International Journal of Hybrid Intelligent Systems 15 (2019) 67–76
DOI 10.3233/HIS-190261
IOS Press

Solving class imbalance problem using bagging, boosting techniques, with and without using noise filtering method

G. Rekha[a], Amit Kumar Tyagi[b,*] and V. Krishna Reddy[a]
[a] *Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Guntur 522502, India*
[b] *Lingaya's Vidyapeeth, Faridabad 121002, India*

Abstract. In numerous real-world applications/domains, the class imbalance problem is prevalent/hot topic to focus. In various existing work, for solving class imbalance problem, almost data is labeled as one class called majority class, while fewer data is labeled as the other class, called minority class (more important class to focus). But, none of the work has performed efficiently (in terms of accuracy). This work presents a comparison of the performance of several boosting and bagging techniques from imbalanced datasets. The wide range of application of data mining and machine learning encounters class imbalance problem. An imbalanced datasets consists of samples with skewed distribution and traditional methods show biased towards the negative (majority) samples. Note that popular pre-processing technique for handling class imbalance problems is called over-sampling. It balances the datasets to achieve a high classification rate and also avoids the bias towards majority class samples. Over-sampling technique takes full minority samples in the training data into consideration while performing classification. But, the presence of some noise (in the minority samples and majority samples) may degrade the classification performance. Hence, the work presents a performance comparison using boosting and bagging (i.e., with both techniques) with and without using noise filtering. This work evaluates the performance with the state of-the-art methods based on ensemble learning like AdaBoost, RUSBoost, SMOTEBoost, Bagging, OverBagging, SMOTEBagging on 25 imbalance binary class datasets with various Imbalance Ratios (IR). The experimental results show that our approach works as promising and effective for dealing with imbalanced datasets using metrics like F-Measure and AUC.

Keywords: Class imbalance problem, ensemble learning method, noise filter, boosting, bagging

1. Introduction

The skewed distribution in the datasets frequently appears in the field of financial systems, health science, information science, and mechanical engineering [2,3,20,21]. The skew distribution occurs when the samples of one of the class (majority/negative) are severely outnumbered by those of other class (minority/positive). But traditional algorithms when trained on imbalanced (skewed) datasets tend to favour the majority/negative class, resulting in high overall classification accuracy. While the minority class will typically be the class of interest and often ignored by the traditional algorithm. For example, when a classifier is trained on a binary class data sets with 1% samples from the minority class and 99% with majority class, a 99% accuracy can be achieved by the classification models. But these models are practically not valuable. As the minority class are mostly the classes of interest. In the literature, several techniques for dealing class imbalance problem have been proposed. The common method for dealing with class imbalance problem is data sampling [11,18]. Using data sampling, the data is balanced by adding samples to mi-

[*] Corresponding author: Amit Kumar Tyagi, Lingaya's Vidyapeeth, Faridabad 121002, Haryana, India. Tel.: +91 9487868518; E-mail: amitkrtyagi025@gmail.com.

nority class called over sampling or removing the samples from the majority class called under-sampling. Two popular data sampling techniques are Random Under-Sampling (RUS) and Random Over-Sampling techniques (ROS). Apart, Synthetic Minority Oversampling TEchnique (SMOTE) [26] is most popular technique under oversampling technique. In the literature, Boosting and Bagging are effective techniques for training the imbalanced datasets [1,38]. These techniques are commonly known as ensemble of classifiers, combining the individual classifiers for improving the accuracy of the classifiers. Boosting adaptively re-samples according to the weights attached to the samples to produces high accuracy. At each stage for the incorrect classified samples the weights are adjusted and trained on by the classifier. Whereas Bagging or Bootstrap Aggregating method trains multiple classifiers and combined into one predictor. Hence, the organization (remaining part) of this work/paper is followed as: Section 2 discusses the work related to class imbalance problem in brief. Further, the motivation behind working related to this problem is discussed in Section 3. Further ensemble technique is being discussed in Section 4. Further, our proposed method (with noise filtering and sampling technique) is discussed in Section 5 in detail. Later, experiments or simulation results have been discussed in Section 6 with several parameters like AUC, and F-Measure. Finally, Section 7 concludes this paper with some future enhancement (in brief).

2. Related work

Existing approaches (for dealing with class imbalanced problem) can be roughly categorized into algorithm level approaches and data level approaches. The former directly modifies the traditional algorithms to achieve cost sensitivity by taking different misclassification costs into consideration. The algorithms are designed such that the misclassification cost of positive examples is higher than that of negative ones. The goal of a classifier is to minimize the cost instead of classification error, and therefore the classification algorithms will bias towards the small class. At the data level, a data pre-processing step is added to rebalance the class distribution by undersampling the negative class, oversampling the positive class, or creating synthetic positive examples. Wang et al. [4] proposed an online cost-sensitive ensemble learning framework for their online version. They generalized a batch of widely used cost sensitive learning like bagging and boosting. The performance of online bagging and boosting cost-sensitive ensemble learning are determined largely by their batch mode consistency and also the batch ensemble algorithm performance. The results show an outstanding performance by bagging based algorithms both in terms of accuracy and consistency. The high comparable performance was achieved by AdaC2 and CSB2 on the other hand worse performance was achieved by RUSBoost and SMOTEBoost algorithms. To use the ensemble pruning methodology (in the context of imbalanced classification) or for improving the behavior, we used ensemble-based solutions in this work. The author developed a novel ordering-based pruning metrics to address the class imbalance problem [23]. The five most popular ordering-based pruning techniques are: Reduce-Error pruning with Geometric Mean (RE-GM), Kappa pruning (Kappa), Complementarity Measure (Comp), Margin Distance Minimization for imbalanced problems (MDM-Imb), Boosting-Based pruning for imbalanced problems (BB-Imb) have been adapted. The performance of Under-Bagging with RE-GM and Roughly-Based Bagging with BBImb pruning approaches are best among others. A novel ensemble method, called Bagging of Extrapolation Borderline-SMOTE SVM (BEBS), has been proposed in dealing with Imbalanced Data Learning (IDL) problems [30]. The BEBS framework employed an adaptive sampling method called Extrapolation Borderline-SMOTE and bootstrapping aggregation by taking the boundary information derived from the initial SVM and bagging mechanism. It contributes to the relief of overfitting and promotes the capability of models generalization. Learning from imbalanced data is still one of challenging tasks in machine learning and data mining.

The paper [30] discusses the different data difficulty factors which deteriorate classification performance for an imbalanced dataset. The general problems in class imbalance are decomposition of the minority class into rare sub-concepts, overlapping of classes and distinguishing different types of examples. Stefanowski [17] presented new experiment which shows the influence of these factors on classifiers. They also include critical discussions of methods for their identification in real world data. The author [32] proposed an effective Cost-Sensitive Classifier with GentleBoost Ensemble (Can-CSC-GBE) for the classification of breast cancer. The author [10] proposed Binary Particle Swarm Optimization (BPSO)-Adaptive BoostingK Nearest Neighbour (BPSO-Adaboost-KNN) ensemble learning algorithm for addressing the class imbalance problem for multi-class datasets. The BPSO are applied for selecting the important features and then Adaboost-KNN classifier is used to train the model. The performance is measure using AUC. The author [35] proposed an effective ensemble learning framework called adaptive Ensemble Under-Sampling (EUS). They applied Adaboost technique as individual classifier to train the EUS by changing the data distribution and acquiring balanced subsets. By using Real Adaboost they observed an improvement in the accuracy. The output by these classifiers is aggregated by a weighted voting system that is based on the error rate of each individual classifier. Finally, they proposed an adaptive threshold selection method based on OTSU [28] algorithm to find the optimal threshold for the final decision. The authors [5] mainly focused on four types of ensemble solutions such as bagging-based, boosting-based, random-forest-based and hybrid ensemble for customer relationship management churn prediction. They compared the solutions using AUC as general evaluation metric and Expected Maximum Profit (EMP) for domain specific from the perspective of costs and benefits. The original Bagging and random forest learning algorithms performed well with respect to the profit-based measure. Farid et al. [8] proposed multi class imbalance method based on clustering. They trained the classifiers on genomic data. The genomic data is considered noisy, high dimensional, and also with small sample size.

Initially, imbalanced data is divided into majority and minority clusters. The majority class in turn grouped into several clusters. Then, they find the most informative instances in each cluster by combining the instances of minority classes. Finally, multiple classifiers are used to train the different groups. C4.5 decision tree algorithm has been used as base classifier. Recently, some researchers have evidenced that not all the samples are valuable and donated to a classifiers learning [19]. Some samples may be redundant and tend to increase the computational cost. Some may even worsen a classifiers performance, which should be treated as noises and need to be removed/cleaned in both majority and minority classes. However, to the best of our knowledge, there is not much work available to verify the influence of noise (available in both, i.e., in the minority/majority class). Thus, we intend to propose a framework to deal with the noisy examples in both minority and majority classes via a noise filter combined with over-sampling. Note that in this work, we focus only on binary classification problems. This represents the first attempt to combine the noise filter with re-sampling methods. In order to verify the efficiency, we choose six popular sampling methods with ensemble classifiers, i.e., AdaBoost, RUSBoost, SMOTEBoost, Bagging, OverBagging and SMOTE-Bagging to implement the proposed framework with a K-Nearest Neighbor (KNN)-based noise filter. We design several experiments to test our proposed method with collected datasets (from KEEL Machine Learning Repository). In last, the propose framework is compared with the two metrics, i.e., Area Under the Curve (AUC) and F-measure. Hence, this section discusses about related work for handling class imbalance problem at data level. Now, next section will deal with different ensemble techniques to handle class imbalance problem.

3. Motivation

In the past decade, machine learning techniques have been used for solving several problems with respect to big data or class imbalance problem. In machine learning, decision making process or an output matter a lot to people or human-beings. Based on such output, further decision is made/taken for next/further process. But, when this accuracy or results (of collected datasets) differs based on a size or features of dataset, then it may create a serious problem to validity and originality of data/decision. So, to provide validity a data, we need to trained more minority sample than majority samples. For example, if a patient received false information based on a false result/inaccurate result or less features/data-sets, then it is a serious problem in e-healthcare application. We need to provide exact or accurate information to patients using appropriate tools/or efficient methods (based on analyzing collected data (of patients)). Hence, with keeping accuracy in our mind, we try to solve this problem of class imbalance using boosting and bagging techniques; with and without noise filtering method.

4. Ensemble techniques to handle class imbalance problem

As discussed, the ensemble methods are used to handle class imbalance problems. In [16,24], the author proven that ensembles of classifiers will yield better performance/accuracy and are robust at handling the datasets that are imbalanced in nature. However, ensemble techniques with combination of either pre-processing or cost-sensitive approaches lead to promising results. The pre-processing techniques are embeded in to an ensemble learning algorithm [7,33,36]. The cost sensitive ensemble techniques are similar with that of cost-sensitive approaches. These techniques apply different misclassification cost for different classes using boosting algorithms. Apart, the most popular techniques used in the literature is ensemble methods combined with pre-processing techniques. In [22], author classified ensemble methods into bagging, boosting and hybrid-based approaches. These approaches adopted data-level techniques, such as undersampling and oversampling approaches to balance the data before training it with the base classifiers. The combination of data level techniques with ensemble algorithms (Bagging and Boosting) resulted in better performance [12,13,22]. The some of the ensemble methods used in this paper are discussed below.

- SMOTEBoost [27]: It is an oversampling method based on Synthetic Minority Oversampling Technique (SMOTE). SMOTE uses K-Nearest Neighbour (K-NN) to generate the synthetic samples of the minority class. SMOTEBoost adopts SMOTE with Boosting technique. At each iteration, the newly generated samples using SMOTE are used to train the boosting algorithm. The newly generated synthetic samples are assigned weights which are proportional to the total number of overall samples and the weights of the other samples (original samples) are normalized. During the entire process, the weights of the samples are updated according to the Adaboost.M2 algorithm.
- RUSBoost [7]: It is similar to SMOTEBoost, but it adopts a random undersampling technique to remove the samples from the majority class in each iteration. Thus, it will not generate new samples, but it is enough to normalizing the weights of the samples using Adaboost.M2 algorithm.
- Underbagging [31]: It uses undersampling method applied to majority class samples. In each iteration, the majority samples are reduced randomly to that of the minority class and the base classifier are trained on this balanced dataset.
- OverBagging [33]: It is similar to random oversampling method applied to minority class. In each iteration, the oversampling process is applied to increase the minority samples to that of the majority class and the base classifier are trained on this balanced dataset.
- SMOTEBagging [33]: It is a combination of SMOTE with bagging ensemble algorithm. At each iteration, the method generates samples that has the two times the number of majority samples, wherein half of the samples are randomly generated with replacement from majority class and the remaining half is generated through SMOTE and Random Over-Sampling (ROS) on the minority class.

Hence, this section discusses about different ensemble algorithms for handling class imbalance problem at data level. Now, next section will deal with proposed data level framework for class imbalance problem.

5. Noise filtering and sampling technique: A proposed framework

The two common problems in the data quality are existence of noise and class imbalance which degrade the performance of the classifiers. For overcoming the noise, a pre-processing technique called noise filtering is used. Noise filtering detect the noise existing in the data and removes it. Noise basically present in most the real world data [37] and can degrade the system performance in terms of classification accuracy, time in building a classifier and the size of the classifier. Sáez et al. [15] proposed Iterative-Partitioning Filter (IPF) using oversampling algorithm called, SMOTE-IPF. This technique is used to pre-process the data before training it on to the classifier. They divided the training dataset into n subsets and used a set of n base classifiers to train it independently. To the best of our knowledge, existing noise filters in the literature are always combined with under sampling, oversampling techniques or only deal with the noisy examples in the majority class. No noise filtering attempt focuses on entire dataset using ensemble approach in the process of solving class imbalance problem. Can such challenge boost a classifiers performance? Answering this question, this work proposes a Noise-Filtered Over-Sampling technique with Boosting (NF-OS with Boosting) and Bagging (NF-OS with Bagging) techniques, as shown in Fig. 1. In this work, we consider both boosting and bagging ensemble techniques with and without applying NF-OS noise filtering. NF-OS with boosting and NF-OS with bagging are based on the combination of noise filtering with sampling and ensemble (Adaboost, Bagging) algorithm. It is similar to RUSBoost, SMOTEBoost, SMOTEBagging with the critical difference with removal of noise

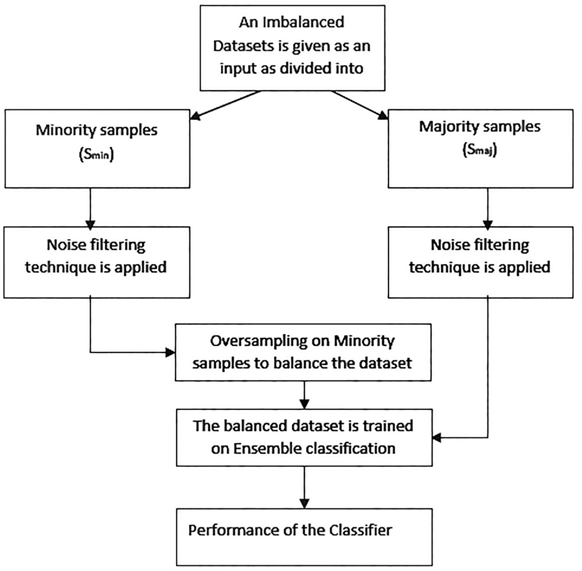
An Imbalanced Datasets is given as an input as divided into
Minority samples (Smin)
Majority samples (Smaj)
Noise filtering technique is applied
Noise filtering technique is applied
Oversampling on Minority samples to balance the dataset
The balanced dataset is trained on Ensemble classification
Performance of the Classifier

Fig. 1. A Noise-filtered Over-Sampling Technique with Boosting (NF-OS with Boosting).

occurring in the datasets. SMOTEBoost uses SMOTE method to oversample the minority class examples, RUSBoost uses random under-sampling on the majority class while SMOTEBagging uses SMOTE method to oversample the minority class. In comparison, our proposed NF-OS uses noise filtering with sampling from the majority class. Considering a given dataset D, I. We define subsets Smaj ⊂ D and Smin ⊂ D, where Smin is the set of minority samples in D, and Smaj is the majority class. II. The noise in minority and majority samples are removed using K-Nearest Neighbours as follows (shown in the Fig. 2).

In this work, each sample falls into any of the six categories based on nearest neighbours.

a) Extremely useful majority sample: All the K-NN are majority class label (labelled as ‘A’ in the Fig. 2).
b) Extremely useful minority sample: All the K-NN are minority class label (labelled as ‘a’ in the Fig. 2).
c) Relatively useful majority sample: Most of the K-NN samples belong to majority class label (labelled as ‘B’ in the Fig. 2).
d) Relatively useful minority sample: Most of the K-NN samples belong to minority class label (labelled as ‘b’ in the Fig. 2).
e) Noisy sample: All the K-NN belongs to different class label (both for majority and minority samples) (labelled as ‘C’ and ‘c’ in the Fig. 2).

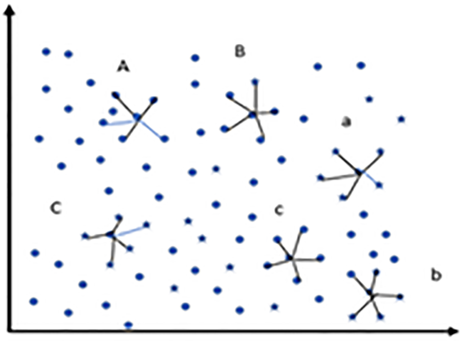

Fig. 2. Six categories of samples. ‘*’ represents minority data. ‘o’ represents majority data.

In this work, ‘noisy samples’ are identified using K-NN algorithm. Note that the choice of K will be highly influenced to find whether a sample is a noise or not. If K is too small then a sample can be classified as a noise and if K is too large then it is considered as a useful one. The best value of K is this work is considered as 5. The strength of our approach lies in the fact that it considers examples after removal of noise in the entire dataset. After, NF-OS applies SMOTE technique to oversampling the minority samples in order to balance the imbalanced dataset. Once the dataset is balanced, classification is done using Boosting and Bagging method. Note that, the Boosting algorithm considers a series of decision trees using C4.5 algorithm and combines the votes of each individual tree to classify new sample, whereas Bagging creates individual classifiers for its ensemble by training each classifier on a random redistribution of the training set. Each classifier’s training set is generated by randomly drawing, with replacement. Hence, this section discusses the proposed framework for handling class imbalance problem at data level. Now, next section will deal with the simulation and result for the proposed framework.

6. Experimental scenarios and simulation results

This section presents a series of experiments conducted to test the performance of our proposed method, pre-processing using noise filtering based ensemble learning for imbalanced datasets. We have selected 25 imbalanced binary datasets from the KEEL repository [14]. We analyse the performance of our proposed method with RUSBoost, AdaBoost, SMOTE-Boost, Bagging, OverBagging and SMOTEBagging. Here first, we present the different evaluation metrics, benchmark datasets used and experimental settings for

Table 1
Confusion matrix

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

Table 2
Data-set used

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
vowel0	988	13	9.98
glass 4	214	9	15.46
ecoli4	336	7	15.8

class imbalanced learning. Then, we show the results in form of two performance metrics for imbalanced learning, i.e., the AUC and F-measure.

6.1. Experimental setting

In our experiment work, we tested the proposed scheme on 25 benchmark datasets shown in Table 2. For every dataset, we used C4.5 decision tree algorithms as a base learner in boosting and bagging. Here, each experiment is done with 20 independent runs with 10-fold cross validation and acquires the average results in terms of AUC and F-measure, respectively.

6.2. Assessment metrics

To evaluate the performance of the classifier, the way it has been evaluated play an important role. We have to use specific evaluation metrics based on the distribution of the data. Traditionally, the common metric used for evaluating the performance of balanced classification algorithms is the accuracy of the classifier. It is defined as 'ratio of number of correctly classified samples to the total number of samples'.

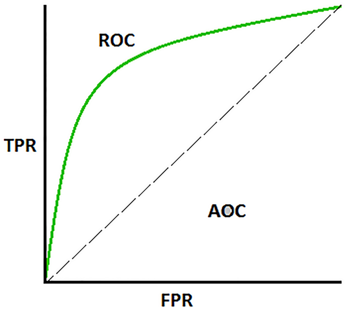

Fig. 3. AUC-ROC curve

$$\text{Accuracy} = (\text{number of correctly classified samples})/(\text{Total number of samples}) \quad (1)$$

However, it is not appropriate for imbalanced data sets. For example, there is a binary-class imbalanced problem with an imbalanced rate of 99:1, with 99 majority samples and only 1 minority one. The goal of traditional learning algorithm is to minimize the error rate and for imbalanced data set with 99:1 rate, which may simply group all the samples into the majority class and thus attains 99% accuracy. So, these learning algorithms are not a good approach to this problem. Since the only minority sample to which we should pay more attention is incorrectly classified. For this reason, different methods must be defined and used to validate the algorithms for handling class imbalance problems appropriately. In this paper, we study the two-class problems, in which the minority class is considered to be the positive class. Hence, the confusion matrix of a two-class problem shows the results of correctly and incorrectly classified samples of each class [6], as shown in Table 1.

In the literature, Receiver Operating Characteristics (ROC) curve evaluation metric has been proposed by many researchers for class imbalance problem. ROC makes use of the proportion of True Positive Rate (TPR) and False Positive Rate (FPR) (refer to Eqs (2) and (3)).

$$\text{TPR} = \text{TP}/(\text{TP} + \text{FN}) \quad (2)$$

$$\text{FPR} = \text{FP}/(\text{FP} + \text{TN}) \quad (3)$$

The ROC graph is formed by plotting TPR over FPR as shown in Fig. 3, and any point in ROC space corresponds to the performance of a single classifier on a given distribution. The ROC curve is useful because it

Table 3
Classification performance of Boosting algorithms using AUC metric

Datasets	AdaBoost	RUSBoost	SMOTEBoost	NF-OS Boost
ecoli0vs1	0.6354	0.794	0.799	0.992
ecoli1	0.6354	0.794	0.799	0.992
ecoli2	0.6354	0.794	0.799	0.992
ecoli3	0.6354	0.794	0.799	0.992
glass0	0.947	0.947	0.947	0.948
glass0123vs456	0.946	0.946	0.946	0.946
glass1	0.946	0.946	0.946	0.946
glass6	0.947	0.918	0.991	0.997
Haberman	0.947	0.656	0.947	0.942
iris0	0.949	0.98	1	0.99
new-thyroid1	0.947	0.975	0.947	0.986
new-thyroid2	0.948	0.948	0.948	0.948
page-blocks0	0.637	0.953	0.967	0.996
Pima	0.6223	0.751	0.897	1
segment0	0.996	0.994	0.998	0.998
vehicle0	0.754	0.754	0.754	0.8754
vehicle1	0.754	0.768	0.897	1
vehicle2	0.854	0.966	0.967	1
vehicle3	0.745	0.763	0.894	1
Wisconsin	0.9	0.96	0.994	1
yeast1	0.7589	0.7382	0.741	0.996
yeast3	0.93	0.944	0.944	0.994
vowel0	0.996	0.9922	0.993	0.991
glass 4	0.8533	0.866	0.971	0.971
ecoli4	0.854	0.938	0.981	0.99

Table 4
Classification performance of Boosting algorithm using F-measure metric

Datasets	AdaBoost	RUSBoost	SMOTEBoost	NF-OS Boost
ecoli0vs1	0.6354	0.794	0.799	0.992
ecoli1	0.6354	0.794	0.799	0.992
ecoli2	0.635	0.794	0.799	0.992
ecoli3	0.6354	0.793	0.799	0.992
glass0	0.947	0.947	0.947	0.947
glass0123vs456	0.946	0.946	0.946	0.946
glass1	0.946	0.946	0.946	0.946
glass6	0.947	0.918	0.991	0.997
Haberman	0.947	0.656	0.947	0.942
iris0	0.949	0.98	1	0.98
new-thyroid1	0.947	0.975	0.947	0.986
new-thyroid2	0.948	0.948	0.948	0.948
page-blocks0	0.637	0.953	0.967	0.996
Pima	0.6223	0.751	0.897	1
segment0	0.996	0.994	0.998	0.998
vehicle0	0.754	0.754	0.754	0.8754
vehicle1	0.754	0.768	0.897	1
vehicle2	0.854	0.966	0.967	1
vehicle3	0.745	0.763	0.894	1
Wisconsin	0.9	0.96	0.994	1
yeast1	0.7589	0.7382	0.741	0.996
yeast3	0.93	0.944	0.944	0.994
vowel0	0.996	0.9922	0.992	0.991
glass 4	0.8533	0.856	0.971	0.971
ecoli4	0.854	0.938	0.981	0.99

provides a visual representation of the relative tradeoffs between the benefits (represented by true positives) and costs (represented by false positives) of classification with respect to data distributions [29]. Here, AUC is defined as the area under the ROC curve, which has been proved to be a reliable evaluation criterion and used as a metric to measure the efficiency against imbalanced classification problems. Two other important evaluation metrics [6] for imbalanced classification problems are defined as follows:

$$\text{F-measure} = \begin{aligned}[t] &(2 * \text{precision} * \text{recall})/ \\ &(\text{precision} + \text{recall}) \end{aligned} \tag{4}$$

Where precision is defined as TP by TP and FP and recall is defined as TP by TP and FN. Note that TPR and FPR has been discussed in detail in [9,34].

6.3. Dataset used

In this work, we test the proposed method on 25 benchmark datasets from KEEL-dataset repository with different imbalance ratio shown in Table 2.

6.4. Results

This section discusses the results of the proposed metric in terms of AUC and F-measure.

Table 5
Classification performance of Bagging algorithm using AUC metric

Datasets	BAG	OverBAG	SMOTEBAG	NF-OS BAG
ecoli0vs1	0.973	0.985	0.971	0.992
ecoli1	0.955	0.963	0.9636	0.992
ecoli2	0.9406	0.9406	0.952	0.952
ecoli3	0.935	0.9356	0.93	0.933
glass0	0.947	0.947	0.947	0.948
glass0123vs456	0.946	0.946	0.946	0.946
glass1	0.946	0.946	0.946	0.946
glass6	0.947	0.918	0.991	0.997
Haberman	0.947	0.656	0.947	0.942
iris0	0.949	0.98	1	0.99
new-thyroid1	0.947	0.975	0.947	0.986
new-thyroid2	0.948	0.948	0.948	0.948
page-blocks0	0.637	0.953	0.967	0.996
Pima	0.6223	0.751	0.897	1
segment0	0.996	0.994	0.998	0.998
vehicle0	0.754	0.754	0.754	0.8754
vehicle1	0.754	0.768	0.897	1
vehicle2	0.854	0.966	0.967	1
vehicle3	0.745	0.763	0.894	1
Wisconsin	0.9	0.96	0.994	1
yeast1	0.7589	0.7382	0.741	0.996
yeast3	0.93	0.944	0.944	0.994
vowel0	0.996	0.9922	0.993	0.991
glass 4	0.8533	0.866	0.971	0.971
ecoli4	0.854	0.99	0.981	0.99

6.4.1. AUC as a metric

Here, Table 3 shows the classification performance in terms of AUC obtained using different classifica-

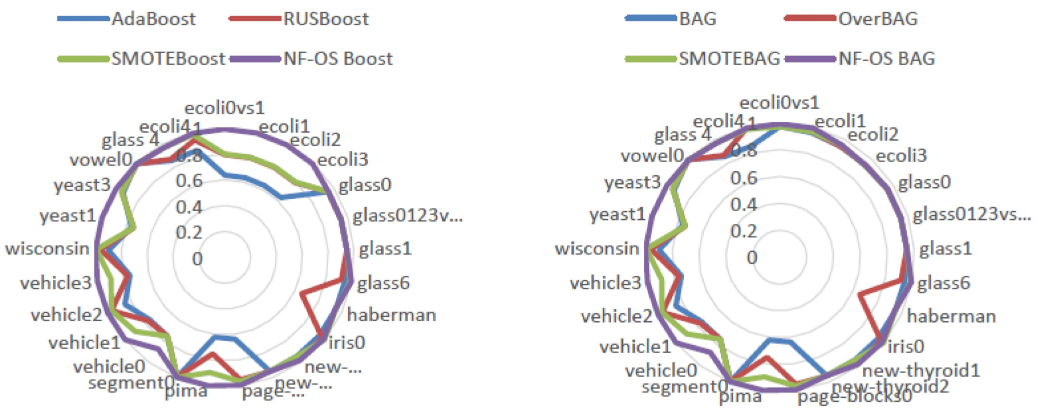

Fig. 4. AUC graph for NF-OS with Boosting and Bagging.

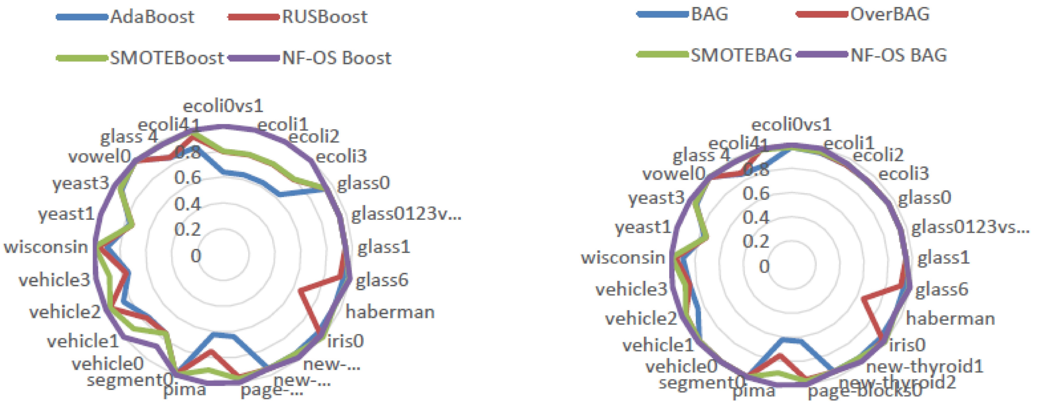

Fig. 5. F-measure graph for NF-OS with Boosting and Bagging.

tion techniques using Boosting. As indicated by the results, the proposed NF-OS with boosting demonstrated the best performance on 17 out of 25 datasets in terms of AUC for almost many datasets. Similarly, Table 5 shows the classification performance with respect to Bagging. The AUC results show a better performance for 14 datasets out of 25. The Fig. 4 presents the AUC results of both Boosting and Bagging in the graph representation.

6.4.2. *F-measure as a metric*

Here, Table 4 shows the classification performance in terms of F-measure obtained using different classification techniques using Boosting. As indicated by the results, the proposed NF-OS with boosting demonstrated the best performance on 16 out of 25 datasets in terms of F-measure for almost many datasets. Similarly, Table 6 shows the classification performance with respect to Bagging. The F-measure results better for 12 datasets out of 25. The Fig. 5 presents the F-Measure results with Boosting and Bagging in the graph representation.

All these experiments were tested on the Weka [25]. These ensemble algorithms are combined with oversampling techniques. One group includes oversampling with bagging (Bagging, OverBag, SMOTEBag) and another group includes both undersampling and oversampling with boosting (AdaBoost, RUSBoost, SMOTEBoost). We compared the results with our proposed methods with NF-OS Bagging and NF-OS Boosting. The results reveal that application of noise filtering with ensemble learning to imbalanced data will improve the performance of the classifiers. We also observed that on noisy data bagging technique outperformed than boosting. Hence, this section discussed several simulation results with several parameters like AUC, F-Measure, etc., and provides efficient and scalable results for class imbalance problems. Now next section will conclude this work with some future enhancements in brief.

Table 6
Classification performance of Bagging algorithm using F-measure metric

Datasets	BAG	OverBAG	SMOTEBAG	NF-OS BAG
ecoli0vs1	0.973	0.985	0.971	0.992
ecoli1	0.955	0.963	0.9636	0.992
ecoli2	0.9406	0.9406	0.952	0.952
ecoli3	0.935	0.9356	0.93	0.933
glass0	0.947	0.947	0.947	0.948
glass0123vs456	0.946	0.946	0.946	0.946
glass1	0.946	0.946	0.946	0.946
glass6	0.947	0.918	0.991	0.997
Haberman	0.947	0.656	0.947	0.942
iris0	0.949	0.98	1	0.99
new-thyroid1	0.947	0.975	0.947	0.986
new-thyroid2	0.948	0.948	0.948	0.948
page-blocks0	0.637	0.953	0.967	0.996
pima	0.6223	0.751	0.897	1
segment0	0.996	0.994	0.998	0.998
vehicle0	0.989	0.989	0.991	0.991
vehicle1	0.989	0.989	0.989	1
vehicle2	0.854	0.966	0.967	1
vehicle3	0.857	0.857	0.894	1
Wisconsin	0.9	0.96	0.994	1
yeast1	0.7589	0.7382	0.741	0.996
yeast3	0.93	0.944	0.944	0.994
vowel0	0.996	0.9922	0.993	0.991
glass 4	0.8533	0.866	0.971	0.971
ecoli4	0.854	0.99	0.981	0.99

7. Conclusion and future work

Due to generating a lot of data virtually or online, balancing this huge data or analysing this data have raised several problems. In literature, we have discussed that no major work has been done with respect to/overcome this problem. Hence, this work presents a novel approach for removing noise from the datasets using Noise Filtering (NF) and also to deal with an imbalanced (classification) problem by performing SMOTE after NF. Hence in this work, before training a classifier, NF first filter the noisy samples from the original dataset using K-NN technique, and then use the new minority and majority dataset to train a classifier. The simulation results (using Weka tool) over 25 datasets shows outperform of NF-OS with Boosting and bagging on AUC and F-measure. Also, this work provides comparison among AdaBoost, RUSBoost, SMOTEBoost, Bagging, OverBagging, and SMOTEBagging techniques, which produces the best results. Further for future work, we can extend this/our work with several real world problems like balancing the Facebook data/twitter data, Google searched/communicated data, etc. So, all the researchers, who are working in/related to this problem/area are kindly invited to do their research in this area.

Acknowledgments

This research work is funded by the Koneru Lakshmaiah Education Foundation, Lingaya's Vidyapeeth and Anumit Academy's Research and Innovation Network (AARIN), India. The authors would like to thank Koneru Lakshmaiah Education Foundation, Lingaya's Vidyapeeth and AARIN, India, an education foundation body and a research network for supporting the project through its financial assistance.

Conflict of interest

The authors have used reference number 9 and 34 as self-citation of their (his/her) work. No author has an objection of citing their work.

References

[1] A. Estabrooks, T. Jo and N. Japkowicz, A multiple resampling method for learning from imbalanced data sets, *Computational intelligence* **20**(1) (2004), 18–36.

[2] A. Fernández, S. García, M.J. del Jesus, and F. Herrera, A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets, *Fuzzy Sets and Systems* **159**(18) (2008), 2378–2398.

[3] A. FernáNdez, V. LóPez, M. Galar, M.J. Del Jesus and F. Herrera, Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches, *Knowledge-based systems* **42** (2013), 97–110.

[4] B. Wang and J. Pineau, Online bagging and boosting for imbalanced data streams, *IEEE Transactions on Knowledge and Data Engineering* (2016), 1–1.

[5] B. Zhu, B. Baesens and S.K. van den Broucke, An empirical comparison of techniques for the class imbalance problem in churn prediction, *Information sciences* **408** (2017), 84–99.

[6] C. Elkan, The foundations of cost-sensitive learning, in: *International Joint Conference on Artificial Intelligence*, Vol. 17, Lawrence Erlbaum Associates Ltd, 2001, pp. 973–978.

[7] 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**(1) (2010), 185–197.

[8] D.M. Farid, A. Nowé and B. Manderick, A new data balancing method for classifying multi-class imbalanced genomic data, in: *25th Belgian-Dutch Conference on Machine Learning (Benelearn)*, 2016, pp. 1–2.

[9] G. Rekha, A.K. Tyagi and V.K. Reddy, A novel approach to solve class imbalance problem using noise filter method, in: *ISDA 2018*, VIT Vellore, India.

[10] G. Haixiang, L. Yijing, L. Yanan, L. Xiao and L. Jinling, BPSO-Adaboost-KNN ensemble learning algorithm for multiclass imbalanced data classification, *Engineering Applications of Artificial Intelligence* **49** (2016), 176–193.

[11] G.M. Weiss, Mining with rarity: A unifying framework, *ACM SIGKDD Explorations Newsletter* **6**(1) (2004), 7–19.

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

[13] I. Visentini, L. Snidaro and G.L. Foresti, Diversity-aware classifier ensemble selection via f-score, *Information Fusion* **28** (2016), 24–43.

[14] J. Alcalá-Fdez, L. Sánchez, M.J. del Jesus, S. Garcia, S. Ventura, J.M. Garrell, J. Otero, C. Romero, J. Bacardit, V.M. Rivas et al., KEEL: A software tool to assess evolutionary algorithms for data mining problems, *Soft Computing* **13**(3) (2009), 307–318.

[15] J.A. Sáez, J. Luengo, J. Stefanowski and F. Herrera, SMOTE-IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering, *Information Sciences* **291** (2015), 184–203.

[16] J.F. Díez-Pastor J.J. Rodríguez, C.I. García-Osorio and L.I. Kuncheva, Diversity techniques improve the performance of the best imbalance learning ensembles, *Information Sciences* **325** (2015), 98–117.

[17] J. Stefanowski, Dealing with data difficulty factors while learning from imbalanced data, in: *Challenges in Computational Statistics and Data Mining*, Springer, 2016, pp. 333–363.

[18] J. Van Hulse, T.M. Khoshgoftaar and A. Napolitano, Experimental perspectives on learning from imbalanced data, in: *Proceedings of the 24th international conference on Machine learning*, ACM, 2007, pp. 935–942.

[19] J. Van Hulse, T.M. Khoshgoftaar and A. Napolitano, A novel noise filtering algorithm for imbalanced data, in: *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on*, IEEE, 2010, pp. 9–14.

[20] M. Alibeigi, S. Hashemi and A. Hamzeh, DBFS: An effective density based feature selection scheme for small sample size and high dimensional imbalanced data sets, *Data and Knowledge Engineering* **81** (2012), 67–103.

[21] M. Fanrong, G. Chunxiao and L. Bing, Fuzzy possibilistic support vector machines for class imbalance learning, *Journal of Convergence Information Technology* **8**(3) (2013).

[22] 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**(4) (2012), 463–484.

[23] M. Galar, A. Fernández, E. Barrenechea, H. Bustince and F. Herrera, Ordering-based pruning for improving the performance of ensembles of classifiers in the framework of imbalanced datasets, *Information Sciences* **354** (2016), 178–196.

[24] M. Galar, A. Fernández, E. Barrenechea and F. Herrera, EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling, *Pattern Recognition* **46**(12) (2013), 3460–3471.

[25] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten, The WEKA data mining software: An update, *ACM SIGKDD Explorations Newsletter* **11**(1) (2009), 10–18.

[26] 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.

[27] 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, pp. 107–119.

[28] N. Otsu, A threshold selection method from gray-level histograms, *IEEE Transactions on Systems, Man, and Cybernetics* **9**(1) (1979), 62–66.

[29] Q. Kang, B. Huang and M. Zhou, Dynamic behavior of artificial Hodgkin-Huxley neuron model subject to additive noise, *IEEE Transactions on Cybernetics* **46**(9) (2016), 2083–2093.

[30] Q. Wang, Z. Luo, J. Huang, Y. Feng and Z. Liu, A novel ensemble method for imbalanced data learning: Bagging of extrapolation-SMOTE SVM, *Computational Intelligence and Neuroscience* **2017** (2017).

[31] R. Barandela, R.M. Valdovinos and J.S. Sánchez, New applications of ensembles of classifiers, *Pattern Analysis and Applications* **6**(3) (2003), 245–256.

[32] S. Ali, A. Majid, S.G. Javed and M. Sattar, Can-CSC-GBE: Developing cost-sensitive classifier with gentleboost ensemble for breast cancer classification using protein amino acids and imbalanced data, *Computers in Biology And Medicine* **73** (2016), 38–46.

[33] S. Wang and X. Yao, Diversity analysis on imbalanced data sets by using ensemble models, in: *Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on*, IEEE, 2009, pp. 324–331.

[34] T. Amit Kumar and G. Rekha, Machine learning with big data, in: *Proceedings of International Conference on Sustainable Computing in Science, Technology and Management*, Elsevier, 2019.

[35] W. Lu, Z. Li and J. Chu, Adaptive ensemble undersampling-boost: A novel learning framework for imbalanced data, *Journal of Systems and Software* **132** (2017), 272–282.

[36] X.-Y. Liu, J. Wu and Z.-H. Zhou, Exploratory undersampling for class-imbalance learning, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **39**(2) (2009), 539–550.

[37] X. Zhu and X. Wu, Class noise vs. attribute noise: A quantitative study, *Artificial Intelligence Review* **22**(3) (2004), 177–210.

[38] Y. Freund, R.E. Schapire et al., Experiments with a new boosting algorithm, in: *Icml*, Vol. 96, Citeseer, 1996, pp. 148–156.