Distance-based Bootstrap Sampling in Bagging for Imbalanced Data-Set

G.Rekha
Department of Computer Science
and Engineering,
Koneru Lakshmaiah Education Foundation,
Deemed to be University,
Hyderabad,
Telangana -500075
Email: gillala.rekha@klh.edu.in

V. Krishna Reddy
Department of Computer Science
and Engineering,
Koneru Lakshmaiah Education Foundation,
Deemed to be University,
Vijayawada,
Guntur-522502
Email: vkrishnareddy@kluniversity.in

Amit Kumar Tyagi
School of Computing Science
and Engineering,
Vellore Institute of Technology,
Chennai Campus,
Chennai,
Tamilnadu, India-600 127
Email: amitkrtyagi025@gmail.com

Meghna Manoj Nair
School of Computing Science
and Engineering,
Vellore Institute of Technology,
Chennai Campus,
Chennai,
Tamilnadu, India-600 127
Email: mnairmeghna@gmail.com

2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) 978-1-7281-4142-8/20/$31.00 ©2020 IEEE 10.1109/ic-ETITE47903.2020.345

***Abstract*—In the recent decade, many new technologies and problems have attracted attention from research community/ scientists. Some problems like imbalanced data-set, security and privacy concerns in various computing environments like internet connected thing (IoTs), cloud computing, distributed computing, etc., have received many innovative ways as efficient answer. But, some problems are still unsolved. Learning from Imbalance data-set with higher accuracy is an essential task/ higher priority work in many applications. For handling the class imbalance problems, many extended approaches have been considered for bagging ensembles. In our study we show that application of distance-based approach (DistBagging) for balancing the distribution of each bag in ensemble bagging provides better results for addressing the class imbalance problems. In this work, we propose distance-based approaches for selecting the group of data for each bootstrap method to improve the performance of the classification in terms of accuracy for minority class in the imbalanced class distribution environment. The experimental results show that our distance-based approach outperforms the other ensemble techniques in the previous studies.**

***Index Terms*—Class Imbalance Problem, Ensemble Techniques, Bagging, Sampling.**

I. Introduction

The real-world classification problems still be challenging while trained on traditional classifiers. One of the common difficulties is the nature of the data i.e., Skewed class distribution (imbalanced class distribution), wherein one of the class(es) contains fewer samples than the other class(es). For example, in medical diagnosis, among 1000 patients, 10 patients may identify with cancer and the rest are healthier and don't need any treatment. Similar cases can be found like credit fraud detection, image recognition, risk management, oil detection and many more. In all the above cases, the class(es) with fewer samples/examples are crucially important. Such classes are called as minority/positive class(es) and the rest are represented as majority/negative class(es). The key role of the classifier is to recognize the minority classes while classification. However, traditional classification algorithms are biased towards negative class(es) and ignore positive class(es) leads to great difficulty while learning. Apart, ensemble techniques such as bagging, boosting are used to handle complex classification problem but will not perform well to this problem. However, may specialized techniques/methods/approaches have proposed by the researchers in the past decade for class imbalance problems [1, 2, 3, 4 and 24]. In overall, they may classified broadly into data level methods, algorithm level methods and ensemble methods. The former method try to re-balance the skewed distribution by either generating new samples for minority/positive class(es) called over-sampling or by discarding majority/negative class(es) called under-sampling. The widely used data-level techniques are Random OverSampling (ROS), Random Under-Sampling (RUS), Synthetic Minority Oversampling Technique (SMOTE), SPIDER framework [23, 22]. Next algorithm-level category provides solutions to improve the existing learning algorithms/classifier to accurately handle imbalanced data. They usually modify the existing algorithms or adapt cost sensitive techniques. The last approach, ensemble techniques, where in multiple classifier are used while training. For designing the solution for class imbalance problems, these techniques mostly employ pre-processing techniques before learning [5, 6]. The popular techniques proposed in literature and most often used in complex classification task are bagging, boosting and random forest. However, there is still a wider study needed for handling the data difficulty problems using ensemble techniques.

Hence, our paper contribution (in further sections) are sum-

978-1-7281-4142-8/$31.00 ©2020 IEEE

marized as follows:

- A new over-sampling technique using distance-based method for bagging on skewed data distribution.
- Extensive experimental evidence with appropriate metrics to evaluate the performance.

In summary, the rest of the paper is organized as follows: Section 2 elaborates the related work on ensemble algorithms for class imbalance problems. We explain our proposed method Distance-based bootstrap sampling along with the algorithm and also present different performance metrics considered for evaluation in Section 3. Section 4 describes the experimental results and analysis continued with final mark of conclusion in Section 5.

II. Related Work on Ensembles for Imbalanced Data

In the past decade several authors inspected the class imbalance problem in classification. The recent book [18] provides complete overview of several methods proposed in the literature. Below, a brief summary of existing methods relevant to our work have been presented. As specified, most often ensemble methods are combined with pre-processing techniques, like random oversampling and under-sampling. But, comes with drawback like oversampling may leads to overfitting whereas under-sampling may discard important information. Thus, informed methods came into existence like Synthetic Minority Oversampling TEchnique (SMOTE) [23]. In SMOTE the synthetic samples for minority class are generated by interpolating minority instances that lie close to each other using Euclidean distance. According to Galar et al [5] the ensemble techniques for imbalanced data are broadly classified into cost-sensitive and pre-processing combined with ensemble approaches. The former techniques mainly concentrate on cost minimization with boosting algorithms, for example, AdaCost, RareBoost. The latter integrates pre-processing techniques to balance the distribution along with ensemble learning (Boosting and bagging) algorithms.

Further, Liu et al [29] categorized ensemble techniques into bagging, boosting and hybrid approaches for class imbalance problems. The existing studies [8, 9, 5, 10, and 6] indicate good classification performance of bagging with respect to skewed distribution. So, we mainly focus on bagging techniques and further they are considered in our study.

Bagging proposed by Breimen [11] is an ensemble learning algorithm trained on original training data by splitting into 'T' bootstrap samples using same classifier. It considers majority voting method with equal weight for final prediction. The main component is bootstrap aggregation, wherein the training samples are uniformly sampled (with replacement) to train each classifier. While training the imbalanced dataset, the bootstrap sampling may bias towards the negative/majority class. To overcome this drawback, most of the research work applied pre-processing/ data level techniques to balance each bootstrap sample. In general, the two data level/ pre-processing techniques proposed are a). Underbagging b). Overbagging. In the former approach, the majority/ negative class samples in each bootstrap sample are reduced to the size of minority/positive samples. In [15], Exactly Balanced Bagging (EBBag) considers entire minority samples combined with subset of majority class samples to balance each bootstrap sample. The main drawback is constant sample size. To overcome, Roughly Balanced Bagging (RBBag) [16], proposed bagging based on negative binomial distribution for solving imbalanced data distribution. At each iteration, the size of the majority class in each bootstrap is set based on negative binomial distribution. According to [19], this method provides better performance compared to EBBag. Another way to balance the bootstrap samples is to oversample the minority class before training called Overbagging. Overbagging is the simplest method to balance the bootstrap by oversampling the minority class to that of majority class. The most important variant used is SMOTEBagging [14]. SMOTEBagging increases the diversity of bootstrap classifier by oversampling the minority class using SMOTE technique [17]. The number of samples to be generated at each iteration varies from smaller to higher values. According to [18], SMOTEBagging provides better performance(to some extent) over other random bagging-based ensemble techniques.

Finally, other two variants of underbagging proposed by Chan et al [13] was to partition the negative samples into non-overlapping subgroups and combine each subgroup with complete set of positive samples to form bag for building the bootstrap classifier. Next, in [14] authors proposed Balanced Random Forest (with replacement) from minority class samples and retaining the same number of majority class for each bootstrap. Then to train all the component bootstrap classifier, CART algorithms is used. The various approaches proposed in the literature for ensemble techniques are presented in the Table I. In the next section, we describe our proposed approach in detail (with explaining an algorithm).

III. Our Proposed Approach

One of the popular meta-learning algorithms is bagging, which build multiple base learners by considering sample subsets from the training data and finally aggregates all the base learners to make final prediction. Let 'T' be the amount of base learners. Given training data 'D' is divided into 'T' equal size subsets D_1, D_2,..., D_T using bootstrap sampling strategy. Let $f^T(x)$ be the base classifier trained on T^{th} subset and $f^s(x)$ be the final ensemble. The set of base models $f^1(x)$, $f^2(x)$,..., $f^T(x)$ is aggregated to generate the final ensemble $f^s(x)$.

Input:
- D-Training set
- LA- Base learner (C4.5)
- K- number of base learners which depends on the ratio of majority samples to minority samples divided by 2.

Output:
Build distance-based bootstrapping model (D,LA,K):

TABLE I
THE LIST OF CLASSIFICATION AND APPROACHES FOR HYBRID LEVEL

Technique	Algorithm	Metric
SMOTEBoost	C4.5, NB, RIPPER	AUC(ROC) and area under the PRC curve.
AdaBoost.NC	AdaBoost	Precision, F-measure, AUC, and G-mean.
Ensemble-based methods	C4.5	Accuracy,AUC
Bagging+SMOTE and RUS	NB, Sequential minimal optimization(SMO), and RBF	AUC and F-measure
DT	C4.5 and CART	AUC(ROC)
MEMMOT, MMMmOT, and CMEOT	RF, NB and AdaBoostM1	F-Measures and ROC
SMOTE-DGC	DGC	AUC, Specificity, Sensitivity, and G-Mean
Hybrid sampling and Bootstrapping	SVM, LR, k-Neural network(KNN) and Gaussian classifier	ROC(AUC)
Hybrid approach	Artifical NN, K-means and GA	Accuracy
Ensemble method	Modified SVM algorithm	Accuracy, Precision, F-measure and G-Mean
RUSBoost	C4.5	F-measure
RB-Boost ensembles approach	k-NN and SVM	ROC (AUC)
AdaOUBoost	SVM	Accuracy
Ensemble construction algorith (EUSBoost)	C4.5	AUC(ROC)
Ensemble classifier (NULCOEC)	SVM	Accuracy, G-means, Diversity
Sample Selection (SS)	MLP	Precision
RUS, ROS, SMOTE	RF	G-Mean
Ensemble method (AdaBoost)	G-mean Optimized Boosting	F-measure and G-Mean
SMOTE-ICS-Bagging	Iterative Classifier with Bagging	AUC
Resampling ensemble algorithm(REA)	NB	Precision, Recall, G-Mean, and F-measure
Ensemble Algorithm	M-Bagging	AUC (ROC) and G-Mean
Boosting and MDOBoost	C4.5	MAUC, G-Mean, and Recall
Undersampling techniques	AdaBoost	AUC, F-measure, and G-Mean
AdaBoost and SMOTE	AdaBoost	Precision, Recall, Specificity, Accuracy, and F-measure
Hybrid Approach	SVM	TN and TP
Cost-Sensitive techniques	Ensemble Classifier	Sensitivity, Specificity, and Accuracy
Emsemble Method-RUS	SVM	Accuracy
SMOTE	SVM	Accuracy, Sensitivity, Specificity, and G-Mean
Improved SMOTE	AdaBoost	Accuracy, Sensitivity, recall, Specificity, Precision, G – Mean, F –Measure
SMOTE	Adaboost.M1 and Bagging	Normalised Popt
Hybrid Ensemble (AdaBoost.M2 and SMOTE)	Bayesian Network (BN), DT, and Tree-J48	Sensitivity, Specificity, and Accuracy
Cost-sensitive learning	Ensemble method(AdaBoost)	F-measure
EasyEnsemble	EasyEnsemble	AUC
SMOTEBoost	AdaBoost.M2	Recall, Precision, and F-value
ensemble learning	1-NN(Multi-Layer Perceptron)	accuracy
ensemble learning	DataBoost-IM	F-measures, G-mean and overall accuracy

• Divide the given dataset ‘D’ into majority samples D_{maj} and minority samples D_{min}
For K= 1 to K
• Draw N_{maj} from ‘D’ randomly without replacement equal to the size of D_{min}
• Set N_{min} to D_{min}
• Now for each majority class find its average distance to all other majority class samples.
• Next discard the majority samples with the closest distance and retain only those majority class whose distance are farthest N'_{maj}
• Build a model $f^k(LA)$ by combining N'_{maj} and N_{min}
• Combine all $f^k(LA)$ into a aggregate model f(LA).
Return f(LA).

In the past there are several works explained the predictive performance of bagging. However, application of bagging to imbalanced data depends on the sampling strategy and size of the bootstrap bag ‘T’, as original bagging chooses bootstrap samples independent of class labels. To overcome the sampling problem in bagging, we proposed distance based bootstrapping approach.

A. Distance-Based Bootstrap Sampling in Bagging (DistBagging)

As mentioned, the problem of original bagging, we propose to provide bagging extension for handling skewed class distribution. To improve the performance of bagging on imbalanced dataset, a better generation of bootstrap subsets should be considered. Also, we need to avoid overfitting of data that usual happen while bootstrapping the samples of negative examples. We propose the distance based bootstrapping sampling for bagging to address above drawbacks. Figure 1 presents the algorithm. As with the original bagging algorithm, our algorithm consists of a training set 'D', a base learner LA, and K is the number of base learners based on the ratio of majority samples to minority samples for a given dataset as inputs. The bootstrap sampling starts with drawing majority samples from D randomly without replacement to avoid overfitting of minority samples. Next, to maintain the diversity among the

TABLE II
CONFUSION MATRIX

	Predicted negative	Predicted positive
Actual negative	True negative (TN)	False positive (FP)
Actual positive	False negative (FN)	True positive (TP)

samples and also to avoid noise and overlapping of samples we find the distance among each majority samples to all other minority samples and started selecting the majority samples whose distance is larger and discarding the samples whose distance is lesser. To compute the distance, we applied Euclidean distance formula.

IV. PERFORMANCE METRICS

In this section, we reconsider four metrics commonly used to assess the performance of the classification algorithms on imbalanced data. Confusion matrix is important to evaluate different metrics for classification performance. It consists of TP (True Positive), TN (True Negative), FP (False Positive), FN (False Negative) refer Table II. Accuracy: the simplest metric to predict the overall performance of a classifier is accuracy. For imbalanced datasets, the overall accuracy may lead to bias towards majority samples. So, many studies specific to class imbalance problems will not consider accuracy as good metric to predict the classification performance. However, accuracy is defined as equation 1.

$$Accuracy = \frac{TN + TP}{TN + FN + TP + FP} \tag{1}$$

Geometric Mean (G-Mean): G-Mean is calculated by taking the product of prediction accuracies for both positive and negative classes (refer equation 2). It mainly used to measure the amount of ignoring positive class and overfitting of negative class.

$$G - Mean = \sqrt{(}\frac{TN}{TN + FP} \times \frac{TP}{TP + FN}) \tag{2}$$

F-measure: The F-measure combines Precision and recall to predict the positive samples correctly. The higher the F-measure value provides better performance of model on positive class.F-measure value is calculated by equation 3.

$$Fmeasure = 2 \times \frac{precision \times recall}{precision + recall} \tag{3}$$

AUC (Area Under Curve): The AUC provides a main metric to assess the algorithms performance. It represents the trade-off between True Positive (TP) and False Positive (FP). The better performance is represented by the upper curve in learning from class imbalanced datasets. AUC is calculated by using Mann Whitney statistics by equation 4.

$$AUC = \frac{1 + TPrate - FPrate}{2} \tag{4}$$

Hence, this section describes the working model of the proposed approach along with different performance metric. In next section, we will present the experimental results.

TABLE III
CHARACTERISTICS OF DATA SETS USED IN THE EXPERIMENT

Data set	No.of attributes	Minority class labels	IR	Total No.of samples
Pima	8	1	1.87	768
breast-w	9	Malignant	1.9	699
new-thyroid	5	2	5.14	215
Vehicle	18	Van	3.25	846
credit-g	20	Bad	2.33	1000
Ecoli	7	imU	8.6	336
haberman	4	2	2.78	306
Yeast	8	ME2	28.1	1484
breast-cancer	9	recurrence-events	2.36	286

V. EXPERIMENT AND RESULT ANALYSIS

To evaluate the proposed approach, datasets, algorithm used are discussed in this section. Results are obtained by using 10-fold cross validation. For each dataset, F-measure, G-Mean and AUC is calculated.

A. Data-Sets used

We have used nine benchmark imbalanced data from UCI repository [20]. The Table III provides the details of datasets like dataset name, number of features, its Imbalance Ratio (IR), minority class label and total number of samples. All of the experiments were conducted using R programming language environment [21]. In the following experiments, we conduted ten-fold cross validation to preserve the similar class distribution for training and test set. The three metrics used to test the performance of the classifier are as follows AUC, F-measure, and G-mean.

B. Algorithm used in Bagging for Imbalanced Data-Set

Decision tree algorithms, especially C4.5 [19], are well known and most widely used approaches for classification problems. We employed base learner as C4.5 for distance based bootstrap sampling.

C. Result Analysis

In order to assess the performance of the proposed model, we compared the state-of-the-art technique like simple bagging, EBBagging, RBBagging, SMOTEBagging on nine datasets using G-Mean, AUC and F-Measure. From Table IV, V and VI, we observe that proposed method DistBagging outperformed in terms of G-Mean, AUC and F-Measure compared with state-of-the-art methods. The results are presented in the graphical form (Figure 1 - 3). In summary the performance of distance based bootstrap sampling method was consistent compared to the state-of-the-art techniques.

VI. CONCLUSION

We propose a new Distance based technique for bootstrapping (DistBagging) in bagging to deal with proper class distribution for imbalanced data. The DistBagging improve the performance of bagging by better generation of bootstrap subsets. It maintains the diversity among the samples and also avoid noise and overlapping of samples by finding the distance

TABLE IV
CLASSIFICATION PERFORMANCE USING G-MEAN

Data set	Bagging	EBBagging	RBBagging	SMOTEBagging	DistBagging
Pima	71.58	74.21	75.8	72.43	76.43
breast-w	95.88	96.1	96.4	95.88	95.88
new-thyroid	92.41	96.9	96.54	95.19	98.19
vehicle	93.89	94.58	95.5	94.35	95.35
credit-g	63.98	65.88	67.82	80.6	73.6
ecoli	68.67	72.25	88.87	58.39	87.39
haberman	62.81	78.98	78.69	68.47	86.47
yeast	51.49	84.78	84.78	59.4	64.4
breast-cancer	54.3	58.82	59.34	52.56	78.56

TABLE V
CLASSIFICATION PERFORMANCE USING AUC

Data set	Bagging	EBBagging	RBBagging	SMOTEBagging	DistBagging
Pima	61.29	76.71	78.55	58.89	78.67
breast-w	94.88	96	96.99	95.01	95.99
new-thyroid	87.51	95.45	95.72	92.14	95.72
Vehicle	91.29	91.17	97.05	92.14	98.52
credit-g	48.98	72.89	75.58	65.17	76.08
Ecoli	56.66	78.19	91.15	55	81.23
haberman	26.38	60.65	55.68	49.81	55.68
Yeast	32.22	90.22	87.66	51.54	88.66
breast-cancer	35.94	60.57	57.51	34.35	57.6

TABLE VI
CLASSIFICATION PERFORMANCE USING F-MEASURE

Data set	Bagging	EBBagging	RBBagging	SMOTEBagging	DistBagging
Pima	63.34	78.14	68.65	64.75	78.65
breast-w	95.55	97.01	94.92	95.01	91.02
new-thyroid	86.71	91.7	91.01	92.14	92.81
Vehicle	88.12	90.17	89.44	90.84	90.44
credit-g	48.98	72.89	55.87	65.17	65.87
Ecoli	58.67	67.12	59.56	56.78	59.56
haberman	45.56	67.76	58.65	52.61	58.65
Yeast	67.89	89.23	87.66	78.21	89.66
breast-cancer	56.76	62.34	59.34	67.78	69.34

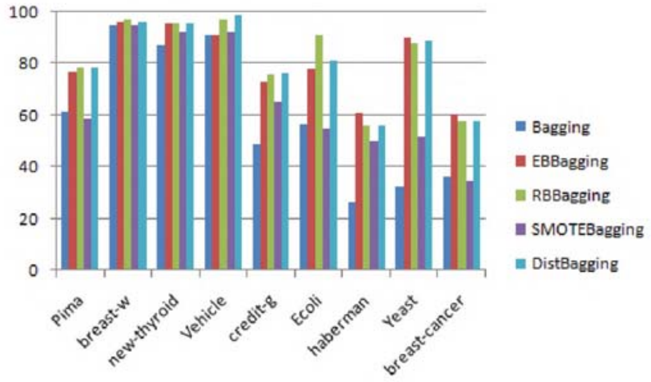

Fig. 1. Comparison of AUC performance on all data-sets using various Bagging techniques

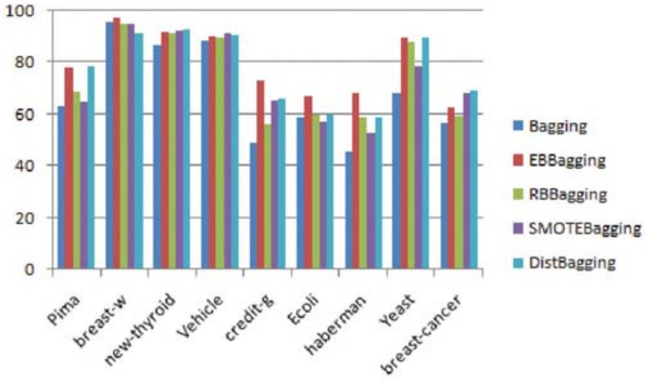

Fig. 2. Comparison of F-Measure performance on all data-sets using various Bagging techniques

among each majority samples to all other minority samples and started selecting the majority samples whose distance is larger and discarding the samples whose distance is lesser.

The experimental results of the proposed model show a better performance to deal with skewed class distribution problem and are statistically comparable to other well-known bagging

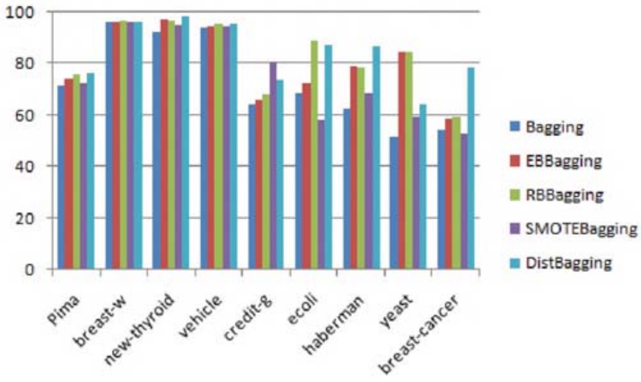

Fig. 3. Comparison of G-Mean performance on all data-sets using various Bagging techniques

approaches. Further work may focus on applying different distance measures on various real-world problems.

References

[1] Chawla N.: Data mining for imbalanced datasets: An overview. In: Maimon O., Rokach L., The Data Mining and Knowledge Discovery Handbook, 853867, (2005).

[2] He H., Garcia E.: Learning from imbalanced data. IEEE Transactions on Data and Knowledge Engineering, 21 (9), 1263–1284 (2009).

[3] He H., Yungian Ma: Imbalanced Learning. Foundations, Algorithms and Applications. Wiley, (2013).

[4] Weiss G.M.: Mining with rarity: a unifying framework. ACM SIGKDD Explorations Newsletter, vol. 6 (1), 7-19 (2004).

[5] Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. Herrera, F.: 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, 99, 1–22 (2011).

[6] Liu A., Zhu Zh: Ensemble methods for class imbalance learning. In He H., Yungian Ma., Imbalanced Learning. Foundations, Algorithms and Applications. Wiley, 6182 (2013).

[7] Stefanowski, J., Wilk, Sz.: Selective Pre-processing of Imbalanced Data for Improving Classification Performance. In: Proc. of 10th Int. Conference DaWaK 2008, Springer Verlag, LNCS vol. 5182, 283-292 (2008).

[8] Anyfantis, D., Karagiannopoulos, M., Kotsiantis, S., Pintelas, P. : Creating ensembles of classifiers by distributing an imbalance dataset to reach balance in each resulting training set. In Proc. of the IEEE International Conference on Distributed Human-Machine Systems Comf. – DHMS, (2008).

[9] Błaszczynski, J., Stefanowski, J., Idkowiak L.: Extending bagging for ´ imbalanced data. Proc. of the 8th CORES 2013, Springer Series on Advances in Intelligent Systems and Computing 226, 269–278 (2013).

[10] Khoshgoftaar T., Van Hulse J., Napolitano A.: Comparing boosting and bagging techniques with noisy and imbalanced data. IEEE Transactions on Systems, Man, and Cybernetics–Part A, 41 (3), 552–568 (2011).

[11] Breiman, L.: Bagging predictors. Machine Learning, 24 (2), 123–140 (1996).

[12] Hido S., Kashima H.: Roughly balanced bagging for imbalance data. Statistical Analysis and Data Mining, 2 (5-6), 412–426 (2009).

[13] Chen C., Liaw A., Breiman, L.: Using random forest to learn imbalanced data. Technical Report, University of California, Berkley, (2004).

[14] Wang, S., Yao, T.: Diversity analysis on imbalanced data sets by using ensemble models. In Proc. IEEE Symp. Comput. Intell. Data Mining, 324- 331 (2009).

[15] Sun, B., Chen, H., Wang, J. and Xie, H., 2018. Evolutionary undersampling based bagging ensemble method for imbalanced data classification. Frontiers of Computer Science, 12(2), pp.331-350.

[16] Błaszczyński, Jerzy, and Mateusz Lango. ”Diversity analysis on imbalanced data using neighbourhood and roughly balanced bagging ensembles.” International Conference on Artificial Intelligence and Soft Computing. Springer, Cham, (2016).

[17] Chawla, Nitesh V., et al. ”SMOTE: synthetic minority over-sampling technique.” Journal of artificial intelligence research 16 (2002): 321-357.

[18] G. Batista, D. F. Silva. How k-nearest neighbor parameters affect its performance. In Proc. of Argentine Symposium on Artificial Intelligence, Mar del Plata, Argentina, 1–12, (2009).

[19] Quinlan, J. Ross. ”Bagging, boosting, and C4. 5.” AAAI/IAAI, Vol. 1. (1996).

[20] D. Newman, S. Hettich, C. Blake, and C. Merz. UCI repository of machine learning databases, 1998.

[21] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, 2005.

[22] Stefanowski, J.; Wilk, S. Selective pre-processing of imbalanced data for improving classification performance. In Proceedings of the 10thInternational Conference in Data Warehousing and Knowledge Discovery (DaWaK 2008), Turin, Italy, 1–5 September 2008; pp. 283–292.

[23] Sun, J., Lang, J., Fujita, H. and Li, H., 2018. Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. Information Sciences, 425, pp.76-91.

[24] G. Rekha, K. Tyagi, 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.

[25] Rekha, G., Tyagi, A.K. and Krishna Reddy, V., 2019. Solving class imbalance problem using bagging, boosting techniques, with and without using noise filtering method. International Journal of Hybrid Intelligent Systems, (Preprint), pp.1-10.