Applied Soft Computing Journal 78 (2019) 141–149

Contents lists available at ScienceDirect

Applied Soft Computing Journal

journal homepage: www.elsevier.com/locate/asoc

SSO_{Maj}-SMOTE-SSO_{Min}: Three-step intelligent pruning of majority and minority samples for learning from imbalanced datasets

Seba Susan [*,1], Amitesh Kumar [2]

Department of Information Technology, Delhi Technological University, Delhi-42, India

HIGHLIGHTS

- A new undersampling–oversampling hybrid learning technique for Imbalanced datasets.
- Three-phase process involving intelligent undersampling twice applied with an intermediate oversampling stage.
- Particle swarm optimization used to optimize the search space for the locating the relevant majority class samples.
- Higher accuracies achieved by the proposed hybrid technique.

ARTICLE INFO

Article history:
Received 8 August 2018
Received in revised form 25 November 2018
Accepted 15 February 2019
Available online 19 February 2019

无能为力
求助于

Keywords:
Imbalanced learning
Sample subset optimization
Particle swarm optimization
SMOTE
Oversampling
Undersampling

ABSTRACT

Real world datasets, particularly in the current context of Big Data applications, suffer from the problem of imbalanced representation of samples from different categories. Most classifiers and learning techniques are inept to deal with this problem, with the majority of them tending to overlook the issue. Typical data balancing methods in literature resort to data sampling that constitutes of either undersampling the majority class samples or oversampling the minority class samples. An intelligent combination of undersampling the majority class and oversampling the minority class is expected to improve the learning performance. In this paper, data balancing is achieved prior to classification, through a novel three-step sequence of intelligent undersampling of the majority class followed by the oversampling of the minority class, which is further followed by the intelligent undersampling of the minority class that has now become the majority class due to the oversampling. The recently proposed Sample Subspace Optimization (SSO) that uses Particle Swarm Optimization (PSO) as an intelligent agent to find globally optimum solutions in the search space, is our choice for the intelligent undersampling technique. The oversampling in the second step is achieved through Synthetic Minority Oversampling (SMOTE) as well as intelligent variants such as Borderline SMOTE, ADASYN and MWMOTE. The increase in computational complexity is compensated by the higher performance achieved due to relevant sampling. Experiments on benchmark datasets from the UCI repository establish the efficiency of our three-step approach SSO_{Maj}–SMOTE–SSO_{Min} as observed from the higher AUC scores from the Receiver Operating Characteristics.

1. Introduction

Real world datasets, particularly in the current context of Big Data applications, suffer from the problem of imbalanced representation of samples from different categories. Most classifiers and learning techniques are inept to deal with this problem, with the majority of them tending to overlook the issue. Examples of imbalanced datasets could be observed from the ever-transient social media platforms like Facebook, where some users have thousands of friends or followers while some others have just a handful few. Learning from such imbalanced data is challenging and requires preprocessing and special treatment of the training data that will restore the balance between the majority and the minority classes prior to the machine learning phase [1]. Sampling techniques are the conventional and most primitive treatment for imbalanced data [2]. Some of the earliest solutions suggested were random oversampling of the minority class and random undersampling of the majority class [3]. Synthetic Minority Oversampling (SMOTE) proposed by Chawla et al. [4], is an extension of this technique, in which extra minority samples are generated along the line segment between two available minority samples albeit with no reference whatsoever to the samples available in the adversarial majority class. The advent

[*] Corresponding author.
E-mail address: seba_406@yahoo.in (S. Susan).
[1] Associate Professor.
[2] Post-graduate student.

https://doi.org/10.1016/j.asoc.2019.02.028
1568-4946/© 2019 Elsevier B.V. All rights reserved.

of SMOTE was followed by a phase of intelligent sampling techniques in which the sampling in one class (minority or majority) is done with the prior knowledge of the samples available in the opposite class, at least those samples that are in proximity with the decision boundary [5,6]. Such a precaution was proved necessary to generate or retain only those samples that would discriminate against their counterparts in the opposite class. Some of the popular intelligent oversampling techniques that succeeded SMOTE were Borderline SMOTE [7], Adaptive Synthetic sampling (ADASYN) [8] and Majority weighted minority oversampling technique (MWMOTE) [9]. In Borderline SMOTE, only the samples near the decision boundary are oversampled. ADASYN involves identifying the minority samples that are more difficult to learn and shifting the decision boundary towards the difficult-to-learn samples. MWMOTE shortlists informative minority samples based on their distances to majority samples, and then applies clustering to generate synthetic minority samples from the shortlisted ones. Solutions to intelligent undersampling of the minority class are also available in literature such as the One-Sided-Selection (OSS) that involves data cleansing [10] and evolutionary algorithm based undersampling in [11]. The performance of the Receiver Operating Characteristics (ROC) for different sample subsets in different folds are compared in [12] to decide on the optimal subset that contains equal samples from the minority and majority class and yet yields the highest accuracy. This method called Sample Subspace Optimization (SSO) uses the Particle Swarm Optimization (PSO) [13] as the intelligent agent for optimizing the search space. Intelligent swarm algorithms that seek to find the global solution to optimization problems have been used for feature selection algorithms [14–16] and spam detection [17]. SSO serves as a precursor to a whole lot of recent researches that seeks the globally optimum solution for the ideal sample subspace for the imbalanced dataset that could be applied for training the machine learning algorithms. Interestingly, new trends have emerged in recent times that do not rely on sampling for data imbalance treatment but instead integrate the data balancing in the learning algorithm itself which is however specific to the classifier being used. Classifiers such as Extreme Learning Machines (ELM) [18] and Support Vector Machines (SVM) [19] have incorporated changes in their learning algorithms for dealing with imbalanced datasets. Recently, imbalanced evolving self-organizing maps (IESOMs) were proposed in [20], that extracts useful information from the minority class using separate positive and negative SOMs for training the minority and majority classes respectively. Cost Sensitive Learning (CSL) is another alternative to sampling, that computes the cost matrix for misclassifying observations [21] that works fine for rare classes [22].

In our work, we propose a new technique $\mathrm{SSO_{Maj}}$–SMOTE–$\mathrm{SSO_{Min}}$ that presents a unique combination of hierarchical undersampling with oversampling that will intelligently cut down and prune the training data in a manner such that both/all classes are equally represented and at the same time also provide discriminative information against each other. SSO optimized by PSO takes care of the intelligent undersampling (both of the majority class in the first step and the minority class after it becomes the new majority in the third step). The SMOTE technique and other intelligent variants are explored for the task of oversampling of the minority samples in the second step. The proposed sequential combination of sampling techniques provides an effective solution to the data imbalance problem prevalent in real-world datasets, since it prunes both the majority and minority classes and retains only samples in every class that contribute to class-discriminative information. Our paper is organized as follows: a literature survey pertaining to related work is presented in Section 2, the steps of the proposed intelligent pruning of the unbalanced data are discussed in detail in Section 3, the experimental setup and the results are analyzed in Section 4 and the overall conclusions are drawn in Section 5.

2. Literature survey

There are several works in literature that explore intelligent techniques for remedying the data imbalance problem. Estabrooks et al. in [23] conclude from their experimental studies that trying out different combinations of resampling is an ideal solution for correcting the imbalance problem. Also, in the same paper it was observed that the resampling rate is dependent on the specific dataset. More specifically, the paper suggested that the parallel outputs of oversampling and undersampling achieved under different sampling rates be applied to learn a variety of different classifiers. The outputs of the classifiers are then compared and only a single classifier among these would make the decision for deciding the class of the testing point. Two popular research directions taken while singularly investigating sampling approaches are: (a) to find the best possible data subset that could be applied for training (such as [12] that uses PSO for the search space and [24] based on the Geometric Mean) and (b) to find the best proportion of *majority:minority* samples that could be applied to classifiers for achieving high performance [25,26]. A recent trend in the arena of imbalanced learning is to find new intelligent sampling techniques [27] that would discriminate against the available samples of the opposite class in a binary classification problem.

In [28], sparsely sampled boundaries in the minority class are found in an iterative process to mitigate the dominant effect of the majority class in the deep learning framework. A cost function namely, Class Rectification Loss (CRL) was coined specially for the purpose. Liu et al. in [29] propose two intelligent undersampling techniques:- EasyEnsemble that selects the best optimal subset of majority samples using separate classifiers, and BalanceCascade which is a step-by step sequential process for selecting the most useful majority samples. A PYTHON language specific tool *imbalanced-learn* was developed in [30] for the Imbalanced Learning problem that tries out a combination of undersampling followed by oversampling in order to overcome the overfitting that happens due to the oversampling [31]. Software packages in R that incorporate solutions to data imbalance include ROSE [32] and CARET [33]. ROSE or Random Oversampling Examples package generates synthetic samples using a smoothed bootstrap approach and handles parallelly both issues of model estimation and accuracy evaluation. An intelligent oversampling technique Synthetic Informative Minority Oversampling (SIMO) was introduced in [19] specifically for the Support Vector Machine (SVM) classifiers wherein only the minority samples close to the SVM decision boundary are oversampled. Some other hybrid approaches for intelligent sampling that have been introduced recently are Random Hybrid Sampling [34], Bagging of Extrapolation Borderline-SMOTE SVM [35] and Neighbors Progressive Competition (NPC) algorithm [36] that build up on the k-NN classifier strategy for searching out those samples that are closest to the query sample. Some hybrid techniques for sampling with boosting are RUSBoost [37] and SMOTEBoost [38] that combine random undersampling with SMOTE and AdaBoost respectively.

Some other popular hybrid oversampling and undersampling combinations that have gained popularity over time are:- SMOTE–RSB [39] that constructs new samples using SMOTE followed by editing with Rough Set Theory, Chris et al.'s hybrid sampling approach [40] and SMOTE–ENN [41] and Random Undersampling with Tomek Link [42] combinations. SMOTE–Tomek combination [43] is popular in data science experiments since it finds the nearest neighbor from the opposite class and removes the majority instance as an undersampling measure.

3. Proposed intelligent pruning of the imbalanced dataset

相互影响

基本成分

In this Section, we describe the interplay of SSO and SMOTE that are the basic constituents in our proposed hybrid $\mathrm{SSO_{Maj}}$–SMOTE–$\mathrm{SSO_{Min}}$. Intelligent variants of oversampling are also substituted instead of SMOTE as the intermediary step. Motivated by several works discussed in previous sections on creating hybrids of oversampling and undersampling procedures for imbalance treatment, we create an intelligent pruning process for the majority and minority classes through a unique sequential three-step process by: *undersampling the majority----oversampling the minority----undersampling the oversampled minority*, that we call aptly by the name $\mathrm{SSO_{Maj}}$–SMOTE–$\mathrm{SSO_{Min}}$. The detailed steps of this new technique are elucidated below.

最佳选择

STEP 1: Sample Subspace Optimization (SSO) [12] is an intelligent Majority undersampling technique that finds the optimum choice of majority samples which in combination with all the existing minority samples provide discriminative information against each other. The Particle Swarm Optimization (PSO) algorithm [13] is the intelligent agent embedded in SSO that achieves this task. PSO identifies those majority samples in fold *i* that would take part in the classification. In Binary PSO (BPSO), each particle position is an indicator function $\left\{x_{i,d}\right\} = I_1, I_2, \ldots I_d$ where each value equals 1 or 0 to indicate if this majority sample is included in the training or not, assuming *d* is the number of majority samples in the imbalanced training set **TRAIN**. All the minority samples are included as it is, in the training. The BPSO is used for optimizing the SSO fitness function which is the k-fold cross-validation error term defined by

$$\hat{\varepsilon} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n/k}\left\|y_j^{(i)} - p\left(t_j^{(i)}|\Theta^{(i)}, x_j^{(i)}\right)\right\| \tag{1}$$

For *n* number of samples and *k* number of folds. $y_j^{(i)}$ is the target class label of the sample *j* from the test fold *i*. $p(t_j^{(i)}|\Theta^{(i)}, x_j^{(i)})$ indicates the prediction from the model p(.) whose parameters $\Theta^{(i)}$ are derived from the training phase given the feature vector $x_j^{(i)}$. The notation $t_j^{(i)}$ denotes the predicted class label for sample *j*.

The problem in (1) can be interpreted as a maximization one, by modifying the cost function in (1) from $\hat{\varepsilon}$ to $1 - \hat{\varepsilon}$. For each test fold *i*, the optimal sample space subset is the combination $\left\{Smaj^{(i)}, Smin\right\}$ where *Smin* is the original minority sample set to be retained in every fold, and $Smaj^{(i)}$ is the undersampled majority subset for the *ith* fold that is optimized from the search space by BPSO. The particles in the PSO constitute the population in the course of a run, and they move with updated velocities. The velocity update formula of BPSO seeks to move each particle position towards the local best solution for each particle *lbest* and towards the global best solution *gbest* among all particles as shown below.

$$v_{i,d} \longleftarrow \omega v_{i,d} + \phi_p r_p \left(lbest - x_{i,d}\right) + \phi_g r_g \left(gbest - x_{i,d}\right) \tag{2}$$

In (2), ω is a constant called the inertia weight, ϕ_p and ϕ_g are the cognitive and social coefficients respectively, and r_p and r_g are random numbers in the range [0, 1].

The PSO parameters $\left\{\omega, \phi_p, \phi_g, r_p, r_g\right\}$ in (2) are preset to fixed values as per the guidelines in [12], and choices of these parameters for our experiments are explained in our discussion on Results in Section 4 as well. The iterative optimization procedure is prescribed for a fixed number of iterations and the optimal sample subset is defined as the sample subset from the very last iteration.

STEP 2: Synthetic Minority Oversampling technique (SMOTE) which is our intermediate step, finds the k-nearest neighbors $\{\boldsymbol{s_{k-nn}}\}$ of each minority sample $\boldsymbol{s_{current}}$ and creates extra samples along the line segment joining the two samples as per the following equation

$$\boldsymbol{s_{new}} = \boldsymbol{s_{current}} + (\boldsymbol{s_{current}} - \boldsymbol{s_{k-nn}}) \times \delta \tag{3}$$

where s ∈ Smin the minority class and δ is a random value in the range of 0 to 1.

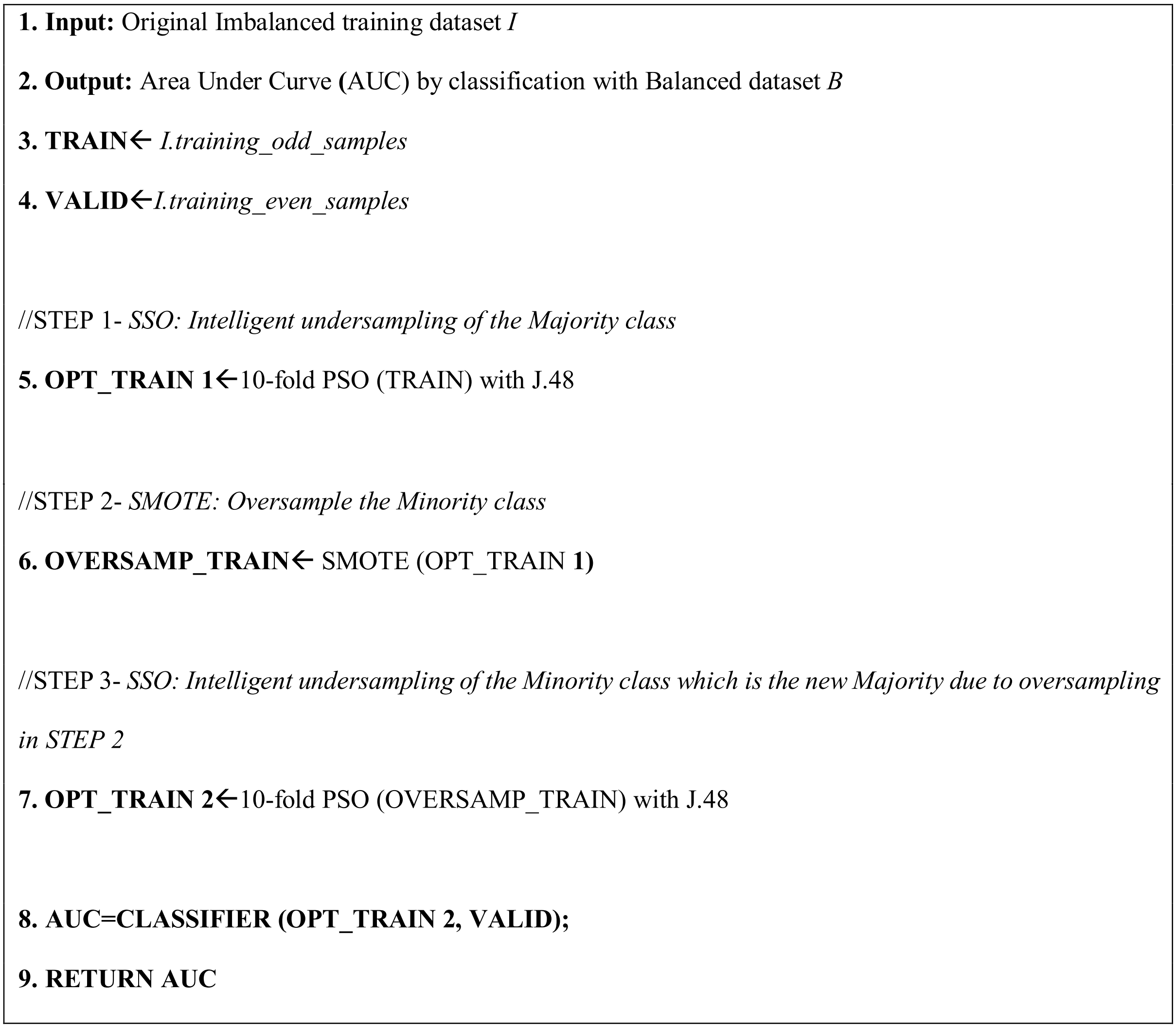
STEP 3: Oversampling the data leads to overfitting, hence we need some kind of pruning for retaining only those oversampled minority instances in the end that could discriminate against the majority class. So we proceed for intelligent undersampling of the oversampled minority class as the third and final step in our three-step procedure. The fitness function is the same as in (1), only that this time, the sample search space is that belonging to the oversampled minority class. The overall pseudo-code for the proposed $\mathrm{SSO_{Maj}}$–SMOTE–$\mathrm{SSO_{Min}}$ hybrid technique is summarized in Fig. 1.

The 10-fold cross-validation procedure is followed to optimize the search space and evaluate accuracies. The Area Under Curve (AUC) of the Receiver Operating Characteristics curve is the performance criteria for the selection of the optimum sample subset. Since the samples near the decision boundary contain discriminative information and creation of synthetic samples correct the skewness in the decision boundary [35], hence Borderline SMOTE is also investigated as an option for the oversampling instead of SMOTE in the second step of our process. ADASYN and MWMOTE intelligent oversampling techniques are also tested as a substitution to SMOTE. This augments our proposed list of hybrids to include $\mathrm{SSO_{Maj}}$–Borderline SMOTE–$\mathrm{SSO_{Min}}$, $\mathrm{SSO_{Maj}}$–ADASYN–$\mathrm{SSO_{Min}}$, and $\mathrm{SSO_{Maj}}$–MWMOTE–$\mathrm{SSO_{Min}}$.

4. Experimental results and discussions

This section analyzes the effectiveness of the proposed hybrid of *undersampling–oversampling–undersampling* in correcting the data imbalance problem. The techniques introduced in the previous section are applied to balance the training set with equal number of majority and minority samples. The classifier used is J.48 decision trees since they are found apt by researchers for oversampling and undersampling experiments [44]. The experiments were conducted in JAVA programming platform and the IDE: Eclipse Oxygen and R interpreter along with the Weka 3.8 data mining and visualization tool. The processor used was an Intel Core i3-4005U CPU at a clock frequency of 1.70 GHz. The parameters of PSO observed in Eq. (2) are preset to the following values: 100 iterations of the Particle Swarm Optimization are used with the cognitive and social acceleration constants ϕ_p, ϕ_g each set to 1.43 for an initial population of 100, as per the experimental guidelines for SSO (PSO) in [12]. The inertia weight ω is set to 0.689 and the velocity boundary is 0.018–0.982. Datasets from the UCI repository [45] selected for the experiments are- *Ionosphere*, *Diabetes*, *Bank Note*, *EEG Eye*, *Yeast*, *Abalone*, *Glass*, *Image Segmentation*, *Haberman's survival*, and from the ROSE package- *Hacide* dataset (ten datasets in all). The details of the ten datasets are outlined in Table 1. The imbalance ratio among the majority and minority classes can be observed from this Table.

The Area Under Curve (AUC) of the ROC curves are summarized in Table 2 for all the datasets for the proposed hybrids of $\mathrm{SSO_{Maj}}$–SMOTE–$\mathrm{SSO_{Min}}$, $\mathrm{SSO_{Maj}}$–Borderline SMOTE–$\mathrm{SSO_{Min}}$, $\mathrm{SSO_{Maj}}$–ADASYN–$\mathrm{SSO_{Min}}$, and $\mathrm{SSO_{Maj}}$–MWMOTE–$\mathrm{SSO_{Min}}$. The AUC scores are used universally for studying the performance of the imbalanced datasets [12,43,44]. For the oversampling by SMOTE the number of nearest neighbors considered are K=9. Both 25% and



```
1. Input: Original Imbalanced training dataset I

2. Output: Area Under Curve (AUC) by classification with Balanced dataset B

3. TRAIN← I.training_odd_samples

4. VALID←I.training_even_samples


//STEP 1- SSO: Intelligent undersampling of the Majority class

5. OPT_TRAIN 1←10-fold PSO (TRAIN) with J.48


//STEP 2- SMOTE: Oversample the Minority class

6. OVERSAMP_TRAIN← SMOTE (OPT_TRAIN 1)


//STEP 3- SSO: Intelligent undersampling of the Minority class which is the new Majority due to oversampling
in STEP 2

7. OPT_TRAIN 2←10-fold PSO (OVERSAMP_TRAIN) with J.48


8. AUC=CLASSIFIER (OPT_TRAIN 2, VALID);

9. RETURN AUC
```

Fig. 1. Pseudo-code for the proposed $\mathrm{SSO_{Maj}}$–SMOTE–$\mathrm{SSO_{Min}}$ hybrid.

Table 1
Datasets used for the experimentation (Majority and Minority classes shown).

S. no.	Datasets	Majority samples	Minority samples	Imbalance ratio *Majority:Minority*
1	*Ionosphere*	127	48	2.64:1
2	*Bank Note*	381	305	1.25:1
3	*EEG Eye*	4128	3362	1.227:1
4	*Diabetes*	400	214	1.87:1
5	*Hacide*	980	20	49:1
6	*Yeast*	582	160	3.64:1
7	*Abalone*	343	23	14.9:1
8	*Glass*	82	25	3.28:1
9	*Image Segmentation*	990	165	6:1
10	*Haberman's survival*	115	38	3.03:1

85% increase in minority samples are considered for SMOTE as seen from Table 2.

Some specific observations made from the results in Table 2 are:

1. We compare our proposed methods with that of popular sampling techniques of SSO incorporating PSO [12] (--Undersampling), ADASYN [8] (--Oversampling) and MWMOTE [9] (--Oversampling) proposed recently for imbalance correction in datasets.
2. All methods outperform the baseline method of SSO [12] that performs intelligent undersampling of the majority class using PSO as the search space algorithm. In case of *Yeast*, SSO gives comparable performance to the proposed $\mathrm{SSO_{Maj}}$–Borderline SMOTE–$\mathrm{SSO_{Min}}$.
3. The proposed hybrids formed by the three-step pruning process give consistently high results for all the ten datasets. The proposed hybrid sampling of $\mathrm{SSO_{Maj}}$–SMOTE–$\mathrm{SSO_{Min}}$ (25%) and $\mathrm{SSO_{Maj}}$–SMOTE–$\mathrm{SSO_{Min}}$ (85%) give overall the best performance.

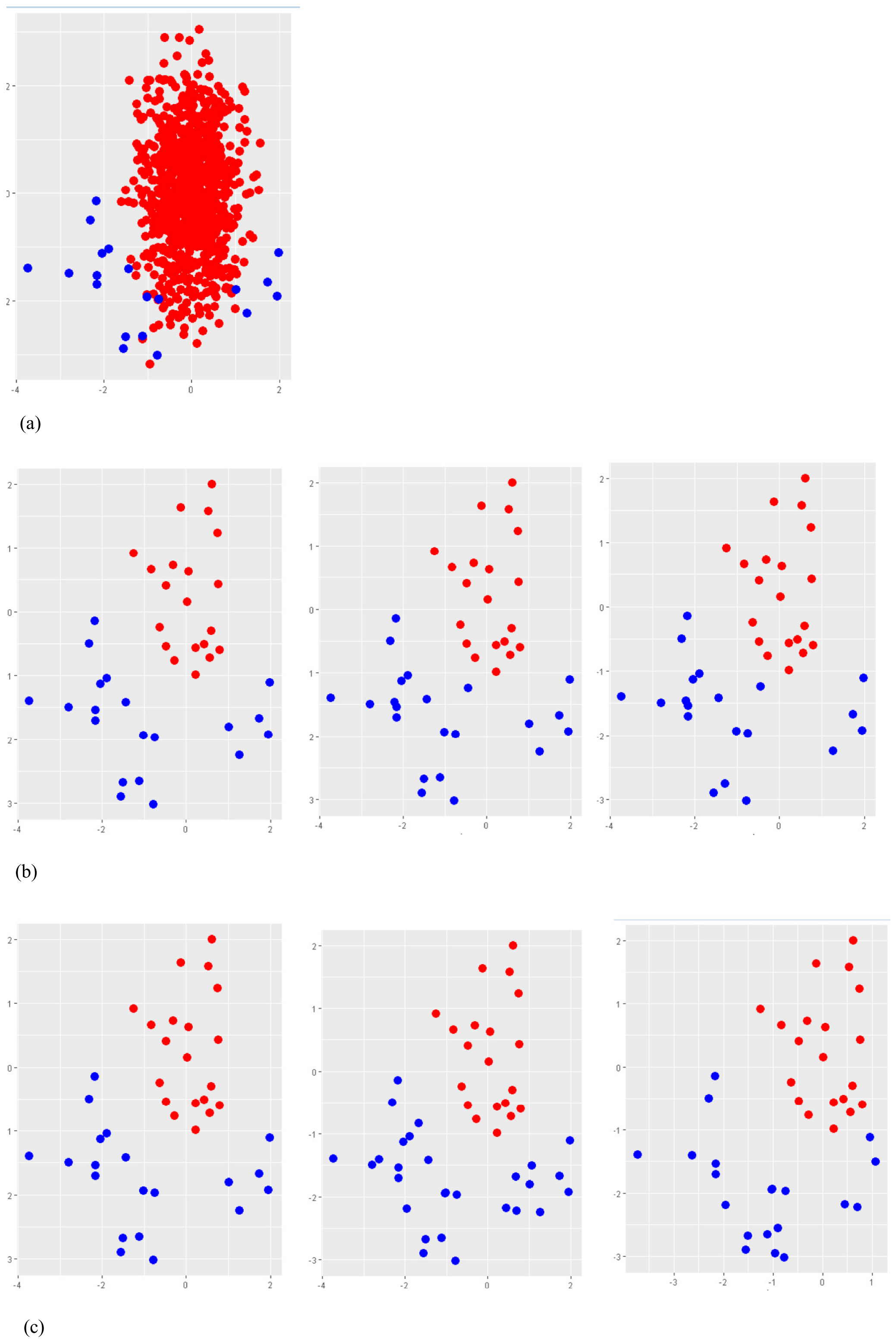


Fig. 2. The data distribution of *Hacide* before and after the balancing procedure for the proposed technique (RED: Majority sample, BLUE: Minority samples) (a) Original distribution (b) $\mathrm{SSO_{Maj}}$-SMOTE-$\mathrm{SSO_{Min}}$ (25% oversampling) (c) $\mathrm{SSO_{Maj}}$-SMOTE-$\mathrm{SSO_{Min}}$ (85% oversampling) (d) $\mathrm{SSO_{Maj}}$-Borderline SMOTE-$\mathrm{SSO_{Min}}$ (e) $\mathrm{SSO_{Maj}}$-ADASYN-$\mathrm{SSO_{Min}}$ (f) $\mathrm{SSO_{Maj}}$-MWMOTE-$\mathrm{SSO_{Min}}$ (*Stepwise for the proposed method in the order from left to right: After* $\mathrm{SSO_{Maj}}$ *- After SMOTE - After* $\mathrm{SSO_{Min}}$*).*

4. The improvement in AUC scores for our three-step sampling process is quite significant, over and above the baseline method, for all datasets especially for the severely unbalanced cases, the severity of imbalance depicted in Table 1. In case of the severely unbalanced *Hacide* dataset in Table 2, the experiments give the highest accuracies

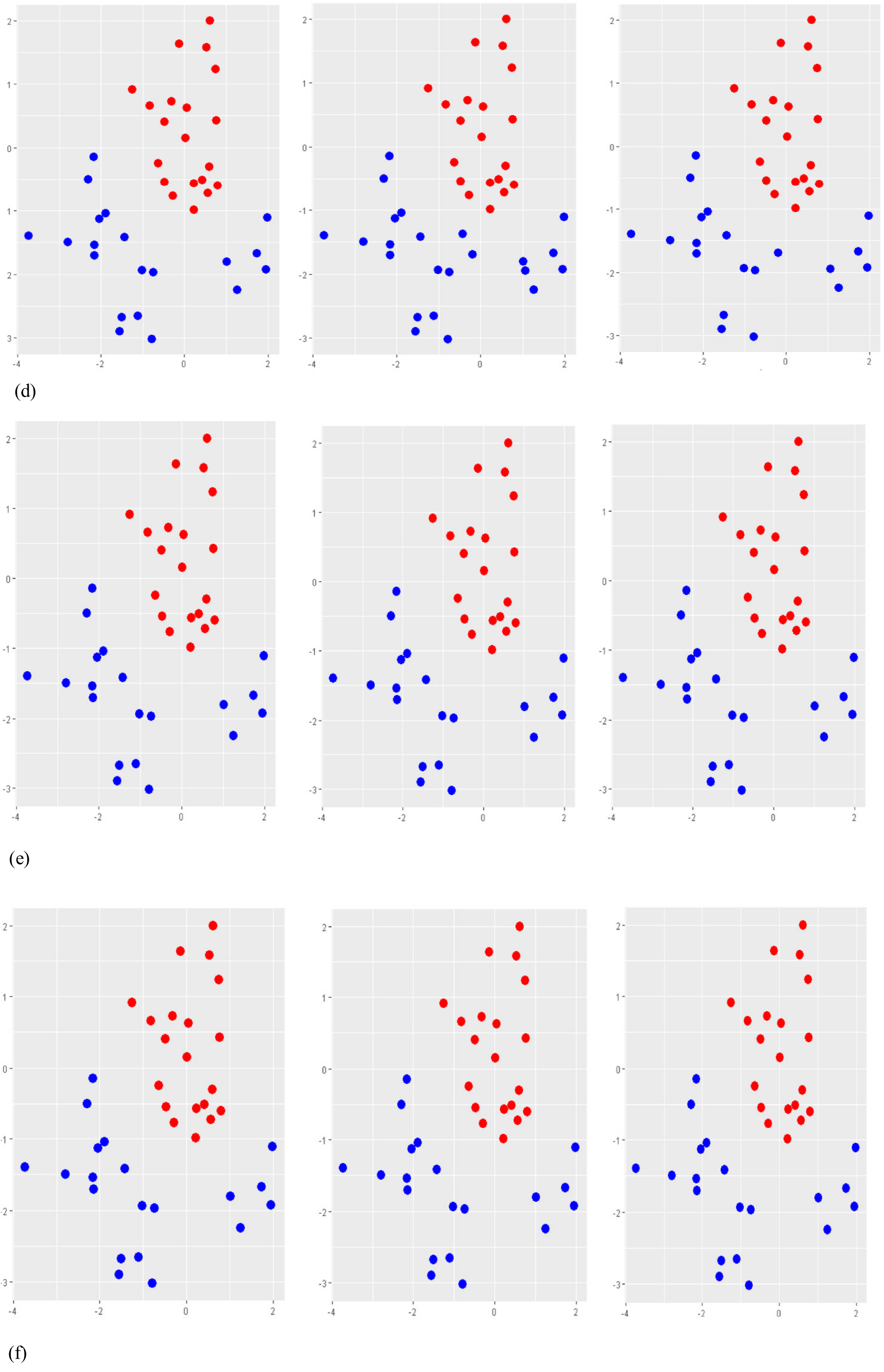


Fig. 2. (*continued*).

(AUC=1) for SSO_{Maj}–SMOTE–SSO_{Min} (for an 85%increase in the minority samples). Fig. 2 shows the visualization of this balancing act for *Hacide* for the three-step process of SSO_{Maj}–SMOTE–SSO_{Min}, SSO_{Maj}–Borderline SMOTE–SSO_{Min}, SSO_{Maj}–ADASYN–SSO_{Min}, and SSO_{Maj}–MWMOTE–SSO_{Min}.

5. In case of the lesser unbalanced dataset of *Ionosphere*, the best performance is noted for the SSO_{Maj}–SMOTE–SSO_{Min} (25%) from Table 2. A similar result is observed for the *Diabetes* dataset in Table 2 for which SSO_{Maj}–SMOTE–SSO_{Min} (25%) performs best.
6. For the larger-sized *EEG Eye* dataset, for which the degree of unbalance between classes is not so high but the number of samples are huge, in thousands range, the highest accuracies are obtained for the 85% SSO_{Maj}–SMOTE–SSO_{Min} case as observed from Table 2.

Table 2
Classification accuracies: Area Under Curve (AUC) from ROC curve analysis for different datasets (Proposed method: Highlighted in gray; Highest AUC scores for each dataset Highlighted in bold).

Methods	*Ionosphere*	*Bank Note*	*EEG Eye*	*Diabetes*	*Hacide*	*Yeast*	*Abalone*	*Glass*	*Image Seg.*	*Haberman's survival*
SSO [12]	0.8869	0.968	0.8096	0.7807	0.932	**0.879**	0.5989	0.9437	0.993	0.7122
ADASYN [8]	0.86	**0.9999**	0.7804	0.782	0.9674	0.8117	0.6395	0.8987	0.9989	0.6506
MWMOTE [9]	0.8822	0.9395	0.7953	0.7473	0.7644	0.8780	0.5034	0.9380	0.9929	0.6375
SSO_{Maj}-SMOTE-SSO_{Min} (25%)	**0.9163**	0.9932	0.8058	**0.8101**	0.95	0.8576	0.7365	0.9609	**1**	**0.7904**
SSO_{Maj}-SMOTE-SSO_{Min} (85%)	0.8945	0.9935	**0.8122**	0.7982	**1**	0.874	0.678	0.9585	**1**	0.7784
SSO_{Maj}-Borderline SMOTE-SSO_{Min}	0.8737	0.9955	0.8089	0.7909	0.95	**0.879**	0.7017	0.9612	0.988	0.7565
SSO_{Maj}-ADASYN-SSO_{Min}	0.8985	0.9926	0.8057	0.7859	0.95	0.8536	**0.7509**	0.9613	**1**	0.7677
SSO_{Maj}-MWMOTE-SSO_{Min}	0.8937	0.9948	0.8036	0.8091	0.95	0.8527	0.7481	**0.9975**	**1**	0.7807

7. For the lesser unbalanced case of *Bank Note* dataset, the ADASYN oversampling technique gives the best results as seen from Table 2 even though the proposed hybrids like SSO_{Maj}–Borderline SMOTE–SSO_{Min} also perform equally well.

The general observations about our approach, made from these tables are as follows: After the first step SSO_{Maj}, the number of majority samples are limited to the population of the minority class. After the second step of SMOTE, the number of minority samples increase (by a fixed proportion in SMOTE and a variable amount in Borderline SMOTE) more than the majority class. Thus the roles and definitions of the majority and minority are swapped due to the oversampling. Finally, the intelligent undersampling of the oversampled minority denoted by SSO_{Min} is carried out to complete the three-step process that defines our method. The optimal subset $\left\{Smaj^{(opt)}, Smin\right\}$ so obtained is next applied to the J.48 decision tree classifier for cross-validating the results with the test data, the AUC scores of which are summarized in Table 2. Our programs took hardly a second to execute varying slightly for different datasets with the timing complexity demonstrated in Fig. 3 for the *Hacide* dataset. The correction in imbalance achieved by SSO_{Maj}–SMOTE–SSO_{Min} on the imbalanced *Yeast* dataset is shown in Fig. 4. Fig. 4 shows the change in balance ratio for our approach shown for each step sequentially. The proposed method thus provides a viable solution to the data imbalance problem assured of high accuracies due to the global search strategy applied.

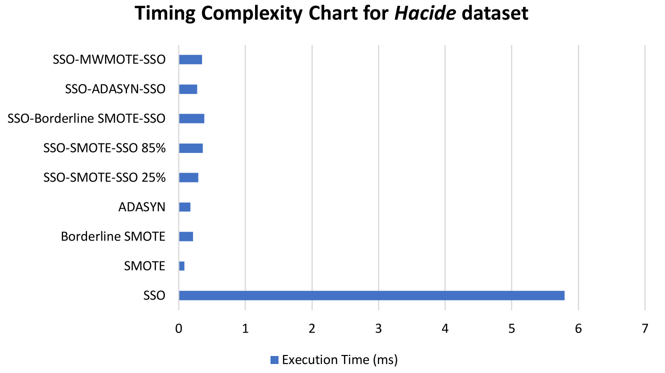


Fig. 3. Timing Complexity of different methods (in secs) for the *Hacide* dataset.

5. Conclusion and future directions

A novel three-step sampling for imbalanced datasets is proposed in this paper, as a correction treatment for data imbalance prior to the classification phase, a condition under which rarer classes have much fewer samples as compared to the others. Our method that we call SSO_{Maj}–SMOTE–SSO_{Min} follows the three-step procedure of *undersampling the majority----oversampling the minority----undersampling the oversampled minority*. Intelligent variants of oversampling such as Borderline SMOTE, ADASYN and MWMOTE are substituted instead of SMOTE for testing the performance. Experiments on benchmark datasets give higher Area Under the Curve (AUC) scores from the Receiver Operating Characteristics (ROC) for the proposed intelligent data pruning technique which prunes samples from the majority and minority classes into a refined sample space that provides class-discriminative information. The proposed sequential combination of sampling techniques provides an effective solution to the data imbalance problem prevalent in real-world datasets since it prunes both the majority and minority classes. Our algorithm could be directly applied to real-world imbalanced image datasets such as the LFW Face database where the facial images of some celebrities are huge in number while those of some lesser known individuals are too few. Our idea of pruning both classes in a step-by-step manner could be extended to several research directions in future such as general data cleansing and treatment, cost sensitive learning fundamentals and modern machine learning

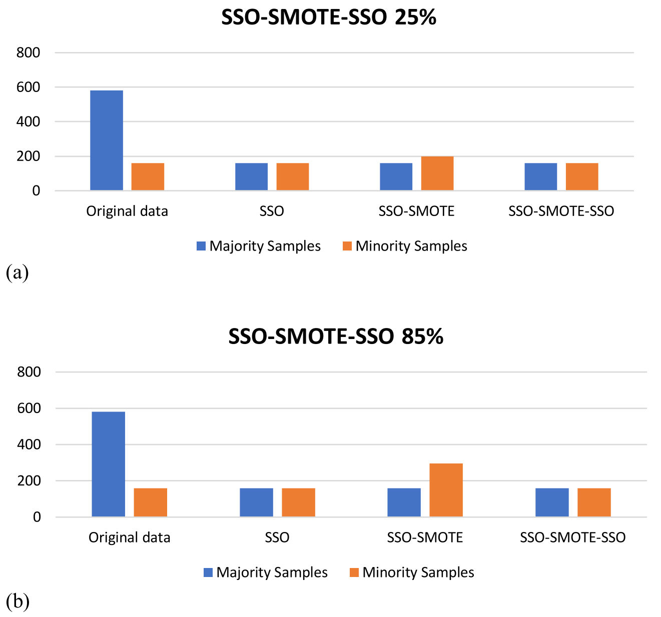


Fig. 4. Data Imbalance treatment by the proposed hybrid techniques of (a) SSO$_{Maj}$–SMOTE–SSO$_{Min}$ (25% oversampling) (b) SSO$_{Maj}$–SMOTE–SSO$_{Min}$ (85% oversampling) for the *Yeast* dataset.

algorithms like that of ELM, SVM, SOMs that incorporate some form of imbalance treatment. The cue obtained from this paper is that both the majority and minority classes contain redundant information, though often, it is only the majority classes that get pruned and the minority classes are either usually retained as it is. Even while oversampling, the synthetic samples generated are sometimes redundant and downsizing is never done. The work done in this paper thus encourages future research in providing preferential treatment sequentially to all classes whether rare or not, and retain only samples that provide good class discrimination for high performance classification.

References

[1] Haibo He, Yunqian Ma (Eds.), Imbalanced Learning: Foundations, Algorithms, and Applications, John Wiley & Sons, 2013.

[2] Nitesh V. Chawla, Data mining for imbalanced datasets: An overview, in: Data Mining and Knowledge Discovery Handbook, Springer, Boston, MA, 2009, pp. 875–886.

[3] Nathalie Japkowicz, Shaju Stephen, The class imbalance problem: A systematic study, Intell. Data Anal. 6 (5) (2002) 429–449.

[4] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.

[5] Seyda Ertekin, Jian Huang, Leon Bottou, Lee Giles, Learning on the border: active learning in imbalanced data classification, in: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, ACM, 2007, pp. 127–136.

[6] Alberto Fernández, Salvador Garcia, Francisco Herrera, Nitesh V. Chawla, SMOTE For learning from imbalanced data: Progress and challenges, marking the 15-year anniversary, J. Artificial Intelligence Res. 61 (2018) 863–905.

[7] Hui Han, Wen-Yuan Wang, Bing-Huan Mao, Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning, in: International Conference on Intelligent Computing, Springer, Berlin, Heidelberg, 2005, pp. 878–887.

[8] Haibo He, Yang Bai, Edwardo A. Garcia, Shutao Li, ADASYN: Adaptive synthetic sampling approach for imbalanced learning, in: Neural Networks, 2008 IJCNN 2008(IEEE World Congress on Computational Intelligence) IEEE International Joint Conference on, IEEE, 2008, pp. 1322–1328.

[9] Sukarna Barua, Md Monirul Islam, Xin Yao, Kazuyuki Murase, MWMOTE–Majority weighted minority oversampling technique for imbalanced data set learning, IEEE Trans. Knowl. Data Eng. 26 (2) (2014) 405–425.

[10] Miroslav Kubat, Stan Matwin, Addressing the curse of imbalanced training sets: one-sided selection, in: ICML, vol. 97, 1997, pp. 179–186.

[11] García Salvador, Francisco Herrera, Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy, Evol. Comput. 17 (3) (2009) 275–306.

[12] Pengyi Yang, Paul D. Yoo, Juanita Fernando, Bing B. Zhou, Zili Zhang, Albert Y. Zomaya, Sample subset optimization techniques for imbalanced and ensemble learning problems in bioinformatics applications, IEEE Trans. Cybern. 44 (3) (2014) 445–455.

[13] Russell Eberhart, James Kennedy, A new optimizer using particle swarm theory, in: Micro Machine and Human Science, 1995 MHS'95 Proceedings of the Sixth International Symposium on, IEEE, 1995, pp. 39–43.

[14] Ibrahim Aljarah, Majdi Mafarja, Ali Asghar Heidari, Hossam Faris, Yong Zhang, Seyedali Mirjalili, Asynchronous accelerating multi-leader salp chains for feature selection, Appl. Soft Comput. 71 (2018) 964–979.

[15] Hossam Faris, Majdi M. Mafarja, Ali Asghar Heidari, Ibrahim Aljarah, Al-Zoubi Ala'M, Seyedali Mirjalili, Hamido Fujita, An efficient binary salp swarm algorithm with crossover scheme for feature selection problems, Knowl.-Based Syst. 154 (2018) 43–67.

[16] Majdi Mafarja, Ibrahim Aljarah, Ali Asghar Heidari, Abdelaziz I. Hammouri, Hossam Faris, Al-Zoubi Ala'M, Seyedali Mirjalili, Evolutionary population dynamics and grasshopper optimization approaches for feature selection problems, Knowl.-Based Syst. 145 (2018) 25–45.

[17] Hossam Faris, Al-Zoubi Ala'M, Ali Asghar Heidari, Ibrahim Aljarah, Majdi Mafarja, Mohammad A. Hassonah, Hamido Fujita, An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks, Inf. Fusion 48 (2019) 67–83.

[18] Wentao Mao, Jinwan Wang, Zhanao Xue, 8, An ELM-based model with sparse-weighting strategy for sequential data imbalance problem, Int. J. Mach. Learn. Cybern. (4) (2017) 1333–1345.

[19] Saeed Piri, Dursun Delen, Tieming Liu, A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets, Decis. Support Syst. (2017).

[20] Qiao Cai, Haibo He, Hong Man, Imbalanced evolving self-organizing learning, Neurocomputing 133 (2014) 258–270.

[21] Gary M. Weiss, Kate McCarthy, Bibi Zabar, Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?, DMIN 7 (2007) 35–41.

[22] Kate McCarthy, Bibi Zabar, Gary Weiss, Does cost-sensitive learning beat sampling for classifying rare classes?, in: Proceedings of the 1st International Workshop on Utility-Based Data Mining, ACM, 2005, pp. 69–77.

[23] Andrew Estabrooks, Taeho Jo, Nathalie Japkowicz, A multiple resampling method for learning from imbalanced data sets, Comput. Intell. 20 (1) (2004) 18–36.

[24] Ludmila I. Kuncheva, Álvar Arnaiz-González, José-Francisco Díez-Pastor, Iain A.D. Gunn, Instance selection improves geometric mean accuracy: A study on imbalanced data classification, 2018, arXiv preprint arXiv:1804.07155.

[25] Syed Tanveer Jishan, Raisul Islam Rashu, Naheena Haque, Rashedur M. Rahman, Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique, Decis. Anal. 2 (1) (2015) 1.

[26] Nitesh V. Chawla, Nathalie Japkowicz, Aleksander Kotcz, Special issue on learning from imbalanced data sets, ACM Sigkdd Explorations Newslett. 6 (1) (2004) 1–6.

[27] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, Gong Bing, Learning from class-imbalanced data: Review of methods and applications, Expert Syst. Appl. 73 (2017) 220–239.

[28] Qi Dong, Shaogang Gong, Xiatian Zhu, Imbalanced deep learning by minority class incremental rectification, IEEE Trans. Pattern Anal. Mach. Intell. (2018).

[29] Xu-Ying Liu, Jianxin Wu, Zhi-Hua Zhou, Exploratory undersampling for class-imbalance learning, IEEE Trans. Syst. Man Cybern. B 39 (2) (2009) 539–550.

[30] Guillaume Lemaître, Fernando Nogueira, Christos K. Aridas, Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning, J. Mach. Learn. Res. 18 (17) (2017) 1–5.

[31] Ronaldo C. Prati, Gustavo EAPA Batista, Maria Carolina Monard, Data mining with imbalanced class distributions: concepts and methods, in: IICAI, 2009, pp. 359–376.

[32] N. Lunardon, G. Menardi, N. Torelli, R Package ROSE: Random over-Sampling Examples (Version 0.0-3), Università di Trieste and Università di Padova, Italia, 2013, URL http://cran.r-project.org/web/packages/ROSE/index.html.

[33] M. Kuhn, Caret: Classification and Regression Training, 2014, URL http://CRAN.R-project.org/ package=caret. R package version 6.0-22. Contributions from Jed Wing and Steve Weston and Andre Williams and Chris Keefer and Allan Engelhardt and Tony Cooper and Zachary Mayer and the R Core Team.

[34] S. Jahangeer Sidiq, Majid Zaman, Muheet Butt, A framework for class imbalance problem using hybrid sampling, Artif. Intell. Syst. Mach. Learn. 10 (4) (2018) 83–89.

[35] Qi Wang, ZhiHao Luo, JinCai Huang, YangHe Feng, Zhong Liu, A novel ensemble method for imbalanced data learning: bagging of extrapolation-SMOTE SVM, Comput. Intell. Neurosci. 2017 (2017).

[36] Soroush Saryazdi, Bahareh Nikpour, Hossein Nezamabadi-pour, NPC: Neighbors progressive competition algorithm for classification of imbalanced data sets, 2017, arXiv preprint arXiv:1711.10934.

[37] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, Amri Napolitano, RUSboost: A hybrid approach to alleviating class imbalance, IEEE Trans. Syst. Man Cybern. A 40 (1) (2010) 185–197.

[38] Nitesh V. Chawla, Aleksandar Lazarevic, Lawrence O. Hall, Kevin W. Bowyer, Smoteboost: Improving prediction of the minority class in boosting, in: European Conference on Principles of Data Mining and Knowledge Discovery, Springer, Berlin, Heidelberg, 2003, pp. 107–119.

[39] Enislay Ramentol, Yailé Caballero, Rafael Bello, Francisco Herrera, SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory, Knowl. Inf. Syst. 33 (2) (2012) 245–265.

[40] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, Hybrid sampling for imbalanced data, Integr. Comput.-Aided Eng. 16 (3) (2009) 193–210.

[41] Julián Luengo, Alberto Fernández, Salvador García, Francisco Herrera, Addressing data complexity for imbalanced data sets: analysis of SMOTE-based oversampling and evolutionary undersampling, Soft Comput. 15 (10) (2011) 1909–1936.

[42] T. Elhassan, M. Aljurf, F. Al-Mohanna, M. Shoukri, Classification of imbalance data using tomek link (T-link) combined with random undersampling (RUS) as a data reduction method, J. Inf. Data Min. 1 (2) (2016).

[43] Mohamed S. Kraiem, María N. Moreno, Effectiveness of basic and advanced sampling strategies on the classification of imbalanced data. A comparative study using classical and novel metrics, in: International Conference on Hybrid Artificial Intelligence Systems, Springer, Cham, 2017, pp. 233–245.

[44] Bee Wah Yap, Khatijahhusna Abd Rani, Hezlin Aryani Abd Rahman, Simon Fong, Zuraida Khairudin, Nik Nik Abdullah, An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets, in: Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013), Springer, Singapore, 2014, pp. 13–22.

[45] Arthur Asuncion, David Newman, UCI machine learning repository, 2007.