Naseriparsa *et al. Health Inf Sci Syst (2020) 8:22*
https://doi.org/10.1007/s13755-020-00112-w

Health Information Science and Systems

RESEARCH

RSMOTE: improving classification performance over imbalanced medical datasets

Mehdi Naseriparsa[1*], Ahmed Al-Shammari[1,3], Ming Sheng[2], Yong Zhang[2] and Rui Zhou[1]

Abstract

Introduction: Medical diagnosis is a crucial step for patient treatment. However, diagnosis is prone to bias due to imbalanced datasets. To overcome the imbalanced dataset problem, simple minority oversampling technique (SMOTE) was proposed that can generate new synthetic samples at data level to create the balance between minority and majority classes. However, the synthetic samples are generated on a random basis which causes class mixture problem; thus, resulting in deteriorating the classification performance and biased diagnosis.

Purpose: In order to overcome the SMOTE shortcomings, some modified methods were proposed that try to generate synthetic samples along the line segment of selected minority samples. Most of these methods adopt one of the two policies for selecting minority samples to generate synthetic samples: borderline region sampling or safe region sampling. However, they both suffer from over-generalisation problem. We propose a modified SMOTE-based resampling method called RSMOTE to alleviate the medical imbalanced dataset problem. We provide an in-depth analysis and verify the performance of RSMOTE over imbalanced medical datasets.

Methods: In this paper, the proposed RSMOTE divides the minority sample domain into four regions (normal, seminormal, semi-critical, and critical) based on the minority sample density analysis. RSMOTE discovers the minority sample region globally and applies the resampling near a specific group of samples.

Results: Our analysis and experiments verify that if synthetic samples are generated in the regions with high minority sample density, classification performance will be improved due to low risk of class mixture. Unlike some safe region methods, RSMOTE decides the region of minority samples on a global basis, thus removing the over-generalisation problem. Classic and additional evaluation metrics are considered to measure the effectiveness of the modified method: Recall, FP Rate, Precision, F-Measure, ROC area, and Average Aggregated Metric. We carried out experiments over various imbalanced medical datasets.

Conclusion: Based on the minority sample density analysis, we propose RSMOTE method that divides the minority sample domain into four regions. The proposed RSMOTE includes four re-sampling methods that each of them carries out resampling on a specific region. According to the experimental results, resampling on the regions with high minority sample density obtained better results while those with lower minority sample density got the inferior results. Thus, we conclude that the RSMOTE is a more flexible resampling method for the imbalanced medical datasets that is capable of generating samples with various minority sample densities.

Keywords: Medical diagnosis, Imbalanced learning, SMOTE, Class mixture, Classification performance

Introduction

Many medical diagnosis systems employ machine learning classification algorithms [1, 2], in which the classification algorithms are designed to work with balanced datasets, but often the real world datasets are imbalanced [3]. These datasets may contain many samples belonging to a majority class and a small

*Correspondence: mnaseriparsa@swin.edu.au
[1] Swinburne University of Technology, Hawthorn, Australia
Full list of author information is available at the end of the article

percentage to a minority class while the main goal of medical diagnosis systems is to discover those minority samples that are often referred as abnormal or interesting samples [4, 5], e.g., predicting breast cancer cases. In such circumstances, medical diagnosis systems will probably fail to discover the interesting samples correctly because they have bias over the majority class samples, e.g., non-cancerous cases. Due to the creation of bias over the majority class, models which are made on imbalanced medical datasets are not reliable, but show high accuracy. The high accuracy obtained is misleading because such models are capable of identifying the majority class samples which lead to improve the overall accuracy while they are unable to discover the minority class samples correctly which are usually the critical cases, e.g., the cancerous cases. The problem of finding critical samples among a large number of majority class samples occurred in real world problems such as the detection of malignant tumor from the benign samples [6].

A dataset is considered imbalanced if there is an unequal distribution of instances in its classes. Usually in real world medical datasets, the ratio of the number of minority class samples to the majority class samples is considerably small. For example, in UCI Hepatitis dataset, there are 155 instances. However, only 32 out of 155 samples belong to the minority class. The main problem with the class imbalances is that typical learners are biased towards the majority class and often the data distribution is not taken into consideration. As a result, samples from the majority class are correctly classified while the minority class samples are usually mis-classified which is called class mixture problem. Different methods [6–8] were proposed to solve the problem at the data and algorithmic level [9]. In this paper, we focus on the methods at the data level. These methods are based on the modification of datasets in order to provide a balanced distribution. Resampling is a solution to this problem at the data level along this line [6]. In resampling, a number of minority class samples are selected randomly and are duplicated in the original dataset. In this case, the number of minority samples increases and the class distribution gets balanced.

Contributions

The main contributions of this paper are summarised as follows:

- We propose a modified resampling technique called Region-based SMOTE (RSMOTE) for imbalanced medical datasets to improve the classification performance over imbalanced medical datasets.
- We provide an in-depth analysis for the proposed RSMOTE performance under various circumstances and compare it with other SMOTE versions.
- We carry out extensive experiments on imbalanced medical datasets and verify the effectiveness of the proposed method for medical diagnosis systems.

The remainder of this paper is organised as follows: Sect. 2 presents the related work. Section 3 presents the problem definition, which is followed by the proposed framework in Sect. 4. The experimental results are presented in Sect. 5. Finally, the paper is concluded in Sect. 6.

Related work

This section discusses the findings and drawbacks of the related studies on classification and prediction of medical data. The self-care diagnosis and classification are challenging in most health care systems. The self-care classification problem is computationally expensive and requires an expert system to reduce the cost and time [10, 11]. The majority of the existing works focus on building expert systems that are based on artificial intelligence methods. Zarchi et al. [12] introduced a novel standard dataset called SCADI (Self-Care Activities Dataset based on ICF-CY), where ICF-CY is a conceptual framework which was released by the World Health Organisation (WHO) [13]. The authors proposed two main types of expert systems for the self-care classification problem of children with physical and mental disabilities. Firstly, an Artificial Neural Network (*ANN*) is used as a classifier, which is trained by using SCADI. Then, a decision tree algorithm is employed to extract the self-care classification problem. The results show that the ANN-based system has achieved high accuracy in self-care classification in comparison with other classification methods.

Similarly, Lynch et al. [14] proposed modified classification models to diagnose the patients who have lung cancer in terms of survival. A number of supervised machine learning techniques including linear regression, Gradient Boosting Machines (GBM), Decision Trees (DTs), and Support Vector Machines (SVM) are used to efficiently classify the patient cases and make a good prediction for each case. In these techniques, some data attributes are selected such as tumor size, tumor grade, gender, stage, age, and a number of primaries. The experimental results show that the predicted values of the survival match the actual values of survival.

Most recently, Sanchez et al. [15] made a comprehensive comparison among various selection methods for clinical predictive modelling. Technically, two variable selection methods are implemented: regression-based method and (b) tree-based method. The regression-based method assumes a linear relationship between the given variables and the outcome, and the tree-based method is known to excel well in highly imbalanced datasets. The results on real medical datasets verify the efficiency of the classic regression-based variable selection methods in smaller datasets with < 20 events per variable, whereas the modern tree-based methods have achieved better performance in larger datasets with > 300 events per variable. However, the existing classification methods [14–17] have shown a serious limitation. None of these methods takes the imbalanced medical datasets problem into consideration. For instance, the classification model may be biased in favour of the majority class. Therefore, the goal of this paper is to address the medical imbalanced datasets problem.

Simple Minority Oversampling TEchnique (SMOTE) is a state of the art resampling technique that tries to balance the class distribution by adding new synthetic minority samples to the main dataset rather than duplicating the existing minority samples; however, it has various shortcomings. The main problem with SMOTE is that it creates synthetic samples without considering majority samples; thus, resulting in class mixture [6]. Other modified SMOTE versions were proposed like Borderline SMOTE [18] and Local Neighbourhood SMOTE [8] to tackle the setbacks existed in the original SMOTE but they have their own shortcomings. Borderline SMOTE creates synthetic samples in the borderline area; hence, it increases the risk of class mixture. In Local Neighbourhood SMOTE; new samples are created inside a safe region based on the local investigation of minority sample neighbours which may lead to the over-generalisation problem.

In some SMOTE's extensions, the authors combine the SMOTE technique with filtering methods [19–22]. The filtering scheme is proposed to overcome the noisy and borderline sample problem. In these methods, the noisy and problematic samples are detected by various methods, removed, and the synthetic samples are generated on the rest of the samples. This technique improves the classification performance since the synthetic samples are generated on the absence of noisy samples. For comparison purposes, we selected the iterative ensemble-based noise filter called Iterative-Partitioning Filter (IPF) [19], which is the most representative one. Recently, Cheng et al. [20] assumed data could be captured by Gaussian Mixture model, which may not be always true. Fahrudin et al. [21] used attribute weighting scheme which is likely to be application dependent. The work [22] is a preliminary version of the work [19].

In this paper, we propose a modified method called Region-based SMOTE (RSMOTE) that exploits minority sample density analysis. RSMOTE differentiates between minority samples based on their regions and generates the synthetic samples using the original minority samples of a region. We carry out a comprehensive investigation to verify the classification performance on resampled datasets in different regions. Moreover, we make an in-depth comparison among the proposed RSMOTE, Borderline SMOTE, and Local Neighbourhood SMOTE. In this work, we focus on the binary classification problem for imbalanced medical datasets which turns into a burning issue among the academia, industry researchers, and governmental agencies.

Problem statement

A dataset D is considered imbalanced if there is an unequal distribution between its classes. Class imbalance is primarily a two-class (binary) problem; while in some cases imbalance may also exist between various classes [6].

Definition 1 *Imbalanced Dataset*. A dataset D is imbalanced if the ratio of the minority S_{min} to the majority samples S_{maj} is less than a threshold δ, i.e., $\frac{|S_{min}|}{|S_{maj}|} < \delta$.

Usually classification algorithms are biased towards the majority class. This raises a critical issue with respect to the classification of the imbalanced datasets, specifically, in the medical domain. Thus, the classification results are not reliable because often data distribution is not taken into consideration. As a result, samples from the majority class are correctly classified while minority class samples are usually misclassified. Data and algorithmic level [9] methods were proposed to balance the class distribution. The data methods modify the imbalanced datasets to provide a balanced distribution which is referred to resampling techniques. In this paper, we utilize resampling methods to balance the class distribution of imbalanced medical datasets.

Definition 2 *Synthetic Sample*. Assume samples S belong to the dataset D, then a synthetic sample $s' \notin S$ is produced by using the line segment between two minority samples $s_1, s_2 \in S_{min}$.

In resampling, a number of minority class samples are selected randomly and are duplicated in the original dataset. In this case, the number of minority samples increases and the class distribution gets balanced.

Definition 3 *Simple Minority Oversampling TEchnique (SMOTE)*. Given an imbalanced dataset D, we generate synthetic minority samples S'_{min} such that $\frac{|S_{min} \cup S'_{min}|}{|S_{maj}|} \geq \delta$.

After balancing the dataset *D*, the classifier model would be more reliable because it is not biased against the minority class. The existing SMOTE versions suffer from drawbacks such as over-generalization and ignoring the global density analysis of the minority samples. Thus, there is a need to globally analyze the density of the minority samples before generating the synthetic samples to balance the datasets.

Proposed method

SMOTE was proposed by Chawla et al. [4]. This method generates synthetic samples for minority class to balance the class distribution of the dataset. SMOTE uses Euclidean distance of two minority samples and multiplies this distance to a random number between 0 and 1. Hence, the new sample is generated along the line segment of the two original samples. We upgrade SMOTE to consider the minority sample density for generating the synthetic samples which we call it *RSMOTE (Region-based Simple Minority Oversampling TEchnique)*. Our proposed RSMOTE diagnosis system is presented in Fig. 1. From this figure, RSMOTE works on imbalanced medical dataset and produces synthetic samples based on region analysis. To balance the medical dataset, the system adds the synthetic samples to the original dataset. Then, the diagnosis system analyses the balanced dataset to make a decision.

SMOTE is a powerful method that generates new samples between several existing minority class samples and has shown acceptable performance in various applications [7]. However, SMOTE suffers from serious shortcomings that make it unreliable for medical diagnosis systems over imbalanced datasets.

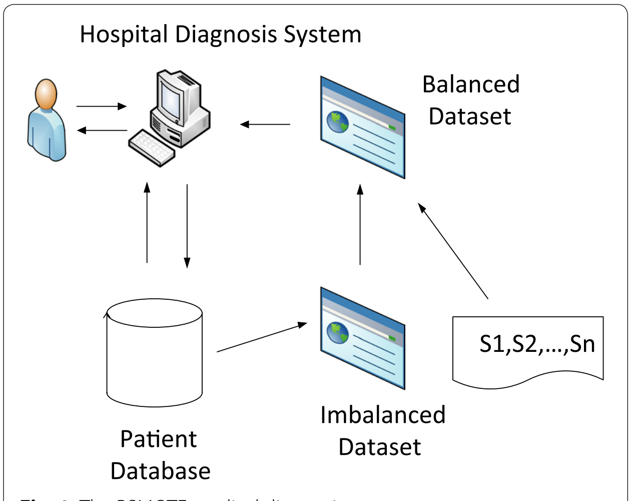


Fig. 1 The RSMOTE medical diagnosis system

SMOTE's shortcomings

SMOTE is inherently dangerous since it blindly generalises the minority area without considering the majority class. This strategy is particularly problematic in the case of highly skewed class distribution since, in such cases, the minority class is very sparse with respect to the majority class; thus, resulting in a higher chance of class mixture. To avoid the risk of class mixture and achieve better performance, several modified SMOTE methods were proposed. Borderline SMOTE which is referred to BL hereafter, uses borderline samples to create synthetic samples. Based on an analysis that considers classification algorithms attempt to learn the borderline of each class as exactly as possible, it selects the borderline and nearby samples and creates synthetic samples based on those samples. According to BL, the samples located far from the borderline may contribute little to classification while the samples in the borderline and nearby are the most important for classifiers. BL considers the samples at the decision border between classes and those located nearby are the most important for classifiers and the samples located inside a class region are easier to be classified. Therefore, the density of samples is increased in the borderline region due to creation of synthetic samples. This trend may have negative impact on the prediction of minority samples inside the minority class region due to lack of concentration on the inside region samples and changing the density in favour of decision region samples. Furthermore, creation of synthetic samples at the borderline region could increase the risk of class mixture because there are more majority class samples near to the minority samples in this region; thus, resulting in reduction of majority class prediction.

Another proposed method to solve the class mixture problem is Safe Level SMOTE [23]. In this method a special coefficient called safe level is calculated which denotes the level of risk for class mixture before generating the synthetic samples. In fact, safe level is defined as the number of minority class samples among *k* nearest neighbours for each minority class sample. Considering a minority sample, if safe level value equals to 0, it is considered as a noise. On the other hand, if safe level value is closer to *k* then the sample is located in a safe region of the minority class. The proposed method aims to generate synthetic samples closer to the safe region. Local neighbourhood SMOTE which is referred to LN hereafter, uses safe level definition to decide the risk level of a minority class sample for creating synthetic samples among its *k* nearest neighbours. LN uses safe level ratio to decide whether a sample is in safe region or not. The problem with LN is that the method calculates the safe level on a local basis, and therefore may result in overgeneralisation in the resampling process.

Region-based SMOTE

To tackle the drawbacks of the existing SMOTE variations, we should selectively consider the minority samples for the generation of synthetic samples. In this paper, we propose Region-based SMOTE (RSMOTE) to solve the existing problem. RSMOTE discovers minority sample region and then adopts an appropriate policy for synthetic samples generation. RSMOTE differentiates minority samples based on their regions and generates the synthetic samples using the original minority samples of a specific region. This step is helpful to tackle the over-generalisation problem which exists in SMOTE. To define the regions of the minority samples in RSMOTE, we carry out minority sample analysis.

Technically, RSMOTE divides the sample domain into four main categories: normal region (safe region), semi-normal region, semi-critical region, and critical region (risky region). This categorisation is based on the minority sample density, i.e., normal region is the most dense region and critical region is the least dense region. The risk of class mixture increases from normal to critical region in the resampling process. For instance, the risk of class mixture for semi-normal region is higher compared to normal region while is smaller compared to semi-critical region. Our goal is to focus on studying the resampling effect on the classification performance on various minority class subdomains and to discover the characteristics of each subdomain during the resampling process. By default, RSMOTE uses half of the number of minority class samples to decide the region of a minority sample. This parameter could be changed depending on the distribution of the minority samples.

To be specific, RSMOTE uses a density threshold ϵ to define the minority region. RSMOTE is not restricted to the following categorisation and any other categorisation may be applied easily by setting ϵ to different values. Let $N(s, r, D)$ return the number of minority neighbours of a minority sample s within the range r with respect to the dataset D.

- If $N(s,r,D) = 0$, then the sample is defined as Noise.
- If $0 < N(s,r,D) \leq \frac{|S_{min}|}{3}$, then the sample is located in critical region.
- If $\frac{|S_{min}|}{3} < N(s,r,D) \leq \frac{|S_{min}|}{2}$, then the sample is located in semi-critical region.
- If $\frac{|S_{min}|}{2} < N(s,r,D) \leq \frac{2\times|S_{min}|}{3}$, then the sample is located in semi-normal region.
- If $\frac{2\times|S_{min}|}{3} < N(s,r,D) \leq |S_{min}|$, then the sample is located in normal region.

As shown in Formula 1, minority sample domain consists of four subdomains. The subdomains are normal domain (ND), semi-normal domain (SND), semi-critical domain (SCD), and critical domain (CD).

$$S_{min} = ND \cup SND \cup SCD \cup CD \tag{1}$$

RSMOTE consists of four methods that each focuses on resampling over a specific region. These methods are as follows: (1) normal region SMOTE (NR), (2) semi-normal region SMOTE (SN), (3) semi-critical region SMOTE (SC), and (4) critical region SMOTE (CR).

```
Algorithm 1: RSMOTE Procedure
  Input: Dataset $D$, Threshold $\epsilon$, Balanced Ratio $\delta$,
         Default Range $r$
  Output: Balanced Dataset $D'$
1 $D' \leftarrow D$;
2 while $s = getNext(D.S_{min}) \neq \emptyset$ do
3     if $N(s,r,D) \geq \epsilon$ then
4         foreach $s_1, s_2 \in N$ do
5             $s' \leftarrow generateSample(s_1, s_2, D)$;
6             $D'.S \leftarrow D'.S \cup s'$;
7     if $|D'|.\frac{|S_{min}|}{|S_{maj}|} \geq \delta$ then
8         break;
9 return $D'$;
```

RSMOTE algorithm

Algorithm 1 presents the procedure of RSMOTE. The threshold ϵ determines the type of RSMOTE procedure since it permits the sample generation on a specific region. In line 1, we initialise the balanced dataset D'. In line 2, we select a minority sample s to analyse the region. If the number of minority samples within the nearest neighbours is bigger than the threshold ϵ , i.e., $N(s, D) \geq \epsilon$, then we are permitted to generate synthetic samples as pseudocoded in line 3. Thus, we generate the new synthetic sample s' alongside of the two minority samples s_1 and s_2 and add it to the balanced dataset D' as pseudocoded in lines 4–6. In line 7, we check whether the dataset D' is balanced, i.e., $|D'|.\frac{|S_{min}|}{|S_{maj}|} \geq \delta$. If it is balanced, we stop generating more samples in line 8; otherwise, we continue to generate more samples.

Analysis and discussion

As shown in Formula 2, $|S|$ is the total number of samples from both minority and majority classes.

$$|S| = |S_{min}| + |S_{maj}| \tag{2}$$

The class mixture is defined as the occurrence of incorrectly classifying a majority sample as minority or critical sample. Assume we have α minority samples among k nearest neighbours and $|S|$ number of samples.

Lemma 1 *The class mixture probability in the original SMOTE is* $1 - \frac{\alpha}{k}$.

Proof

We compute the minority sample density as follows: $|S_{min}| = \frac{\alpha}{k} \times |S|$. *Then, we compute the majority sample density* $|S_{maj}| = |S| - |S_{min}| = \frac{(k-\alpha) \times |S|}{k}$. *Finally, the class mixture probability will be as follows:* $P = \frac{|S_{maj}|}{|S|} = \frac{(k-\alpha) \times |S|}{k \times |S|} = \frac{k-\alpha}{k} = 1 - \frac{\alpha}{k}$. □

When the number of α minority samples within k nearest neighbours increases, the class mixture probability decreases. This would improve the classification performance after SMOTE generates synthetic samples.

Lemma 2 *The maximum class mixture probability in NR, SN, SC and CR are 0.33, 0.5, 0.66, and 1.*

Proof

We have the maximum density of majority samples in the NR when there is minimum density of minority samples. Thus, $Min(S_{min}) = \frac{2 \times |S|}{3}$, $Max(S_{maj}) = |S| - Min(S_{min}) = \frac{|S|}{3}$. *The maximum class mixture occurs when there is maximum density in the region. Hence,* $Max(P) = \frac{Max(S_{maj})}{|S|} = \frac{0.33 \times |S|}{|S|} = 0.33$. *Similarly for SN, SC, and CR, we compute the maximum class mixture probability as 0.5, 0.66, and 1 respectively.* □

Clearly, NR achieves the best performance because it guarantees that the smallest class mixture probability occurs when we select the most highly-dense normal region samples for generating synthetic samples. However, there may be small number of such samples in the normal region. In most cases, we need to generate synthetic samples on other regions in addition to the normal region, i.e, semi-normal, semi-critical or critical to balance the class distribution.

Lemma 3 *The maximum class mixture probability in BL is* $0.5 \leq P \leq 1$.

Proof

Assume the portion of minority samples within the total samples is as follows: $|S_{min}| = \alpha \times |S|$. *Then, the majority samples density becomes* $|S_{maj}| = |S| - |S_{min}| = (1 - \alpha) \times |S|$. *Therefore, we compute the class mixture probability as follows:* $P = \frac{|S_{maj}|}{|S|} = \frac{(1-\alpha) \times |S|}{|S|} = (1 - \alpha)$. *Since the samples are generated in the border region in BL, we have the following relation:* $0 \leq \alpha \leq 0.5$. *Finally, the class mixture probability becomes* $0.5 \leq P \leq 1$. □

Experiments

This Section highlights the experimental results for evaluating the performance of RSMOTE. We also compare RSMOTE with other versions (BL and LN). In Sect. 5.1, we provide the various evaluation metrics for verification of RSMOTE performance. Then, we present a brief description for the experimented datasets in Sect. 5.2. Finally, we present the experimental results in Sect. 5.3. For computing the classification algorithm performance, we split the data into training and testing data. We assign 90% as the training and 10% for testing the classification performance. The experiments are implemented in *C#* and executed by a processor Intel, Core (i3)-3570 CPU 3.40 GHz.

Evaluation metrics

To evaluate the performance for the binary classification problems, we have employed standard metrics. These metrics are produced by computing 4 common parameters shown in the following confusion matrix. Table 1 illustrates this confusion matrix for a binary classification problem. The first column of the table shows the actual class labels of the samples, and the first row presents their predicted class labels. TP denotes the number of minority (Positive) class samples that are correctly classified while FP denotes the number of minority (Positive) class samples that are incorrectly classified. TN denotes the number of majority (Negative) class samples that are correctly classified and FN is the number of majority (Negative) class samples that are misclassified.

$$OverallAccuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (3)$$

To evaluate the performance of a learner for imbalanced datasets, overall accuracy (shown in Eq. 3) is not suitable because it is heavily affected by the large number of correctly classified majority samples; thus, it performs poorly to the minority class [24–26]. The fact is that in imbalanced datasets, there are much more majority class samples comparing to minority class and misclassified

Table 1 Confusion matrix for performance evaluation

	Predicted positive	Predicted negative
True positive	TP	FN
True negative	FP	TN

minority samples have no reliable impact on the high rate of correctly classified majority samples. Therefore, the result shows the high accuracy rate while this rate does not reflect the poor performance for minority class. To conclude, we need more reliable evaluation metrics to evaluate the performance of a learner for imbalanced datasets. From the confusion matrix shown in Table 1, we can derive other metrics based on these four metrics that are presented in Eqs. 4, 5, 6, and 7.

$$FPRate = \frac{FP}{TN + FP} \tag{4}$$

$$TPRate = Recall = \frac{TP}{TP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$F - Measure = \frac{(1 + \beta^2) \times Recall \times Precision}{(\beta^2 \times Recall + Precision)} \tag{7}$$

We present the following example to illustrate more. Suppose we are classifying the patient tumors into malignant and benign tumors. Assume we have 10 samples that 4 of them are malignant while 6 are benign. Also, suppose the algorithm classifies 4 samples as malignant in which 2 of them are correctly malignant, and 6 samples labeled as benign in which 4 of them are actually benign. Then, $TP = 2$, $FP = 2$, $TN = 4$, and $FN = 2$. By having the values of these basic parameters, we can compute the metrics as follows:

- $Accuracy = \frac{TP+TN}{TP+FN+FP+TN} = \frac{2+4}{2+2+4+2} = 0.6.$
- $FPRate = \frac{FP}{TN+FP} = \frac{2}{4+2} = 0.33.$
- $TPRate = \frac{TP}{TP+FN} = \frac{2}{2+2} = 0.5.$
- $Precision = \frac{TP}{TP+FP} = \frac{2}{2+2} = 0.5.$

To achieve a reliable classification performance for imbalanced datasets, we have to improve Recall (TP Rate) (shown in Eq. 5) and Precision (shown in Eq. 6) while decreasing the value of FP Rate (shown in Eq. 4). Therefore, we need to increase TP and TN values while decreasing FP and FN values simultaneously which is conflicting in some cases. To avoid the conflict between Precision and Recall, F-Measure metric is presented which indicates a single number reflecting the merit of a classifier in the presence of rare classes. In Formula 7, β corresponds to the relative importance of Precision versus Recall which is usually set to 1. To avoid the conflict of Recall (TP Rate) and FP Rate, AUC (Area under ROC) is presented. ROC curve is a two-dimensional graph in which TP Rate is plotted on the y-axis and FP Rate is plotted on the x-axis. The point (0, 1) is ideal point on graph for learners [27, 28].

Average aggregated metric

The last metric that we consider for performance evaluation is the Average Aggregated Metric which is derived from the combination of five metrics mentioned in Sect. 5.1 (FP Rate, Recall, Precision, F-Measure and AUC) and is referred to AAM in this paper hereafter. Although these metrics are not homogeneous, their combination could reflect the performance of resampling more fair. In fact, this metric reflects the resampling performance by a broad view and considers several factors simultaneously for a more accurate comparison. We introduce the AAM as a metric that reflects the model performance with respect to the various metrics. That is because some models perform well on a part of metrics while they do not perform well on all metrics at the same time. To reflect the overall performance for the classifier model, we proposed AAM which avoids the performance evaluation to be biased against particular metrics.

In Table 2, metrics and their characteristics for evaluating AAM are presented. In this paper, the weights for all metrics are set to 1. To evaluate this metric, four metrics (Recall, Precision, F-Measure and AUC) have positive sign while FP Rate has negative sign in the aggregation function. In fact, resampling performance is better if Recall, Precision, F-Measure and AUC have higher values while FP Rate has a lower value. That's why we penalise the FP rate weight for computing AAM since the FP Rate shows the misclassification rate of the model. In Eq. 8, Average Aggregated Metric (AAM) is presented.

$$AAM = \frac{\sum_{i=1}^{5} w_i \times M_i}{5} \tag{8}$$

Dataset

To test the resampling performance of RSMOTE and compare it with other SMOTE versions, we use some

Table 2 Metrics required for calculating AAM

Index	Metric	Weight	Sign
M_1	Recall	1	+
M_2	Precision	1	+
M_3	F-Measure	1	+
M_4	AUC	1	+
M_5	FP Rate	1	–

Table 3 SMOTE versions characteristics

Abbr	Method	Strategy
NR	Normal SMOTE	Region-based
SN	Semi-normal SMOTE	Region-based
SC	Semi-critical SMOTE	Region-based
CR	Critical SMOTE	Region-based
BL	Borderline SMOTE	Borderline
LN	Local neighbourhood SMOTE	Safe region
SM	SMOTE	Original
IPF	SMOTE-IPF	Iterative filtering

Table 4 Characteristics of UCI medical datasets

Datasets	Att	$\lvert S_{min}\rvert$	$\lvert S_{maj}\rvert$	Ratio
Diabetes	9	268	500	1.86:1
Heart	14	120	150	1.25:1
Hepatitis	20	32	123	3.84:1
WDBC	569	212	357	1.68:1

imbalanced medical UCI datasets [29] and apply various SMOTE methods (Table 3) on these datasets. We employ and modify the SMOTE-IPF [19] which is an iterative filtering-based resampling methods to compare the results in the experiments. Table 4 presents the datasets and their characteristics. We present and analyse the classification results after the corresponding SMOTE methods have added the synthetic samples to the original dataset in Sects. 5.3 and 5.4 . In Sect. 5.5, we vary the dataset imbalanced ratio to examine the resampling performance by increasing the number of synthetic samples in different steps.

Evaluation and analysis

Table 5 presents the performance evaluation results on Diabetes dataset for various methods. There is no normal region on Diabetes dataset; thus, NR is not applicable. The best performance belongs to SN for all metrics presented in the Table. Considering Recall, SN is the best with the value 0.79 and IPF with 0.75 is in the second place while LN with 0.7 places the third. In Precision, SN with 0.82 goes first while IPF with 0.78 and LN with 0.76 are in the second and third places respectively. SN has the smallest value of 0.18 for FP Rate which is the best and IPF with 0.21 and LN with 0.22 are in the second and third places. For F-Measure, SN with 0.8 places first and IPF with 0.75 is in the second place and LN with 0.73 is in the third place. In AUC metric, SN is the best with 0.88 while IPF and LN with values 0.84 and 0.83 are in the second and third places respectively.

Table 6 presents the evaluation results on Heart dataset for different SMOTE versions. There is no normal region on Heart dataset; thus, NR is not applicable. SN outperforms other versions considering three metrics (FP Rate with 0.11, F-Measure with 0.86 and AUC with 0.91). In Recall metric, the best performance belongs to LN with 0.84. Considering Recall, SN and IPF place the second

Table 5 Results on diabetes dataset

Metric	SN	SC	CR	BL	LN	SM	IPF
M_1	**0.79**	0.70	0.64	0.66	0.7	0.69	0.75
M_2	**0.82**	0.73	0.7	0.72	0.76	0.75	0.78
M_3	**0.80**	0.72	0.67	0.69	0.73	0.72	0.75
M_4	**0.87**	0.81	0.76	0.79	0.83	0.82	0.84
M_5	**0.18**	0.25	0.27	0.25	0.22	0.23	0.21

The bold shows the best performance among various methods

Table 6 Results on heart dataset

Metric	SN	SC	CR	BL	LN	SM	IPF
M_1	0.83	0.82	0.81	0.82	**0.84**	0.8	0.83
M_2	0.88	0.84	0.84	0.84	0.85	0.85	**0.89**
M_3	**0.85**	0.83	0.83	0.83	0.84	0.82	0.84
M_4	**0.92**	0.9	0.89	0.88	0.9	0.89	0.91
M_5	**0.11**	0.15	0.14	0.14	0.14	0.14	0.13

The bold shows the best performance among various methods with 0.83. In Precision metric, IPF is the best with 0.89, SN is in the second with 0.88, LN and SM are in the third place with 0.85. For FP Rate, IPF is in the second place with 0.13. In F-Measure, LN and IPF with 0.84 are in the second place. Finally in AUC metric, IPF with 0.91 has the second place.

In Table 7, performance metrics are evaluated on Hepatitis dataset and the results are shown for different SMOTE versions. There is no normal region on Hepatitis dataset; thus, NR is not applicable. SN outperforms other versions considering three metrics (Precision with 0.93, FP Rate with 0.06 and AUC with 0.943). In Recall metric, the best performance belongs to IPF with 0.88 while LN is in the second place with 0.87, CR is in the third place. Considering Precision, SC and BL with 0.91 are in the second place. In FP Rate metric, both SN and SC with 0.06 are in the first place, and IPF with BL are in the second place. For F-Measure, BL with 0.886 places first while SN with 0.885 is in the second and LN with 0.884 is in the third place. Finally in AUC metric, both SN and SC with 0.943 are the best and then IPF with 0.941 is in the second place and LN with 0.94 is in the third place.

In Table 8, performance metrics are evaluated on WDBC dataset and the results are shown for different SMOTE versions. NR outperforms other versions considering all metrics (Recall with 0.93, Precision with 0.968, FP Rate with 0.03, F-Measure with 0.95 and AUC with 0.99). Considering Recall, SN and IPF place the second with 0.91 and then LN is in the third place with 0.9. In Precision metric, LN and IPF are in the second place with 0.96 and SM with 0.95 is in the third place. For FP Rate, LN with 0.039 places the second while SM and IPF with 0.04 are in the third place. In F-Measure, both NR and IPF place the first while LN with 0.935 is in the second place. Finally in AUC metric, IPF with 0.983 has the second place.

There is no normal region on 3 UCI datasets (Diabetes, Heart, and Hepatitis) due to the data distribution. Therefore, NR is not applicable for these datasets. In WDBC however, we find normal region; thus, NR is applicable and it outperforms other versions with a good margin on all metrics. That is because NR effectively discovers the most highly-dense region and utilises its corresponding samples to generate the synthetic samples. Thus, the synthetic minority samples strengthen the learners model for classification which improves the performance. After NR (which is not applicable in many cases due to its tight conditions), SN outperforms other SMOTE versions considering most metrics since it utilises the highly-dense minority samples region. This leads to better classification results because it avoids the learners from building a model with class mixture problem. After SN, we observe that IPF achieves the best performance due to the fact that it iteratively filters the noise samples; then it applies the resampling technique.

Average aggregated metric evaluation

Figure 2 evaluates the AAM parameter on the experimented UCI datasets and presents the results for different SMOTE versions (SN, SC, CR, LN, BL, SM, and IPF). From the figure on Hepatitis dataset, AAM parameter for SN equals to 0.71 which is the highest compared to

Table 7 Results on hepatitis dataset

Metric	SN	SC	CR	BL	LN	SM	IPF
M_1	0.84	0.83	0.85	0.85	0.87	0.81	**0.88**
M_2	**0.93**	0.91	0.9	0.91	0.89	0.89	0.9
M_3	0.885	0.87	0.88	**0.886**	0.884	0.85	0.88
M_4	**0.943**	**0.943**	0.93	0.93	0.94	0.92	0.941
M_5	**0.06**	**0.06**	0.08	0.07	0.09	0.09	0.07

The bold shows the best performance among various methods

Table 8 Results on WDBC dataset

M	NR	SN	SC	CR	BL	LN	SM	IPF
M_1	**0.93**	0.91	0.89	0.84	0.89	0.9	0.89	0.91
M_1	**0.968**	0.94	0.93	0.93	0.94	0.96	0.95	0.96
M_1	**0.95**	0.93	0.91	0.89	0.91	0.935	0.92	**0.95**
M_1	**0.99**	0.98	0.97	0.96	0.97	0.98	0.98	0.983
M_1	**0.03**	0.05	0.06	0.06	0.06	0.039	0.04	0.04

The bold shows the best performance among various methods

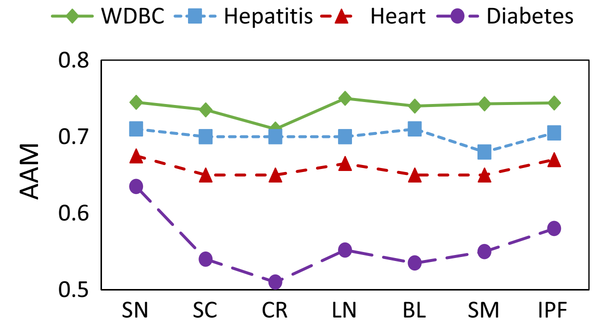


Fig. 2 Average aggregated metric (AAM) using similar weight (w=1) for all metrics

other methods. After that, BL with 0.705 and IPF with 0.704 are in the second and third places respectively. Considering Diabetes dataset, SN has the highest value for AAM parameter which is 0.6238 and then IPF with 0.58 is in the second place; then, LN with 0.5598 succeeds. In WDBC dataset, the best AAM belongs to NR with 0.7616 which is not shown in the figure. After that, SN with 0.745 and IPF with 0.744 have the highest value respectively. In Heart dataset, SN with 0.674 is the best; then, IPF with 0.67 is in the second place, and LN with 0.665 is in the third place.

Based on the results observed above and the findings about the class mixture probability of RSMOTE methods, the order of resampling performance for different methods is as follows: *NR* > *SN* > *SC* > *CR*.

In a broader comparison, LN [8] has a better performance comparing to SC while its performance is lower than IPF in most cases. Considering BL [18], its performance lies between SC and SM. Also the results show that SM has a better performance compared to CR while it has a lower performance compared to BL. Hence, a broader performance comparison for SMOTE versions performance is summarized below:

$$NR > SN > IPF > LN > SC > BL > SM > CR \tag{9}$$

Effect of imbalanced ratio

In this section, we use Cuff-Less Blood Pressure Estimation dataset [29] (BPE) as an imbalanced dataset to show the final performance of various SMOTE versions. This dataset contains a collection of blood pressure values for a group of patients. We classify some records as risky and the rest as normal due to the dataset attributes value. Since the majority of records are normal and the rest are risky (the minority class), this dataset is imbalanced. Thus, we choose it and apply different SMOTE resampling methods to balance the dataset in 5 iterations. In each iteration we generate 1000 synthetic samples and add them to the original dataset and compute the performance metrics. The characteristics of the BPE dataset is presented in Table 9.

In Fig. 3a, we evaluate the Recall parameter on BPE dataset and present the results in 5 steps. The number of synthetic samples is increased in each step by running the resampling methods (NR, LN, BL, SM, and IPF). From the figure, the best Recall values belong to NR in all steps and IPF is in the second place in the first two steps; then, LN gets the second place in the next three iterations. In NR, Recall goes above 0.95 in the second step and gets fixed in the last two steps with the value of 0.964. With respect to IPF, Recall deteriorates in iteration 3; then, it's fixed for the rest of the iterations. Considering LN, Recall is increased by relatively high rates in the second and third steps while in the last two steps this rate is lower and it reaches to 0.928 in the last step. About BL, Recall is higher compared to SM in the first three steps but in the last two steps SM outperforms BL. This shows that by increasing the number of synthetic samples in BPE dataset, SM performs better than BL on Recall metric, but still lower than NR and LN. NR is the best.

Figure 3b presents FP Rate values evaluated on BPE dataset. According to the results, IPF NR and LN have the best performance on this metric in the first iteration. FP Rate value starts from 0.011 and finishes with 0.01 for these methods. BL has a smaller FP Rate compared to SM in the first step but in the next steps SM has a better FP Rate values. Therefore like Recall, the FP Rate improves when the number of synthetic samples increases for SM while BL performance deteriorates.

In Fig. 3c, we evaluate the Precision on BPE dataset. The best Precision belongs to NR in all steps; then, IPF and LN gain the second place on the Precision metric. That is because NR and LN successfully generate the synthetic samples along the minority samples which assist the learners not to mix the minority with majority samples. Also, the filtering process in the IPF has improved the classifier precision because the noisy samples are removed. Considering NR, IPF, and LN, Precision is increased similarly but NR is a little better than IPF and LN. That is because unlike LN, NR applies a global

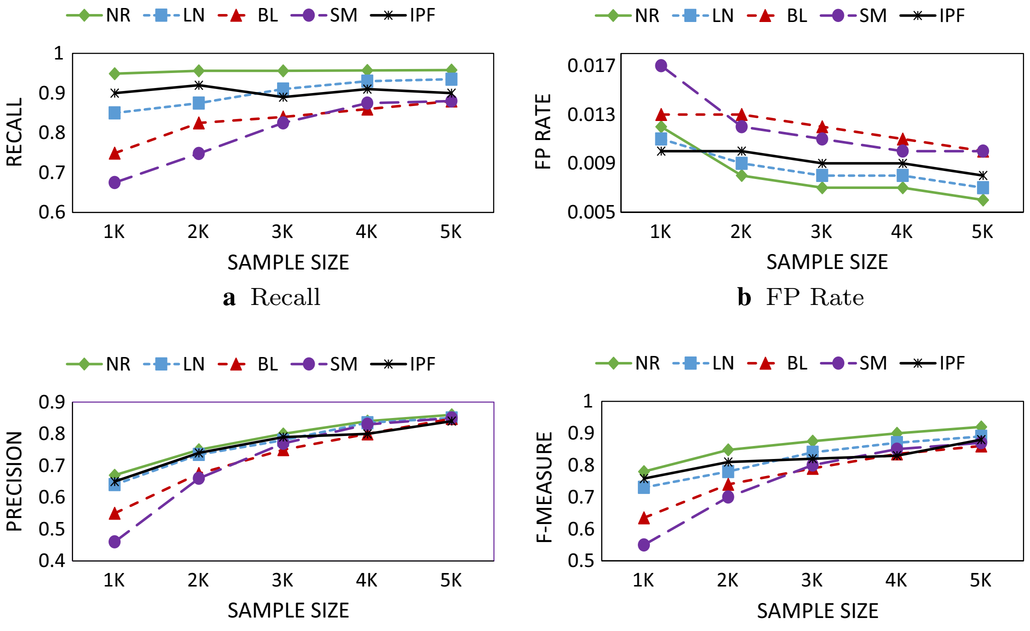
Table 9 Characteristics of BPE dataset

Datasets	Att	$\|S_{min}\|$	$\|S_{maj}\|$	Ratio
BPE	3	1126	97,000	86.15:1
BPE	3	2126	97,000	45.63:1
BPE	3	3126	97,000	31.03:1
BPE	3	4126	97,000	23.51:1
BPE	3	5126	97,000	18.92:1
BPE	3	6126	97,000	15.83:1

region analysis to discover the best minority samples for synthetic sample generation. Moreover, unlike IPF that filters the noisy samples and uses SMOTE without region analysis to generate the synthetic samples, NR thoroughly analyses the sample region globally and only generates the samples in the highly-dense region. This analysis maximises the classifiers performance. In the last step, Precision reaches its highest value to 0.858 for NR, 0.86 for LN, and 0.843 for IPF. For BL, the figure shows that BL has a better Precision compared to SM in the first and second steps. However, SM outperforms BL in the next steps. Also, the result obtained from Precision metric denotes that SM improves the performance better than BL when synthetic samples increase in the dataset. That is because BL generates the minority samples near to the majority region. This leads to increase the class mixture probability; specifically, when the synthetic samples are generated in larger numbers.

Figure 3d presents the outbound results for F-Measure metric on BPE dataset in 5 iterations. From the figure, the best F-Measure belongs to NR in all steps and LN achieves the second place. In NR, F-Measure value starts from 0.779 in the first step and reaches to 0.908 in the last step while in LN it starts from 0.726 and reaches to 0.887 in the last step. Similar to Precision, BL has a better F-Measure comparing to SM in the first two steps while SM is better in the last three steps which confirms this fact that SM performs better when synthetic samples are generated in a larger number.

In Fig. 4a, we evaluate the AUC metric on BPE dataset. The best AUC belongs to NR in all steps, then, IPF is in the second place in the first two steps. However, LN is in the second place in and after the third iterations. That is because both NR and LN utilise the minority samples in the highly-dense minority region to generate synthetic samples. Conversely, the IPF generates the synthetic samples after filtering noisy samples by using SMOTE. The filtering scheme works to improve the performance the first iterations in comparison with LN. However, when the number of synthetic samples increases, the performance is lower than LN since the samples are not generated near the highly-dense samples. Considering NR, AUC starts from 0.988 and reaches to 0.992 in the third step which remains fixed in this number until the last step. In LN, AUC starts from 0.987 and reaches to its highest value in the fourth step to 0.99 and it reduces to 0.989 in the last step. Regarding IPF, the AUC starts from 0.985 and reaches to 0.988 in the fourth step; then, it deteriorates to 0.986 in the last step. For BL, AUC is higher comparing to SM in the first two steps but in the last three steps SM outperforms BL and achieves better AUC. Similar to the results on previous metrics, AUC confirms this fact that SM performs better when synthetic samples are generated in a large number.



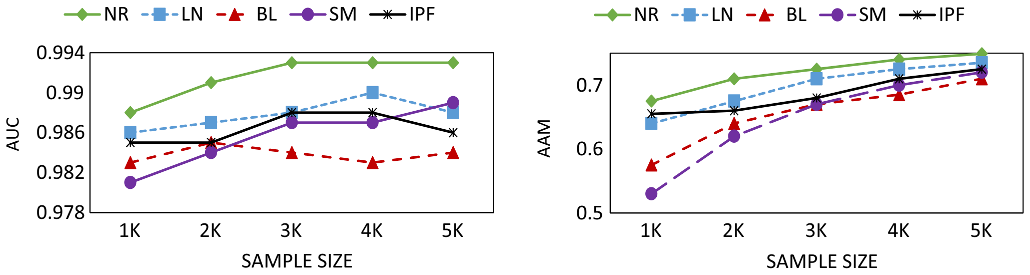
a Recall

b FP Rate

c Precision

d F-Measure

Fig. 3 The effect of imbalanced ratio using Recall, FPRate, Precision, and FMeasure metrics



a AUC

b AAM

Fig. 4 The effect of imbalanced ratio using AUC and AAM metrics

According to the results obtained from implementing different SMOTE methods on BPE dataset, we can make a better comparison between these methods as presented in formula 9. By increasing the number of synthetic samples, SM performance gets better comparing to BL; however, when the number of synthetic samples is small BL has better performance. In summary, NR proves to be the best method when we increase the number of synthetic samples since all these samples are generated in the highly-dense minority sample region.

Figure 4b evaluates the AAM metric on BPE dataset. As it is clear from the figure, NR obtains the best results and after that LN has the best AAM. In NR, AAM starts from 0.6728 and reaches to 0.7424 in the last step. In LN, AAM starts from 0.6368 and reaches to 0.7288 in the last step. With respect to IPF, the AAM is better than LN in the first step; then, LN gets better in the next rounds. Considering BL, AAM is higher comparing to SM in the first two steps while SM outperforms BL in the last three steps.

Conclusion

In this paper, we propose RSMOTE resampling for imbalanced medical datasets. RSMOTE utilizes Safe region resampling technique and provides a flexible minority sample density analysis to generate synthetic samples. The proposed RSMOTE method divides the minority sample domain into four regions, corresponding to four resampling methods (NR, SN, SC, and CR) and each of these methods carries out resampling on a specific region. The impact of resampling on possible regions is investigated thoroughly by evaluating the performance metrics on the corresponding RSMOTE methods. According to the results, resampling on the regions with high minority sample density obtained better results while those with lower minority sample density got the worst results which denote that generating the synthetic samples near the highly-dense minority samples is more effective compared to less-dense minority samples while it has its drawbacks. Safe region resampling techniques such as LN decide the safe level ratio based on local analysis; thus, resulting in over-generalisation. Conversely, our proposed RSMOTE decides the region of minority samples based on a global view of minority class distribution; thus, removing over-generalisation in its judgement.

Author details

[1] Swinburne University of Technology, Hawthorn, Australia. [2] Tsinghua University, Beijing, China. [3] University of Al-Qadisiyah, Al Diwaniyah, Iraq.

Received: 27 October 2019 Accepted: 20 May 2020
Published online: 12 June 2020

References

1. Paiva JS, Cardoso J, Pereira T. Supervised learning methods for pathological arterial pulse wave differentiation: a svm and neural networks approach. Int J Med Inform. 2018;109:30–8.
2. Srivastava SK, Singh SK, Suri JS. Healthcare text classification system and its performance evaluation: A source of better intelligence by characterizing healthcare text. J Med Syst. 2018;42(5):97.
3. Al-Shammari A, Liu C, Naseriparsa M, Vo BQ, Anwar T, Zhou R. A framework for clustering and dynamic maintenance of xml documents. In: International Conference on Advanced Data Mining and Applications, Springer, New York; 2017. pp. 399–412.
4. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
5. Al-Shammari A, Zhou R, Liu C, Naseriparsa M, Vo BQ. A framework for processing cumulative frequency queries over medical data streams. In: International Conference on Web Information Systems Engineering, Springer; 2018. pp. 121–131.
6. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263–84.
7. Batista GEAPA, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor. 2004;6(1):20–9.
8. Maciejewski T, Stefanowski J. Local neighbourhood extension of SMOTE for mining imbalanced data, In: Proceedings of the IEEE symposium on computational intelligence and data mining, CIDM 2011, April 11–15, 2011, Paris, France, 2011; pp. 104–111.
9. Chawla NV, Japkowicz N, Kotcz A. Editorial: special issue on learning from imbalanced data sets. SIGKDD Explor. 2004;6(1):1–6.
10. Yang S, Guo J-Z, Jin J-W. An improved id3 algorithm for medical data classification. Comput Electr Eng. 2018;65:474–87.

11. Al-Shammari A, Zhou R, Naseriparsaa M, Liu C. An effective density-based clustering and dynamic maintenance framework for evolving medical data streams. Int J Med Informatics. 2019;126:176–86.
12. Zarchi M, Bushehri SF, Dehghanizadeh M. Scadi: a standard dataset for self-care problems classification of children with physical and motor disability. Int J Med Informatics. 2018;114:81–7.
13. World Health Organisation, https://www.who.int/, Accessed 20 Nov 2018.
14. Lynch CM, Abdollahi B, Fuqua JD, Alexandra R, Bartholomai JA, Balgemann RN, van Berkel VH, Frieboes HB. Prediction of lung cancer patient survival via supervised machine learning classification techniques. Int J Med Inform. 2017;108:1–8.
15. Sanchez-Pinto LN, Venable LR, Fahrenbach J, Churpek MM. Comparison of variable selection methods for clinical predictive modeling. Int J Med Inform. 2018;116:10–7.
16. Araújo FH, Santana AM, Neto PAS. Using machine learning to support healthcare professionals in making preauthorisation decisions. Int J Med Inform. 2016;94:1–7.
17. Chebouba L, Boughaci D, Guziolowski C. Proteomics versus clinical data and stochastic local search based feature selection for acute myeloid leukemia patients' classification. J Med Syst. 2018;42(7):129.
18. Han H, Wang W, Mao B. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: Advances in Intelligent Computing, International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23–26, 2005, Proceedings, Part I, 2005; pp. 878–887.
19. Sáez JA, Luengo J, Stefanowski J, Herrera F. SMOTE-IPF: addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. Inf Sci. 2015;291:184–203.
20. Cheng K, Zhang C, Yu H, Yang X, Zou H, Gao S. Grouped SMOTE with noise filtering mechanism for classifying imbalanced data. IEEE Access. 2019;7:170668–81.
21. Fahrudin T, Buliali JL, Fatichah C. Enhancing the performance of smote algorithm by using attribute weighting scheme and new selective sampling method for imbalanced data set. Int J Innov Comput Inf Control. 2019;15:423–44.
22. Sáez JA, Luengo J, Stefanowski J, Herrera F. Managing borderline and noisy examples in imbalanced classification by combining SMOTE with ensemble filtering. In: Intelligent Data Engineering and Automated Learning - IDEAL 2014, 15th International Conference, Salamanca, Spain, September 10–12, 2014. Proceedings, Vol. 8669 of Lecture Notes in Computer Science, Springer; 2014. pp. 61–68.
23. Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. Safe-level-smote: safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem, in: Advances in Knowledge Discovery and Data Mining, 13th Pacific-Asia Conference, PAKDD 2009, Bangkok, Thailand, April 27–30, 2009, Proceedings; 2009. pp. 475–82.
24. Weiss GM. Mining with rarity: a unifying framework. SIGKDD Explor. 2004;6(1):7–19.
25. Chawla NV, Lazarevic A, Hall LO, Bowyer KW. Smoteboost: Improving prediction of the minority class in boosting. In: Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, Knowledge Discovery in Databases: PKDD 2003, Cavtat-Dubrovnik, Croatia, September 22–26, 2003, 2003; pp. 107–19.
26. Sun Y, Kamel MS, Wong AKC, Wang Y. Cost-sensitive boosting for classification of imbalanced data. Pattern Recogn. 2007;40(12):3358–78.
27. Chawla NV. Data mining for imbalanced datasets: an overview. In: Maimon O, Rokach L, editors. The data mining and knowledge discovery handbook. New York: Springer; 2005. p. 853–67.
28. Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett. 2006;27(8):861–74.
29. UCI machine learning repository. http://archive.ics.uci.edu/ml/, accessed 7 Feb 2018.