I J C T A, 10(9), 2017, pp. 907-921 ISSN: 0974-5572

Analysing the Multi-class Imbalanced Datasets using Boosting Methods and Relevant Information

Neelam Rout*, Debahuti Mishra and Manas Kumar Mallick****

ABSTRACT

The class imbalance is the major problem for the real-world application because in the presence of the imbalanced datasets the performance of the standard learning algorithms are severely hindered. Normally, it is difficult to handle the multi-class imbalance problem than the binary class imbalance problem. Still now there are different techniques to accept the challenges of the multi-class imbalance datasets and this is the discussion points for this paper. The different kinds of multi-class imbalanced datasets and five types of Boosting methods are taken and by using the KEEL repository, an experimental analysis is shown.

Keywords: Multi-Class Imbalance Problem, Performance metrics, OVO Approach, Survey in Tabular Manner, Important Links, Distribution of the Data, Description of Datasets, Description of Software, KEEL Dataset Repository, Experimental Analysis, Wilcoxon Signed Rank Test

1. INTRODUCTION

A balanced dataset is required for the classical classification like K-Nearest Neighbour (KNN), Naive Bayes, and Support Vector Machine (SVM) but when the dataset is imbalanced then it is the crucial issues for these classifiers and the imbalanced dataset is occurred when one or more than one classes are under-represented in comparison to the other classes [1] [2]. Here, there are two types of classes *i.e.,* majority and minority and the majority class is the most prevalent class and the minority class is the rarest class. Most of the research community have given their attention towards binary class imbalanced problems [3] [4] [5]. But when the multiple classes are found in the imbalanced datasets then the solutions for the binary class imbalance problem may not be directly implemented [6]. Hence, the learning task is very complex when it faces imbalance problem. Many times the imbalanced situation is seen in the real world applications like medical diagnosis [7], text categorization [8], facial age estimation [9], anomaly detection [10], detecting oil spills [11], and fraudulent card transactions [12].

Till now, there are different methods to solve the problem *i. e.,* (i) data pre-processing methods [13] [14], which modify the distribution of dataset to get balance dataset and it is an external method, (ii) algorithmic methods [15], here the classification algorithms are modified to restrict a bias towards the rare or minority class, and (iii) cost-sensitive methods [16], which gives higher costs to misclassified examples. There is another way to tackle the imbalance problems *i.e.,* ensemble method like bagging [17] and boosting [18]. For the performance evaluation, the performance metric *i.e.* accuracy does not work because the traditional classifiers skip or neglect the minority class. So, it is not a standard metric and the standard metrics are precision, F-value, recall [19] [20] and ROC analysis [21] etc.

* Research Scholar, Siksha 'O' Anusandhan University, Bhubaneswar, India, *Email: neelamrout@soauniversity.ac.in*

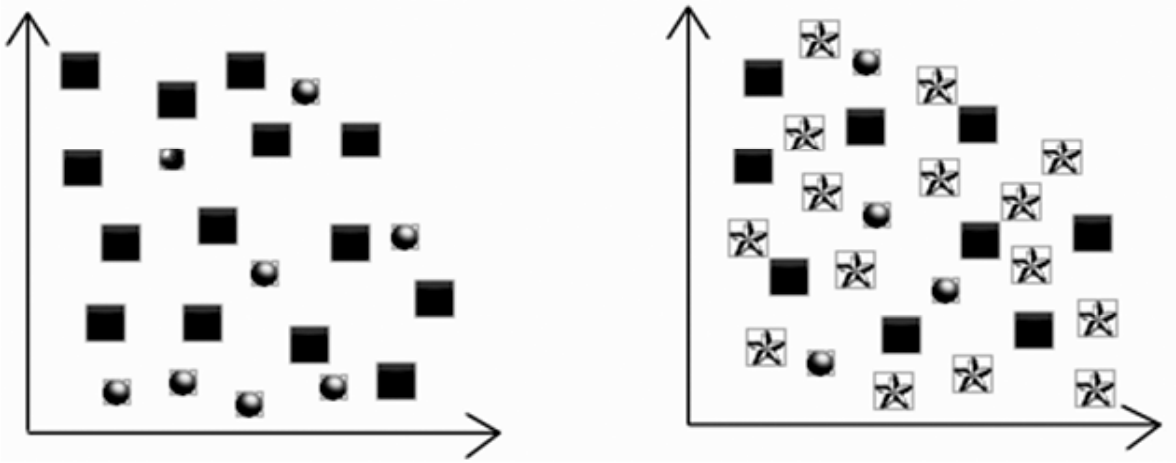
** Professor, Siksha 'O' Anusandhan University, Bhubaneswar, India, *Emails: debahutimishra@soauniversity.ac.in, mkmallick@soauniversity.ac.in*

The remaining part of the paper is as follows. The related work of the imbalanced data is described in section 2. The description of datasets and different kinds of software for data mining is given in section 3. In section 4, experimental framework and result analysis for the multi-class imbalance problem is given. Finally, the section 5 concluded the paper.

2. RELATED WORK

The paper [22] is focused on the multi-class imbalanced data problem because the solutions for the binary class imbalanced problem could not directly apply to the multi-class problem than the binary class imbalance problem.

The authors have addressed the new difficulties due to multi-class imbalanced datasets *i.e.*, many minorities to many majorities, one minority to many majorities, and many majorities to one minority.



(a) (b)

Figure 1: a. Binary class imbalanced problem b. Multi-class imbalanced problem

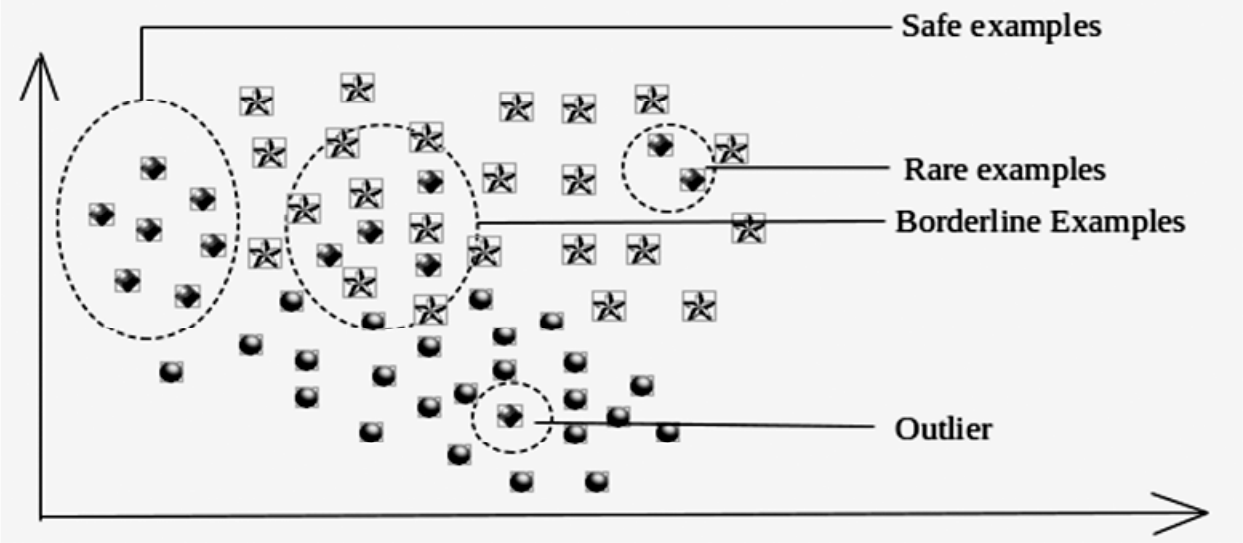


Figure 2: Presence of varieties of examples in a multi-class imbalanced datasets

Safe examples → Examples those are correctly identified by the classifier.

Outlier examples → Present in the other class where it is treated as noise.

Borderline examples → Present in the boundary regions between the different classes.

Rare examples → Small groups of examples that may consist of two or three examples.

◆ ✯ ● → Different types of data class.

The authors have used the function cn(e) to identify different types of examples present in the datasets means when cn(e) >= 4 then example e belongs to the safe zone, if 2 >= cn(e) >= 3 then it is present in borderline zone, if cn(e) = 1 then it belongs to rare class and if cn(e) = 0 then it belongs to the outlier zone. The original multi-class imbalanced datasets are firstly used to examine using C4.5, SVM, and NN classifiers. Secondly, these imbalanced datasets are pre-processed using oversampling method then compared the result using C 4.5, SVM and NN classifiers and finally, from the base multi-class datasets, the authors have chosen the class and types of examples to over sample the new synthetic examples by the help of over sampling method then these concrete classes and examples are input to the C4.5, SVM and NN classifiers. For the performance analysis, the average accuracy metric is used *i.e.*,

$$Avg\ Acc = \frac{\sum_{i=1}^{CL} TPR}{CL} \tag{1}$$

where, CL = Number of classes and TP = True positive rate for the i-th class

The results and graphics for all the 21 datasets could be found on the web-page http:// www.kssk.pwr.edu.pl/krawczyk/multi-over.

To deal with the multi-class imbalanced data, the authors [23] have used adaptive multiple classifier systems (AMCS). The different sampling methods are used in this paper *i.e.,* Adaboost.M1, Under sampling balanced ensemble (USBE), and Oversampling balanced ensemble (OSBE). The boosting is done by the re-sampling algorithm called as FiltEX. For USBE, the randomly under sampling method (RUS) and for OSBE, SMOTE method is used. Here, the wrapper methods (BPSO algorithm) and filter methods (FCBF algorithm) are taken for the feature selection methods to remove the irrelevant features without losing the useful features. The basic PSO is used as an optimization technique because it does not include the crossover and mutation operations. The five weighted ensemble rules are the weighted max, min, product, majority vote, and sum which are accomplished by AUC area i by multiply with pij. In their study, they have selected five base classifiers *i.e.*, C4.5, SVM, RBF-NN, DGC, and KNN. The majority votes have taken as the ensemble rule for Adaboost.M1 ensemble scheme. The AUC area is taken as the performance metric.

In data mining and machine learning [24], not only the binary class imbalanced datasets are affecting the performance of the standard learner classifiers but also the multi-class imbalanced datasets are hampering the classifier's performance very badly. The solutions to the binary class imbalance problem cannot be the

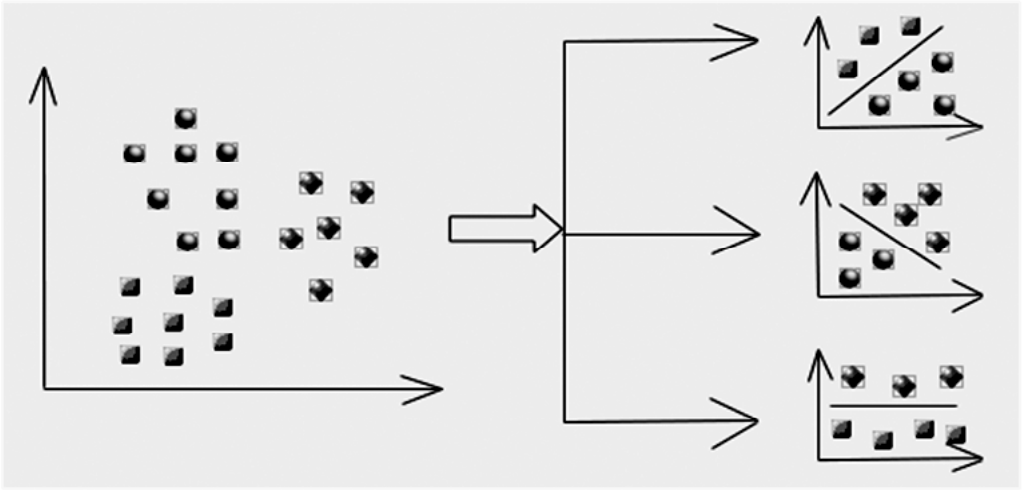


Figure 3: A three class problem converted to three two-class sub-problems using one-*versus*-one (OVO) scheme

direct solutions for the multi-class imbalanced datasets. So, first the multi-class classification problems are converted into the binary class sub-problems and the common methods are One Versus One (OVO) and One Versus All (OVA). For their experiment, The authors have taken six different approaches (RUSBoost, SMOTEBoost, SMOTE+ADABOOST, UnderBagging, SMOTEBagging, and Easy Ensemble) along with OVA scheme for the multi-class imbalance problem. In OVO scheme, an m-class problem is transferred into m(m-1)/2 binary or two class sub-problems and each sub-problem are examined by the different base classifiers and also having distinguishing instances and after that, it seems like the binary class sub-problems (shown in Fig. 3).

The authors have used OVO scheme to get the binary class sub-problem and after that, the pre-processing methods are used i.e., under sampling and over sampling (SMOTE) to balance the imbalanced binary class dataset then by the help of the score matrix the final result is obtained by the six different ensemble learning.

For the multi-class problem, the solutions for the binary class problem cannot be used directly because these techniques do not give good results. The authors have taken four under sampling techniques, four oversampling techniques, and one cost-sensitive learning approach for experiment using the KEEL software tool. For the evaluation purpose, the average accuracy is used for the performance evaluation rather than standard metrics like the accuracy rate. To solve the multi-class imbalanced problem, the authors have used different methods i.e., Static-SMT, Global-CS, AdaBoost NC, and binarization techniques (OVO and OVA). The One-*Versus*-One (OVO) approach is differentiated the class from another class in which examples do not belong to the each other. The code matrix M is used and according to the voting strategy, the instance will be classified. The One-*Versus*-All (OVA) approach, the instances of the one class is taken as positives and the other instances as negatives. The authors have selected the three different classifiers *i.e.*, C4.5, SVM and KNN with the different parameters like 0.25. The non-parametric (Wilcoxon signed-rank) test is used to know whether the algorithms are different from each other significantly and in which manner they differ. To use the different statistical tests, they have suggested a website *i.e.*, http://sci2s.ugr.es/sicidm/, and different information, software is also found and this is a research group of Granada University [6].

The model is complex for the classification when the number of classes *i.e.*, greater than 2 are present. The presence of the more number of classes is called as the multi-class learning. It is very difficult to handle the model, having multiple-class having the presence of the noise. According to the authors [25], there are two ways to handle these types of problem, the first one is to use robust learners and the second one is to use the pre-processing methods. After that, they have explained the advantages and disadvantages of these two approaches. So, they have referred other methods which could reduce the complexity of the problem i.e., One-*Vs*-One (OVO) and One-*Vs*-All (OVA) but mostly, they have focused on OVO strategy because it generally gives the better result than the OVA strategy. The noise can be class noise (contradictory example, misclassifications examples) and attribute noise (erroneous attribute values, miss "do not care" values). An additional metric, the mean f-measure is taken. After doing, the comparison between the OVO and OVA, it is concluded that OVO is the better approach than OVA in the multi-class problem. Many analysis and comparisons are done in this paper for the multi-class imbalanced data problem.

2.1. Literature Survey in the Tabular Manner

Table 1
Literature survey in the tabular manner

References	*Methods*	*Advantages*	*Disadvantages*
1	Over sampling with SMOTE, SVM, C4.5 and NN	To improve the performance and robust one	Computational complexity and difficult to find the optimal configuration

(*contd...Table 1*)

References	*Methods*	*Advantages*	*Disadvantages*
2	Two feature selection methods, three sampling mechanisms, five base classifier and five weighted ensemble rules	Be comparable or outperform with other algorithms	Restrict for the cost sensitive learning
3	One-*vs*-One (OVO) scheme, voting strategy (VOTE), under sampling, over sampling (SMOTE) and six types of ensemble classifiers	Converted to the simpler binary class sub-problems as well as powerful one	Complex architecture
4	Binarization scheme *i.e*., One Versus One and One Versus All and several ad-hoc procedures	The standard approaches for the binary class problem can be used for the multi-class problem, robust for the under sampling and cleaning procedure	OVA strategy not simpler than the OVO strategy and scalability problem
5	One-*vs*-One	Due to its decomposition technique, it is a robust classifiers for the noisy data as well as powerful	It is always not the solution for the multi-class imbalance problem

In the below, some of the important links are given from which the researchers can collect more knowledge about the imbalanced datasets and use in their research work.

1. http://www.kssk.pwr.edu.pl/krawczyk/multi-over
2. http://sci2s.ugr.es/about
3. http://sci2s.ugr.es/ovo_noise
4. http://sci2s.ugr.es/node/26
5. http://www.cwi.ugent.be/sarah.php

3. DESCRIPTION OF DATASETS AND DIFFERENT KINDS OF SOFTWARE FOR DATA MINING

There are many publicly available sites where one can get the imbalanced datasets and some of them are UCI machine learning repository, broad institute, KEEL-dataset repository etc. In this paper, the experiment is done by using KEEL-dataset repository. Lots of data mining software is available (http://www.predictiveanalyticstoday.com/top-free-data-mining-software/) and some of them are weka, KEEL *etc*. Matlab is also very useful and powerful language to do experiment.

By using the glass dataset, it is shown that how the datasets are scattered (Fig. 4 and Fig. 5) in the plane.

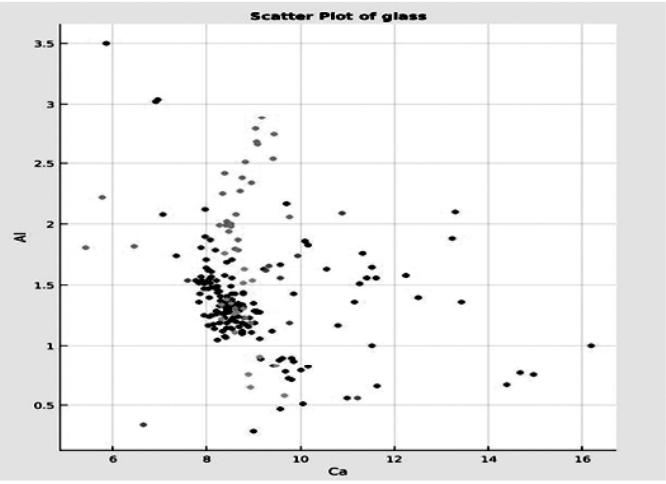


Figure 4: Scatter plot of glass dataset having attributes Al and Ca

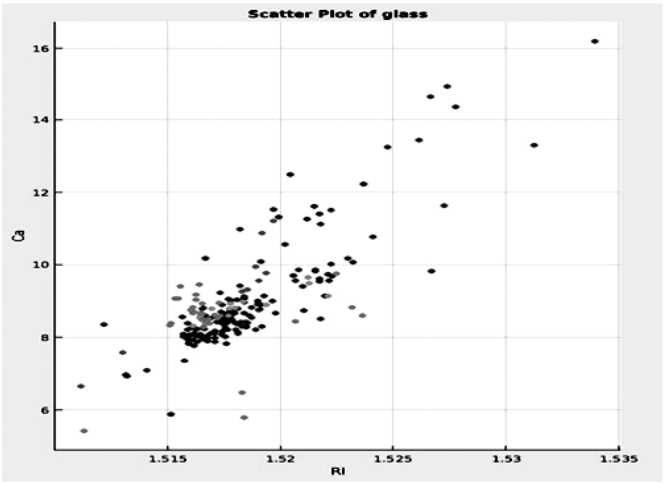


Figure 5: Scatter plot of glass dataset having attributes RI and Ca

In Fig. 4, the scatter plot of glass is shown where the x-axis represents the attribute 'Ca' and the y-axis represents the attribute 'Al'. In Fig. 5, x-axis and y-axis represent the attributes 'Rl' and 'Ca'. There are 7 different types of classes for the glass datasets. In this chapter, the multi-class datasets are collected from the KEEL datasets repository (http://sci2s.ugr.es/keel/imbalanced.php). In the below table, thirteen types of multi-class datasets are taken and the number of attributes, the number of examples are mentioned with their imbalance ratio (IR) in the ascending order.

Table 2
Description of imbalance datasets

Names	*Attributes*	*Examples*	*Imbalance Ratio (IR)*
wine	13	178	1.5
hayes-roth	4	132	1.7
contraceptive	9	1473	1.89
penbased	16	1100	1.95
new-thyroid	5	215	4.84
balance	4	625	5.88
dermatology	34	366	5.55
glass	9	214	8.44
lymphography	18	148	40.5
thyroid	21	720	36.94
ecoli	7	336	71.5
pageblocks	10	548	164
shuttle	9	2175	853

4. EXPERIMENTAL FRAMEWORK AND RESULT ANALYSIS

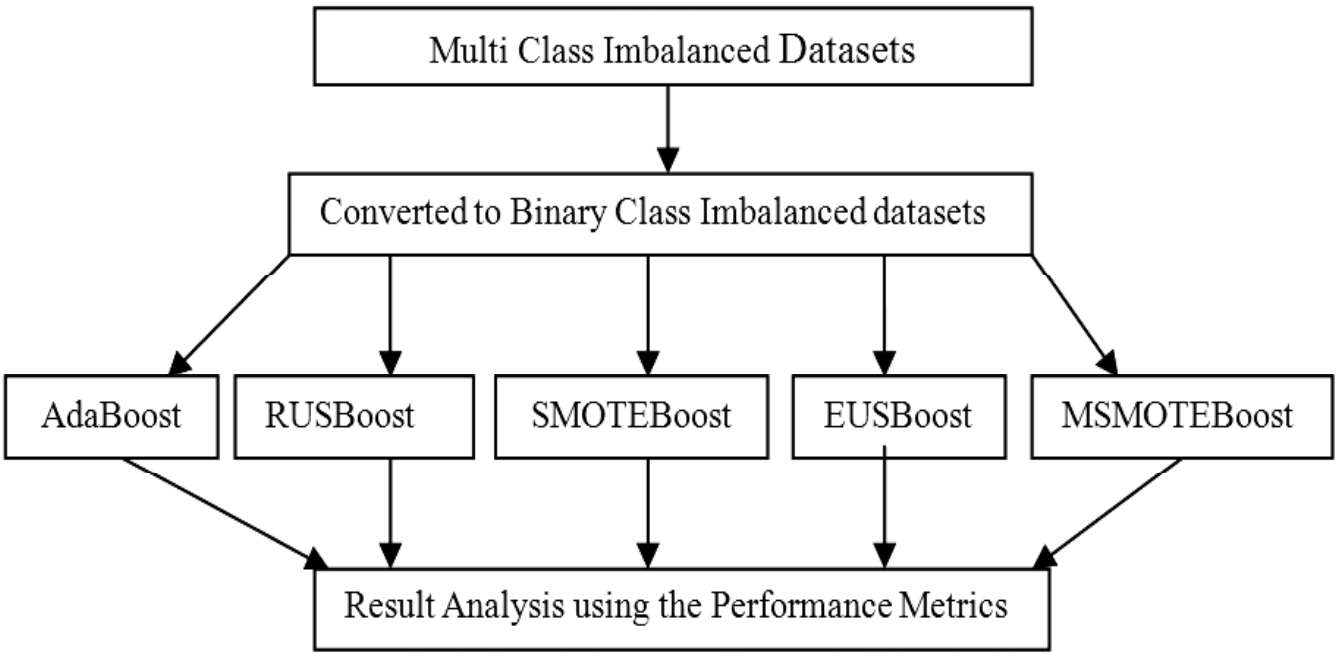


Figure 6: Proposed work model

The multi class imbalanced datasets are converted into binary class imbalanced datasets by merging many of the majorities class into one majority class and many of the minorities class into single minority class. Here, 5 fold cross validation is used. Here, five types of methods are used to solve the multi class

imbalance problem *i.e.,* AdaBoost [26], RUSBoost [27] , SMOTEBoost [28], EUSBoost [2], and MSMOTEBoost [29]. In these methods, C4.5 [30] is chosen as base classifier.

Adaptive Boosting (AdaBoost): It is created from the boosting method and also one of the top ten algorithms in data mining [31]. The objective is to reduce bias. By using the complete datasets, AdaBoost trains each classifier serially. By completion of each round, it gives more focus on the misclassified instances and it is done by giving more weight to incorrect examples than to correct one. In the test phase, each individual classifier is assigned a weighted value. Finally, the class label is chosen by the majority [32].

Synthetic Minority Over-Sampling Boosting(SMOTEBoost): This method is the combination of synthetic minority oversampling technique (SMOTE) and the ensemble Boosting procedure to increase the accuracy. SMOTE is used to create synthesized data for the minority class by randomly selecting k nearest neighbour. After that Boosting method is applied to the balanced datasets.

Modified Synthetic Minority Over-Sampling Boosting (MSMOTEBoost): MSMOTE is the extension of the SMOTE algorithm to improve the performance. In this method, the minority class has three groups i.e., safe, border, and latent noise instances by using the distance formula. If the instance is safe then SMOTE method is used, if the instance is under border group, then it only selects the nearest neighbor, otherwise, the instance is latent noise and does nothing for it. After that Boosting method is used.

Evolutionary Under-Sampling Boosting (EUSBoost): It comes from the applications of evolutionary prototype selection algorithms to improve accuracy and reduce the space necessity. Random under sampling is very powerful for the construction of the ensemble methods because it has good diversity. It is used in this paper for its simplicity, easy to implement, and effective.

Random Under-Sampling Boosting (RUSBoost): It is based upon the technique random under sampling where the instances of the majority class are removed randomly. To form a distribution, the weight for the new set of under sample data is normalized. Other procedure is same as SMOTEBoost.

AdaBoost

Parameter Descriptor	*Value*
Pruned	True
Confidence	0.25
InstancesPerLeaf	2
Number of classifiers	10
Train method	Noresampling

SMOTEBoost

Parameter Descriptor	*Value*
Pruned	True
Confidence	0.25
InstancesPerLeaf	2
Number of classifiers	10
Train method	Noresampling
Quantity of balancing SMOTE	50

MSMOTEBoost

Parameter Descriptor	*Value*
Pruned	True
Confidence	0.25

InstancesPerLeaf	2
Number of classifiers	10
Train method	Noresampling
Quantity of balancing MSMOTE	50

EUSBoost

Parameter Descriptor	*Value*
Pruned	True
Confidence	0.25
InstancesPerLeaf	2
Number of classifiers	10
Train method	Noresampling
% Majority class	50

RUSBoost

Parameter Descriptor	*Value*
Pruned	True
Confidence	0.25
InstancesPerLeaf	2
Number of classifiers	10
Train method	Noresampling
% Majority class	50

4.1. Experimental Setup And Computational Results

For all the methods C4.5 is taken as the base classifier. The confidence value is taken as 0.025, instance per leaf is 2, and the number of the classifier is 10. The percentage of the majority class is 50 except the AdaBoost method.

Table 3
Results of training and testing datasets using Accuracy

Datasets	*Training*					*Testing*				
	Ada Boost	*RUS Boost*	*SMOTE Boost*	*EUS Boost*	*MSMOTE Boost*	*Ada Boost*	*RUS Boost*	*SMOTE Boost*	*EUS Boost*	*MSMOTE Boost*
			Accuracy					*Accuracy*		
wine	0.9986	0.9761	1.0000	0.9846	0.9916	0.9607	0.9663	0.9494	0.9326	0.9326
hayes-roth	0.9962	1.0000	1.0000	0.9072	0.9924	0.9848	1.0000	0.9924	0.9015	0.9773
contraceptive	0.9657	0.7754	0.9180	0.7553	0.7945	0.7475	0.6633	0.7060	0.7401	0.6841
penbased	1.0000	0.9872	1.0000	0.9848	0.9959	0.9764	0.9682	0.9745	0.9718	0.9855
new-thyroid	1.0000	0.9651	1.0000	0.9744	0.9977	0.9674	0.9488	0.9581	0.9674	0.9581
balance	0.9220	0.4680	0.7580	0.8720	0.8700	0.9220	0.4660	0.0780	0.8770	0.6780
dermatology	1.0000	0.9710	0.9960	0.9660	0.9710	0.8100	0.9550	0.9550	0.9500	0.9250
glass	1.0000	0.9521	1.0000	0.9521	0.9673	0.9439	0.9112	0.9206	0.9346	0.9346
lymphography	0.9932	0.8074	1.0000	0.9139	1.0000	0.9720	0.7568	0.9720	0.8378	0.9662
thyroid	1.0000	0.9944	1.0000	0.9260	0.9997	0.9889	0.9917	0.9937	0.9264	1.0000
ecoli	1.0000	0.8770	1.0000	0.8943	0.9860	0.9760	0.8480	0.9700	0.9110	0.9850
pageblocks	1.0000	0.9329	1.0000	0.8489	0.9854	0.9763	0.9252	0.9726	0.8448	0.9252
shuttle	0.9993	0.9891	0.9999	0.9986	0.9989	0.9981	0.9871	0.9959	0.9982	0.9977

Table 4
Results of training and testing datasets using Area Under the ROC Curve

Datasets	*Training*					*Testing*				
	Ada Boost	*RUS Boost*	*SMOTE Boost*	*EUS Boost*	*MSMOTE Boost*	*Ada Boost*	*RUS Boost*	*SMOTE Boost*	*EUS Boost*	*MSMOTE Boost*
	Area Under the ROC Curve					*Area Under the ROC Curve*				
wine	0.9455	0.9455	0.9455	0.9455	0.9455	0.9920	0.9920	0.9920	0.9920	0.9920
hayes-roth	0.9768	0.9768	0.9768	0.9768	0.9768	0.9853	0.9853	0.9853	0.9853	0.9853
contraceptive	0.6231	0.6231	0.6231	0.6231	0.6231	0.8125	0.8125	0.8125	0.8125	0.8125
penbased	0.9647	0.9647	0.9647	0.9647	0.9647	0.9959	0.9959	0.9959	0.9959	0.9959
new-thyroid	0.9377	0.9377	0.9377	0.9377	0.9377	0.9920	0.9920	0.9920	0.9920	0.9920
balance	0.5216	0.5215	0.5215	0.5215	0.5215	0.6051	0.6051	0.6051	0.6051	0.6051
dermatology	0.9325	0.9325	0.9325	0.9325	0.9325	0.9886	0.9886	0.9886	0.9886	0.9886
glass	0.8680	0.8680	0.8680	0.8680	0.8680	0.9817	0.9817	0.9817	0.9817	0.9817
lymphography	0.7197	0.7197	0.7197	0.7197	0.7197	0.9561	0.9561	0.9561	0.9561	0.9561
thyroid	0.9271	0.9271	0.9271	0.9271	0.9271	0.9901	0.9901	0.9901	0.9901	0.9901
ecoli	0.8022	0.8022	0.8022	0.8022	0.8022	0.9550	0.9550	0.9550	0.9550	0.9550
pageblocks	0.9049	0.9049	0.9049	0.9049	0.9049	0.9737	0.9737	0.9737	0.9737	0.9737
shuttle	0.9918	0.9918	0.9918	0.9918	0.9918	0.9965	0.9965	0.9965	0.9965	0.9965

Table 5
Results of training and testing datasets using Specificity

Datasets	*Training*					*Testing*				
	Ada Boost	*RUS Boost*	*SMOTE Boost*	*EUS Boost*	*MSMOTE Boost*	*Ada Boost*	*RUS Boost*	*SMOTE Boost*	*EUS Boost*	*MSMOTE Boost*
	Specificity					*Specificity*				
wine	0.9948	0.1000	0.1000	0.1000	0.1000	0.9583	0.9792	0.8750	0.9167	0.9583
hayes-roth	0.98833	0.1000	0.1000	0.1000	0.1000	0.9333	0.1000	0.1000	0.1000	0.1000
contraceptive	0.9469	0.9069	0.9587	0.1786	0.0.8529	0.3303	0.6877	0.5676	0.1321	0.6426
penbased	0.1000	0.1000	0.1000	0.1000	0.9976	0.9238	0.9524	0.9524	0.9762	0.9333
new-thyroid	0.1000	0.1000	0.1000	0.1000	0.9917	0.8667	0.9000	0.9667	0.9333	0.8667
balance	0.1000	0.4566	0.7378	0.9427	0.9128	0.1000	0.4601	0.6927	0.9427	0.8802
dermatology	0.1000	0.1000	0.1000	0.1000	0.9963	0.8382	0.9559	0.8676	0.9559	0.9118
glass	0.1000	0.1000	0.1000	0.1000	0.9545	0.5909	0.8636	0.8182	0.8636	0.7727
lymphography	0.8333	0.1000	0.1000	0.9583	0.1000	0.3333	0.3333	0.8333	0.5000	0.3333
thyroid	0.1000	0.1000	0.1000	0.1000	0.9853	0.6471	0.8824	0.9412	0.1000	0.9412
ecoli	0.1000	0.1000	1.000	1.0000	0.8056	0.3333	0.7780	0,7778	0.8889	0.7778
pageblocks	0.1000	0.1000	0.1000	0.1000	0.9783	0.6087	0.9565	0.8696	0.9565	0.9565
shuttle	0.9924	0.1000	0.1000	0.1000	0.9866	0.9847	0.9924	0.1000	0.1000	0.9924
Average	0.4274	0.1895	0.2843	0.3061	0.6853	0.6499	0.7570	0.6603	0.6435	0.7744

Table 6
Results of training and testing datasets using Sensitivity

Datasets	*Training*					*Testing*				
	Ada Boost	*RUS Boost*	*SMOTE Boost*	*EUS Boost*	*MSMOTE Boost*	*Ada Boost*	*RUS Boost*	*SMOTE Boost*	*EUS Boost*	*MSMOTE Boost*
	Sensitivity					*Sensitivity*				
wine	0.1000	0.9673	0.1000	0.9885	0.9885	0.9615	0.9615	0.9538	0.9935	0.9462
hayes-roth	0.1000	0.1000	0.1000	0.8799	0.9902	0.1000	0.1000	0.9706	0.8725	0.9902
contraceptive	0.9857	0.7371	0.9061	0.9241	0.7774	0.8693	0.6561	0.7465	0.9026	0.6974
penbased	0.1000	0.9843	0.1000	0.9820	0.9955	0.9888	0.9719	0.9933	0.9708	0.9843
new-thyroid	0.1000	0.9595	0.1000	0.9703	0.9986	0.9838	0.9568	0.9568	0.9730	0.9730
balance	0.0000	0.5969	0.1000	0.0357	0.3673	0.0000	0.5306	0.5102	0.0408	0.1633
dermatology	0.1000	0.9664	0.1000	0.9586	0.9647	0.9793	0.9552	0.9759	0.9483	0.9276
glass	0.1000	0.9466	0.1000	0.9466	0.9688	0.9844	0.9427	0.9323	0.9167	0.9531
lymphography	0.1000	0.7993	0.1000	0.9120	0.1000	0.1000	0.7746	0.9718	0.8521	0.1000
thyroid	0.1000	0.9943	0.1000	0.9243	0.9979	0.9912	0.9943	0.9829	0.9246	0.9886
ecoli	0.1000	0.8739	0.1000	0.8914	0.9908	0.9939	0.8502	0.9755	0.9113	0.9908
pageblocks	0.1000	0.9300	0.1000	0.8424	0.9857	0.9924	0.9238	0.9790	0.8400	0.9733
shuttle	0.9998	0.9884	0.9991	0.9985	0.9998	0.9990	0.9868	0.9976	0.9980	0.9868
Average	0.2297	0.8342	0.2312	0.8657	0.8558	0.7649	0.8157	0.9189	0.8572	0.8211

Table 7
Results of training and testing datasets using Balanced Accuracy (BAcc)

Datasets	*Training*					*Testing*				
	Ada Boost	*RUS Boost*	*SMOTE Boost*	*EUS Boost*	*MSMOTE Boost*	*Ada Boost*	*RUS Boost*	*SMOTE Boost*	*EUS Boost*	*MSMOTE Boost*
	Balanced Accuracy (BAcc)					*Balanced Accuracy (BAcc)*				
wine	0.9974	0.9837	0.1000	0.9942	0.9942	0.9599	0.9704	0.9144	0.9551	0.9522
hayes-roth	0.9917	0.1000	0.1000	0.9400	0.9951	0.9667	0.1000	0.9853	0.9363	0.9951
contraceptive	0.9414	0.8220	0.9324	0.5514	0.8151	0.5998	0.6719	0.6570	0.5174	0.6700
penbased	0.1000	0.9921	0.1000	0.9910	0.9966	0.9563	0.9621	0.9728	0.9735	0.9588
new-thyroid	0.1000	0.9797	0.1000	0.9851	0.9952	0.9252	0.9284	0.9617	0.9532	0.9198
balance	0.5000	0.5268	0.8689	0.4892	0.6401	0.5000	0.4953	0.6015	0.4943	0.5218
dermatology	0.1000	0.9832	0.1000	0.9793	0.9805	0.9088	0.9555	0.9218	0.9521	0.9197
glass	0.1000	0.9733	0.1000	0.9733	0.9616	0.7876	0.9032	0.8752	0.8902	0.8629
lymphography	0.9167	0.8996	0.1000	0.9352	0.1000	0.6667	0.5540	0.9026	0.6761	0.6667
thyroid	0.1000	0.9972	0.1000	0.9621	0.9916	0.8221	0.9383	0.9621	0.9623	0.9649
ecoli	0.1000	0.9369	0.1000	0.9457	0.8982	0.6636	0.8140	0.8767	0.9001	0.8843
pageblocks	0.1000	0.9650	0.1000	0.9212	0.9820	0.8005	0.9402	0.9243	0.8983	0.9649
shuttle	0.9961	0.9942	0.9999	0.9993	0.9932	0.9919	0.9896	0.9988	0.9990	0.9896
Average	0.4649	0.8570	0.2924	0.8975	0.8726	0.8115	0.7864	0.8888	0.8545	0.8660

Table 8
Results of training and testing datasets using F - Measure

Datasets	*Training*					*Testing*				
	Ada Boost	*RUS Boost*	*SMOTE Boost*	*EUS Boost*	*MSMOTE Boost*	*Ada Boost*	*RUS Boost*	*SMOTE Boost*	*EUS Boost*	*MSMOTE Boost*
	F - Measure					*F - Measure*				
wine	0.9990	0.9834	0.1000	0.9942	0.9942	0.9728	0.9766	0.9538	0.9951	0.9647
hayes-roth	0.9976	0.1000	0.1000	0.9361	0.9951	0.9903	0.1000	0.9851	0.9319	0.9951
contraceptive	0.9780	0.8356	0.9448	0.8541	0.8541	0.8420	0.7510	0.7972	0.8373	0.7741
penbased	0.1000	0.9921	0.1000	0.9909	0.9975	0.9854	0.9802	0.9910	0.9824	0.9843
new-thyroid	0.1000	0.8761	0.1000	0.9849	0.9986	0.9811	0.9699	0.9756	0.9809	0.9756
balance	0.0000	0.0289	0.4252	0.0238	0.2414	0.0000	-0.0052	0.1166	-0.0137	0.6035
dermatology	0.1000	0.9829	0.1000	0.9789	0.9816	0.9707	0.9719	0.9725	0.9683	0.9522
glass	0.1000	0.9727	0.1000	0.9726	0.9815	0.9692	0.9628	0.9547	0.9488	0.9632
lymphography	0.9965	0.8885	0.1000	0.9531	0.1000	0.9861	0.8594	0.9822	0.9098	0.9861
thyroid	0.1000	0.9971	0.1000	0.9606	0.9988	0.9943	0.9957	0.9907	0.9608	0.9936
ecoli	0.1000	0.9327	0.1000	0.9426	0.9927	0.9878	0.9160	0.9846	0.9521	0.9923
pageblocks	0.1000	0.9637	0.1000	0.9144	0.9923	0.9877	0.9594	0.9866	0.9121	0.9855
shuttle	0.9996	0.9942	0.9999	0.9993	0.9994	0.9990	0.9931	0.9988	0.9990	0.9931
Average	0.4362	0.8114	0.2592	0.8850	0.8559	0.8974	0.8024	0.8992	0.8742	0.9356

Table 9
Results of training and testing datasets using Matthews Correlation Coefficient (MCC)

Datasets	*Training*					*Testing*				
	Ada Boost	*RUS Boost*	*SMOTE Boost*	*EUS Boost*	*MSMOTE Boost*	*Ada Boost*	*RUS Boost*	*SMOTE Boost*	*EUS Boost*	*MSMOTE Boost*
	Matthews Correlation Coefficient (MCC)					*Matthews Correlation Coefficient (MCC)*				
wine	0.9964	0.9427	0.1000	0.9790	0.9790	0.9029	0.9180	0.8288	0.8759	0.8778
hayes-roth	0.9892	0.1000	0.1000	0.7904	0.9789	0.9578	0.1000	0.9393	0.7802	0.9789
contraceptive	0.9007	0.5480	0.7977	0.1439	0.5437	0.2194	0.2912	0.2806	0.0474	0.0.2931
penbased	0.1000	0.9606	0.1000	0.9552	0.9869	0.9228	0.9006	0.9526	0.9138	0.9176
new-thyroid	0.1000	0.8761	0.1000	0.9055	0.9903	0.8623	0.8041	0.8477	0.8712	0.8282
balance	0.0000	0.2991	0.7871	0.0836	0.6105	0.0000	0.2695	0.3985	0.0987	0.3003
dermatology	0.1000	0.9193	0.1000	0.9027	0.9131	0.8421	0.8657	0.8533	0.8514	0.7802
glass	0.1000	0.8036	0.1000	0.8036	0.8445	0.6643	0.7054	0.6467	0.6407	0.6748
lymphography	0.9097	0.3728	0.1000	0.5220	0.1000	0.5694	0.0506	0.6645	0.1884	0.5694
thyroid	0.1000	0.8972	0.1000	0.4729	0.9497	0.7346	0.8304	0.7258	0.4739	0.7865
ecoli	0.1000	0.3956	0.1000	0.4246	0.7477	0.4363	0.2720	0.5889	0.4127	0.7303
pageblocks	0.1000	0.5983	0.1000	0.4280	0.8497	0.6762	0.5573	0.7366	0.4044	0.7526
shuttle	0.9939	0.9147	0.9990	0.9880	0.9908	0.9838	0.9002	0.9802	0.9841	0.9002
Average	0.4223	0.6633	0.2757	0.6461	0.8065	0.6748	0.5742	0.7264	0.5802	0.7223

Table 10
Results of training and testing datasets using G-Mean

Datasets	*Training*					*Testing*				
	Ada Boost	*RUS Boost*	*SMOTE Boost*	*EUS Boost*	*MSMOTE Boost*	*Ada Boost*	*RUS Boost*	*SMOTE Boost*	*EUS Boost*	*MSMOTE Boost*
			G-Mean					*G-Mean*		
wine	0.9974	0.9673	1.0000	0.9894	0.9942	0.9599	0.9703	0.9522	0.9275	0.9136
hayes-roth	0.7324	1.0000	1.0000	0.9380	0.9951	0.9661	0.9341	0.9852	1.0000	0.9951
contraceptive	0.9410	0.8170	0.9320	0.4060	0.8140	0.5360	0.6720	0.7450	0.3450	0.6830
penbased	1.0000	0.9921	1.0000	0.9901	0.9966	0.9557	0.9620	0.9585	0.9735	0.9726
new-thyroid	1.0000	0.9850	0.9951	0.9795	1.0000	0.3622	0.9279	0.8702	0.9529	0.9617
balance	1.0000	0.5220	0.8590	0.1840	0.5760	1.0000	0.4600	0.3780	0.1950	0.5940
dermatology	1.0000	0.9830	0.9891	0.8367	0.9803	0.9060	0.9556	0.9201	0.9520	0.9196
glass	1.0000	0.9729	1.0000	0.9729	0.9616	0.7627	0.9023	0.8733	0.8898	0.8581
lymphography	0.9129	0.8940	1.0000	0.9349	1.0000	0.5774	0.7186	0.5733	0.6527	0.8991
thyroid	1.0000	0.9972	1.0000	0.9614	0.9832	0.8033	0.3420	0.9646	0.9616	1.0000
ecoli	1.0000	0.9348	1.0000	0.9442	0.8877	0.5756	0.8132	0.8710	0.9000	0.8779
pageblocks	1.0000	0.9644	1.0000	0.9178	0.9643	0.3883	0.9400	0.9649	0.8963	0.9227
shuttle	0.9961	0.9942	0.9999	0.9993	0.9932	0.9624	0.9896	0.9798	0.9694	0.9988
Avearage	0.9677	0.9249	0.9827	0.8503	0.9343	0.7504	0.8144	0.8489	0.8166	0.8920

To analysze the performance of the classifier, the performance metrics *i.e.*, accuracy, area under the ROC curve, and G-Mean [33] are taken in this paper. The equation for the overall accuracy (OA) (2) is as follows:

$$Overall\ Accuracy\left(OA\right)=\frac{TP+TN}{TP+FP+FN+TN} \tag{2}$$

Where TP is true positive, TN is true negative, FP is false positive, and FN is false negative. The results for the training and testing datasets are shown in table 3 but the accuracy is not the standard metric for the imbalanced datasets because it neglects the instances of the minority class. So, it is given good results. In table 4, the results of the area under the ROC (receiver operating characteristic) curve are shown and it is useful when the curve is not clear and used to show the performance of the classifiers. But the results of the single dataset are same for all the methods. The specificity is defined as the ratio of the true negative and true positive with false negative (3) [35]. The ratio true positive and true positive with false negative is called as sensitivity (4) [35].The results are shown in table 3 and 4. The balanced accuracy (BAcc) (5) is measured the quality of the used methods and the results are shown in table 5. The F-Measure (6) is the trade-off between the precision and the recall and the results are shown in table 6. The Matthews Correlation Coefficient (MCC) (7) [36] is summarized the confusion matrix into a single value and introduced by the Matthews in 1975. The range of the values is differed from -1 to +1. The result is shown in table 7. Then the performance metrics G-Mean (8) is used to evaluate the classifier's performance by using positive accuracy and negative accuracy and the results are shown in table 8. The results of each performance metrics are shown and the highest values are underlined. The average values are not calculated for the accuracy and area under the ROC Curve because accuracy is not the standard metric for the imbalanced data and area under the ROC Curve has given the same values for each boosting methods. From the experiment, it is known that for the proposed work, for the training dataset, SMOTEBoost has given the best result using G-mean that is 98.27% and for the testing dataset, MSMOTEBoost has given the best result using G-mean that is 93.56%.

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

$$Balanced\ Accuracy\ \left(BAcc\right) = \frac{1}{2} * Sensitivity + Specificity \tag{5}$$

$$F - Measure = \frac{\left(1 + \beta^2\right) * Precision * Recall}{\beta^2 * Recall + Precision} \tag{6}$$

Where, β = 1, $Precision = \frac{TP}{\left(TP + FP\right)}$, and $Recall = \frac{TP}{\left(TP + FN\right)}$

$$Matthews\ Correlation\ Coefficient\ \left(MCC\right) = \frac{TP * TN - FP * FN}{\sqrt{\left(TP + FP\right)\left(TP + FN\right)\left(TN + FP\right)\left(TN + FN\right)}} \tag{7}$$

$$G - Mean = \sqrt{\frac{TP}{TP + FN} * \frac{TN}{TN + FP}} \tag{8}$$

4.2. Statistical Test

In this paper, the nonparametric (Wilcoxon signed rank) test is used to compare the two samples that are paired or related by using the performance results of the area under the ROC Curve [34]. This method is used when the samples are not in normal distributions and relatively small. First, the differences between the two techniques (having n objects) are calculated after that, calculate the absolute values and then rank it by ignoring the zero values. The values of R+ and R- are calculated where R^+ is the sum of ranks of positive differences and R^- is the sum of ranks of positive differences. Here, the significance value á is taken as 0.05 and the p-value also gives important information about the significance differences of the classifier. The whole calculations are shown in table 6. The selection column shows the selection of the techniques which depends on the hypothesis's rejection or if it is not rejected then rejection is depended on the ranks basis.

Table 6
Wilcoxon test for pair wise comparison

Comparison	R^+	R^-	*p-value*	*Hypothesis (α = 0.05)*	*Selection*
AdaBoost *vs.*EUSBoost	19	59	0.1213	Not rejected	EUSBoost
MSMOTEBoost *vs.*SMOTEBoost	36	42	0.8291	Not Rejected	SMOTEBoost
RUSBoost *vs.* SMOTEBoost	32	100	0.1364	Not Rejected	SMOTEBoost
SMOTEBoost t *vs.* EUSBoost	48	30	0.4925	Rejection for SMOTEBoost	EUSBoost

5. CONCLUSIONS

In this paper, the short description is given regarding the multi-class imbalanced datasets that what is the multi-class imbalanced datasets, how it is generated and handled. The results for the multi-class imbalanced datasets are analyzed using various techniques and also, the Wilcoxon signed rank test is used for the statistical analysis. For future research, there are various challenges for the multi-class imbalanced datasets

like to deal with severe or highly imbalanced datasets, to decrease the cost function, or to improve the performance of the many well-liked classification algorithms. Hence, it is a very broad and sensitive area to do research.

ACKNOWLEDGMENTS

The authors would like to thank our S'O'A university and the reviewers.

REFERENCES

[1] Díez-Pastor JF, Rodríguez JJ, García-Osorio C, Kuncheva LI. Random Balance: Ensembles of variable priors classifiers for imbalanced data. Knowledge-Based Systems. 2015 Sep 30; 85: 96-111.

[2] Galar M, Fernández A, Barrenechea E, Herrera F. EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. Pattern Recognition. 2013 Dec 31; 46(12): 3460-71.

[3] Chawla NV, Japkowicz N, Kotcz A. Editorial: special issue on learning from imbalanced data sets. ACM Sigkdd Explorations Newsletter. 2004 Jun 1; 6(1): 1-6.

[4] Japkowicz N, Stephen S. The class imbalance problem: A systematic study. Intelligent data analysis. 2002 Jan 1; 6(5): 429-49.

[5] Orriols-Puig A, Bernadó-Mansilla E. Evolutionary rule-based systems for imbalanced data sets. Soft Computing. 2009 Feb 1; 13(3): 213-25.

[6] FernáNdez A, LóPez V, Galar M, Del Jesus MJ, Herrera F. Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. Knowledge-based systems. 2013 Apr 30; 42: 97-110.

[7] Krawczyk B, Galar M, Jeleń Ł, Herrera F. Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. Applied Soft Computing. 2016 Jan 31;38:714-26.

[8] Pramokchon P, Piamsa-nga P. Reducing Effects of Class Imbalance Distribution in Multi-class Text Categorization. InRecent Advances in Information and Communication Technology 2014 (pp. 263-272). Springer International Publishing.

[9] Chao WL, Liu JZ, Ding JJ. Facial age estimation based on label-sensitive learning and age-oriented regression. Pattern Recognition. 2013 Mar 31; 46(3): 628-41.

[10] Khreich W, Granger E, Miri A, Sabourin R. Adaptive ROC-based ensembles of HMMs applied to anomaly detection. Pattern Recognition. 2012 Jan 31; 45(1): 208-30.

[11] Kubat M, Holte RC, Matwin S. Machine learning for the detection of oil spills in satellite radar images. Machine learning. 1998 Feb 1; 30(2-3): 195-215.

[12] Fawcett T, Provost F. Adaptive fraud detection. Data mining and knowledge discovery. 1997 Sep 1; 1(3): 291-316.

[13] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002; 16: 321-57.

[14] Batista GE, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. ACM Sigkdd Explorations Newsletter. 2004 Jun 1; 6(1): 20-9.

[15] Huang K, Yang H, King I, Lyu MR. Imbalanced learning with a biased minimax probability machine. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2006 Aug; 36(4): 913-23.

[16] Sun Y, Kamel MS, Wong AK, Wang Y. Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition. 2007 Dec 31; 40(12): 3358-78.

[17] Breiman L. Bagging predictors. Machine learning. 1996 Aug 1; 24(2): 123-40.

[18] Freund Y, Schapire RE. Experiments with a new boosting algorithm. InIcml 1996 Jul 3 (Vol. 96, pp. 148-156).

[19] Buckland M, Gey F. The relationship between recall and precision. Journal of the American society for information science. 1994 Jan 1; 45(1): 12.

[20] Joshi MV, Kumar V, Agarwal RC. Evaluating boosting algorithms to classify rare classes: Comparison and improvements. InData Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on 2001 (pp. 257-264). IEEE.

[21] Provost F, Fawcett T. Robust classification for imprecise environments. Machine learning. 2001 Mar 1; 42(3): 203-31.

[22] Sáez JA, Krawczyk B, WoŸniak M. Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. Pattern Recognition. 2016 Sep 30; 57:164-78.

[23] Yijing L, Haixiang G, Xiao L, Yanan L, Jinling L. Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. Knowledge-Based Systems. 2016 Feb 15; 94: 88-104.

[24] Zhang Z, Krawczyk B, Garcìa S, Rosales-Pérez A, Herrera F. Empowering one-*vs*-one decomposition with ensemble learning for multi-class imbalanced data. Knowledge-Based Systems. 2016 May 25.

[25] Sáez JA, Galar M, Luengo J, Herrera F. Analyzing the presence of noise in multi-class problems: alleviating its influence with the One-vs-One decomposition. Knowledge and information systems. 2014 Jan 1;38(1): 179-206.

[26] Freund Y, Schapire RE. A desicion-theoretic generalization of on-line learning and an application to boosting. InEuropean conference on computational learning theory 1995 Mar 13 (pp. 23-37). Springer Berlin Heidelberg.

[27] Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A. RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans. 2010 Jan; 40(1):185-97.

[28] Chawla NV, Lazarevic A, Hall LO, Bowyer KW. SMOTEBoost: Improving prediction of the minority class in boosting. InEuropean Conference on Principles of Data Mining and Knowledge Discovery 2003 Sep 22 (pp. 107-119). Springer Berlin Heidelberg.

[29] S. Hu, Y. Liang, L. Ma, Y. He. MSMOTE: Improving classification performance when training data is imbalanced. 2nd International Workshop on Computer Science and Engineering (WCSE 2009). Qingdao (China , 2009) 13-17.

[30] Quinlan J. C4. 5: Programs for Machine Learning. C4. 5-programs for machine learning/J. Ross Quinlan. 1993.

[31] Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Philip SY, Zhou ZH. Top 10 algorithms in data mining. Knowledge and information systems. 2008 Jan 1; 14(1): 1-37.

[32] Rudin C, Daubechies I, Schapire RE. The dynamics of AdaBoost: Cyclic behavior and convergence of margins. Journal of Machine Learning Research. 2004; 5 (Dec): 1557-95.

[33] Thanathamathee P, Lursinsap C. Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and AdaBoost techniques. Pattern Recognition Letters. 2013 Sep 1; 34(12): 1339-47.

[34] Wilcoxon F. Individual comparisons by ranking methods. Biometrics bulletin. 1945 Dec 1; 1(6): 80-3.

[35] Zhu W, Zeng N, Wang N. Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations. NESUG proceedings: health care and life sciences, Baltimore, Maryland. 2010 Nov 14: 1-9.

[36] Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure. 1975 Oct 20; 405(2): 442-51.