEL-data repository [2] with different imbalance ratio and results are noted. Table 1 shows the details of data sets, including the number of samples, the number of attributes and their Imbalance Ratio (IR).

TABLE 1. Dataset Characteristics

Datasets	Size	# attr	% IR
ecoli-0_vs_1	220	7	1.82
ecoli1	336	7	3.36
ecoli2	336	7	5.46
ecoli3	336	7	8.6
glass0	214	9	2.06
glass-0-1-2-3_vs_4-5-6	214	9	3.2
glass1	214	9	1.82
glass6	214	9	6.38
haberman	306	3	2.78
iris0	150	4	2
new-thyroid1	215	5	5.14
new-thyroid2	215	5	5.14
page-blocks0	5472	10	8.79
pima	768	8	1.87
segment0	2308	19	6.02
vehicle0	846	18	3.25
vehicle1	846	18	2.9
vehicle2	846	18	2.88
vehicle3	846	18	2.99
wisconsin	683	9	1.86
yeast1	1484	8	2.46
yeast3	1484	8	8.1

4.1. **Performance metrics for evaluating class imbalance problem.** The performance of our proposed method is evaluated using two metrics F-measure and Area Under the ROC Curve (AUC) [5] [24]. As illustrated in most of the research, accuracy is the poor indicator for measuring the performance of the classifier trained on imbalanced data sets. Therefore, the effectiveness of a classifier needed to be evaluated using additional metrics. Some common metrics used for measuring the performance of a classifier include AUC, F-measure and Geometric mean (G-mean) [8]. The reason for choosing AUC and F-measure is AUC evaluates the overall performance of the classifier on both classes [5] and the performance of only minority class is evaluated by F-measure [24].

4.2. **Results.** In this experiment, we have compared the proposed method with AdaBoost, RUSBoost, and SMOTEBoost [8]. The experiment was implemented using R, an open source statistical tool [23]. The two performance metrics are used to evaluate the classification performance. All experiments are performed using tenfold cross-validation. The training dataset is split into ten partitions, out of which nine are used to train the model, while the one hold-out partition is used to test the model. This process is repeated for ten times so that each partition act as test data. We use 22 data sets with various levels of imbalance and size from the KEEL-datasets repository. A decision tree (C4.5) [20] classifier is used as a base learner in boosting to train the data sets. Table 2 and 3 presents the performance

of each method, AdaBoost, RUSBoost, SMOTEBoost and proposed method across all data sets using AUC and F-measure metrics accordingly. From the results, we observe that the proposed method outperformed most of the time.

TABLE 2. Performances of the sampling techniques across all datasets using AUC Metric

Datasets	**AdaBoost**	**RUSBoost**	**SMOTEBoost**	**Cluster+boost**
ecoli-0_vs_1	0.6354	0.794	0.799	0.992
ecoli1	0.778	0.883	0.899	0.985
ecoli2	0.703	0.899	0.967	0.97
ecoli3	0.681	0.856	0.955	0.986
glass0	0.74	0.813	0.912	0.974
glass-0-1-2-3_vs_4-5-6	0.703	0.91	0.987	0.987
glass1	0.952	0.763	0.985	0.987
glass6	0.947	0.918	0.991	0.997
haberman	0.947	0.656	0.947	0.942
iris0	0.949	0.98	0.978	0.981
new-thyroid1	0.947	0.975	0.947	0.986
new-thyroid2	0.687	0.961	0.987	0.994
page-blocks0	0.637	0.953	0.967	0.996
pima	0.6223	0.751	0.897	0.899
segment0	0.996	0.994	0.998	0.998
vehicle0	0.943	0.965	0.968	0.978
vehicle1	0.754	0.768	0.897	0.899
vehicle2	0.854	0.966	0.967	0.978
vehicle3	0.745	0.763	0.894	0.894
wisconsin	0.9	0.96	0.994	0.894
yeast1	0.7589	0.7382	0.741	0.996
yeast3	0.93	0.944	0.944	0.994

5. **Conclusion.** The purpose of this study was to present a cluster-based oversampling with Boosting (Cluster+Boost) method which showed significant improvement over state-of-the-art algorithms to solve class imbalance problems. The approach consists of a novel synthetic data generation method using clustering methods with ensemble classifiers. We compared the performance of the proposed approach (Cluster+Boost) with that of the most effective techniques, AdaBoost, RUSBoost, and SMOTEBoost for learning from class imbalance problems. Based on the experimental results, we found that our proposed method (Cluster+Boost) achieves the best performance. Additionally, the performance of our proposed method produces better performance with datasets having a higher imbalanced ratio.

As a Future work, we investigate the performance of the proposed approach (Cluster+Boost) using additional classifiers and its effectiveness to the specific application domain. Furthermore, we also investigated by extending our method to handle multiclass imbalance problems.

REFERENCES

[1] A. Ali, S. M. Shamsuddin and A. L. Ralescu, Classification with class imbalance problem: A review, *Int J Adv Soft Comput Appl*, **7** (2015), 176–204.

[2] J. Alcalá-Fdez, L. Sánchez, S. Garcia, M. J. del Jesus, S. Ventura, J. M. Garrell, J. Otero, C. Romero, J. Bacardit and V. M. Rivas, et al., Keel: A software tool to assess evolutionary algorithms for data mining problems, *Soft Computing*, **13** (2009), 307–318.

TABLE 3. Performances of the sampling techniques across all datasets using F-measure Metric

Datasets	AdaBoost	RUSBoost	SMOTEBoost	Cluster+Boost
ecoli-0_vs_1	0.632	0.795	0.799	0.995
ecoli1	0.778	0.89	0.899	0.992
ecoli2	0.71	0.899	0.967	0.986
ecoli3	0.681	0.856	0.955	0.964
glass0	0.74	0.813	0.912	0.982
glass-0-1-2-3_vs_4-5-6	0.703	0.91	0.987	0.995
glass1	0.952	0.763	0.985	0.99
glass6	0.947	0.918	0.991	0.994
haberman	0.947	0.656	0.947	0.942
iris0	0.949	0.98	0.894	0.993
new-thyroid1	0.947	0.975	0.947	0.983
new-thyroid2	0.687	0.961	0.987	0.983
page-blocks0	0.637	0.953	0.967	0.998
pima	0.6223	0.751	0.897	0.894
segment0	0.996	0.994	0.998	0.998
vehicle0	0.943	0.965	0.988	0.984
vehicle1	0.754	0.768	0.897	0.894
vehicle2	0.854	0.966	0.967	0.941
vehicle3	0.745	0.763	0.894	0.894
wisconsin	0.9	0.96	0.994	0.997
yeast1	0.7589	0.7382	0.741	0.979
yeast3	0.93	0.944	0.944	0.974

[3] C. Bunkhumpornpat, K. Sinapiromsaran and C. Lursinsap, Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem, in: *Proceedings of the IEEE Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, **5476** (2009), 475–482.

[4] S. Barua, M. M. Islam, X. Yao and K. Murase, Mwmote–majority weighted minority oversampling technique for imbalanced data set learning, *IEEE Transactions on Knowledge and Data Engineering*, **26** (2014), 405–425.

[5] A. P. Bradley, The use of the area under the roc curve in the evaluation of machine learning algorithms, *Pattern Recognition*, **30** (1997), 1145–1159.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, Smote: Synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, **16** (2002), 321–357.

[7] N. V. Chawla, A. Lazarevic, L. O. Hall and K. W. Bowyer, Smoteboost: Improving prediction of the minority class in boosting, in: *European Conference on Principles of Data Mining and Knowledge Discovery*, Springer, 2003, 107–119.

[8] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince and F. Herrera, A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, **42** (2012), 463–484.

[9] V. García, R. A. Mollineda and J. S. Sánchez, On the k-nn performance in a challenging scenario of imbalance and overlapping, *Pattern Analysis and Applications*, **11** (2008), 269–280.

[10] H. He, Y. Bai, E. A. Garcia and S. Li, Adasyn: Adaptive synthetic sampling approach for imbalanced learning, in: *Proceedings of the IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, IEEE, 2008, 1322–1328.

[11] S. Hu, Y. Liang, L. Ma and Y. He, Msmote: Improving classification performance when training data is imbalanced, in: *Proceedings of the Second International Workshop on Computer Science and Engineering*, IEEE, **2** (2009), 13–17.

[12] H. Han, W.-Y. Wang and B.-H. Mao, Borderline-smote: A new over-sampling method in imbalanced data sets learning, in: *Proceedings of the International Conference on Intelligent Computing*, Springer, 2005, 878–887.

[13] M. Krstic and M. Bjelica, Impact of class imbalance on personalized program guide performance, *IEEE Transactions on Consumer Electronics*, **61** (2015), 90–95.
[14] M. Lin, K. Tang and X. Yao, Dynamic sampling approach to training neural networks for multiclass imbalance classification, *IEEE Transactions on Neural Networks and Learning Systems*, **24** (2013), 647–660.
[15] W.-Z. Lu and D. Wang, Ground-level ozone prediction by support vector machine approach with a cost-sensitive classification scheme, *Science of the Total Environment*, **395** (2008), 109–116.
[16] W.-C. Lin, C.-F. Tsai, Y.-H. Hu and J.-S. Jhang, Clustering-based undersampling in class-imbalanced data, *Information Sciences*, **409/410** (2017), 17–26.
[17] G. Rekha, A. K. Tyagi and V. Krishna Reddy, A wide scale classification of class imbalance problem and its solutions: A systematic literature review, *Journal of Computer Science*, **15** (2019), 886–929.
[18] G. Rekha, A. K. Tyagi and V. Krishna Reddy, Solving class imbalance problem using bagging, boosting techniques, with and without using noise filtering method, *International Journal of Hybrid Intelligent Systems*, **15** (2019), 67–76.
[19] F. Rayhan, S. Ahmed, A. Mahbub, M. Jani, S. Shatabda and D. M. Farid, et al., Cusboost: Cluster-based under-sampling with boosting for imbalanced classification, *2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*, (2017), arXiv1712.04356.
[20] S. Ruggieri, Efficient c4. 5 [classification algorithm], *IEEE Transactions on Knowledge and Data Engineering*, **14** (2002), 438–444.
[21] Y. Sun, A. K. Wong and M. S. Kamel, Classification of imbalanced data: A review, *International Journal of Pattern Recognition and Artificial Intelligence*, **23** (2009), 687–719.
[22] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse and A. Napolitano, Rusboost: A hybrid approach to alleviating class imbalance, *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, **40** (2010), 185–197.
[23] R. C. Team, R: A language and environment for statistical computing [internet], vienna (austria): R foundation for statistical computing.[cited 2015 mar 23] (2012).
[24] S. Wang and X. Yao, Diversity analysis on imbalanced data sets by using ensemble models, in: *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining*, IEEE, 2009, 324–331.

Received September 2019; revised December 2019.

E-mail address: gillala.rekha@klh.edu.in
E-mail address: vkrishnareddy@kluniversity.in
E-mail address: amitkrtyagi025@gmail.com