Hindawi
Discrete Dynamics in Nature and Society
Volume 2021, Article ID 6647557, 12 pages
https://doi.org/10.1155/2021/6647557

Research Article

Increasing Minority Recall Support Vector Machine Model for Imbalanced Data Classification

Chunye Wu, Nan Wang, and Yu Wang

School of Mathematical Science, Heilongjiang University, Harbin 150080, China

Correspondence should be addressed to Nan Wang; 2003137@hlju.edu.cn

Received 13 December 2020; Revised 14 April 2021; Accepted 21 April 2021; Published 5 May 2021

Academic Editor: Manuel De la Sen

Imbalanced data classification is gaining importance in data mining and machine learning. The minority class recall rate requires special treatment in fields such as medical diagnosis, information security, industry, and computer vision. This paper proposes a new strategy and algorithm based on a cost-sensitive support vector machine to improve the minority class recall rate to 1 because the misclassification of even a few samples can cause serious losses in some physical problems. In the proposed method, the modification employs a margin compensation to make the margin lopsided, enabling decision boundary drift. When the boundary reaches a certain position, the minority class samples will be more generalized to achieve the requirement of a recall rate of 1. In the experiments, the effects of different parameters on the performance of the algorithm were analyzed, and the optimal parameters for a recall rate of 1 were determined. The experimental results reveal that, for the imbalanced data classification problem, the traditional definite cost classification scheme and the models classified using the area under the receiver operating characteristic curve criterion rarely produce results such as a recall rate of 1. The new strategy can yield a minority recall of 1 for imbalanced data as the loss of the majority class is acceptable; moreover, it improves the g-means index. The proposed algorithm provides superior performance in minority recall compared to the conventional methods. The proposed method has important practical significance in credit card fraud, medical diagnosis, and other areas.

1. Introduction

Imbalanced data classification is gaining importance in data mining and machine learning [1–3]. An imbalance in a particular dataset occurs when the number of instances in one class (the minority class) is significantly smaller than that in the other class (the majority class). The minority class is generally of interest in data classification; hence, it is considered a positive (+) class, whereas the majority class is considered a negative (−) class [4, 5]. For instance, in the medical field, as there is more interest in the disease samples than in the health samples, the disease samples comprise the minority class. Traditional machine learning models mainly consider the accuracy of the two categories as equally important, i.e., they do not consider the accuracy of some categories to be more important than that of other categories, particularly when the sample size of the minority category is small [6, 7]. Although these models have been successfully applied to various balanced data classification problems, their performance decreases substantially when applied to imbalanced datasets [8–10] owing to the lack of sufficient training data in the positive class, which are required to perform an accurate classification of its instances. To address this issue, the positive class of a dataset has attracted increased attention [11], and traditional classifiers have undergone numerous improvements in order to be able to handle imbalanced data applications.

In certain applications, the tolerance to errors in the positive class is extremely low. For instance, in the case of industrial system fault diagnosis [12], the classifier must deal with an imbalanced dataset, i.e., the number of available healthy class instances outnumbers the faulty class ones. Accordingly, it is necessary to develop a classifier that accounts for this type of imbalanced data distribution and warns of all possible failures, even if there may be many false-warning occurrences. Credit card fraud detection is a

well-known classification problem [13]. In order to target a specific customer segment, banks use data mining algorithms to classify customers as buyers and nonbuyers. In this context, if a model correctly detects a potential customer for a campaign, there will be a particular profit related to gaining that customer; if a potential buyer is not identified, the profits that would be gained from him/her might be lost. However, if a potential nonbuyer is identified as a buyer, credit card fraud would occur, causing the bank potential massive losses. Similarly, a 100% recall model reduces some of the profits but does not risk credit card fraud occurrences. Furthermore, failing to diagnose a cancerous lesion is unacceptable and can have devastating effects on a patient, although this situation rarely occurs [14]. In general, the classifier is only used as an aid to manual diagnosis, which means that the classifier can only diagnose patients at risk of cancer and provide an artificial judgment. There are disastrous consequences if the classifier misses any patient who may have cancer. In this study, we developed a cost-sensitive support vector machine (SVM) model to increase the recall of a positive sample from an actual background to 100%. Based on this model, we proposed a medium algorithm strategy to increase the positive-class recall rate to 1. We introduced different penalty factors, namely, C^+ and C^-, for each of the positive and negative SVM slack variables during the training process and adjusted the classification boundary by altering the positive-class margin. This approach ensured that the positive-class samples would be more generalized to achieve the target recall rate of 1 when the boundary reached a certain position. For parameter selection, we employed the grid search approach and selected a model with a recall of 100% and a higher specificity for the negative class. The SVM model was adopted to address the data imbalance and exhibited an acceptable performance [15–21].

Our study makes a significant contribution to the field because we were able to confirm that (1) increasing the recall rate to 100% is a feasible classification indicator; (2) the decision boundary could be altered successfully by correcting the positive margin; and (3) the g-means increases when the recall rate is increased to 1 in some datasets. The experimental results demonstrated that these advantages improve the positive-class classification performance to a greater extent than those achieved in previous studies.

The remainder of this paper is organized as follows. In Section 2, the cost-sensitive SVM model is briefly reviewed. In Section 3, the proposed cost-sensitive model is introduced to improve the recall rate of the positive classes. Section 4 describes the tests performed using the new method on actual data. Finally, Section 5 discusses and summarizes the advantages and disadvantages of the proposed method and suggests future research directions.

2. Related Work

In this section, we will briefly discuss different imbalanced dataset classification problems. The existing classification methods for imbalanced data can be roughly divided into two categories [22, 23]: data-level and algorithm-level methods [24–27]. We will first discuss the most effective approaches and then discuss the advantages and limitation of these proposed approaches.

Data-level approaches, which are also known as sampling methods [28], typically involve data preprocessing. These approaches rebalance highly skewed class distributions using various resampling methods, such as oversampling of the positive instances and undersampling of the negative instances, and at times, both methods are combined as well [29]. The simplest way to balance a dataset is by undersampling (randomly or selectively) the majority class, while keeping the original data of the minority class. However, this method results in loss of information of the majority class [30]. Another approach that can be used is oversampling in which the minority class instances are randomly duplicated to rebalance class distribution. Although oversampling does not result in loss of information of the majority class, it can cause overfitting. To solve this issue, Chawla et al. [31] proposed a method called Synthetic Minority Oversampling Technique (SMOTE) to generate new instances by linear interpolation between closely lying minority class samples. SMOTE generates new minority samples by interpolating between k-nearest minority class neighbors and has a better classification effect than random oversampling. However, the samples generated through this method may cause an overlap between the two categories.

In contrast, using algorithm-level approaches [32, 33], researchers have been able to introduce cost-sensitive learning to reduce the degree of imbalance by assigning a higher learning cost to positive-class samples [34–36]. Algorithm-level approaches directly modify the learning procedure to improve the sensitivity of the classifier toward minority classes. One such crucial approach to class-imbalanced learning was proposed by Veropoulos et al. [37], who used different penalty constants for different classes to assign higher costs to errors in classifying positive-class instances than those in classifying negative-class instances. However, this method does not consider the distance between the two types of samples and the classification hyperplane.

Some studies successfully applied the aforementioned methods in several different fields such as cancer diagnosis [38], sentiment analysis, and text classification [39]. In this study, we improved Veropoulos's method, modified the distance between the positive samples and the classification hyperplane, and developed an SVM classification algorithm with special classification purposes.

3. Cost-Sensitive Support Vector Machine

The goal of classification is to map feature vectors $x \in X$ to class labels $Y = \{-1, 1\}$ [40]. As in previous studies [41, 42], a training set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ is given, where x_i is an instance with an n-tuple of attribute values that belong to a certain instance space X and $y_i \in Y = \{-1, 1\}$ is a label.

A standard classification problem of a linear SVM can be expressed as

$$\arg\min_{w,\xi} \frac{1}{2}\|w^2\| + C\sum_{i=1}^{N}\xi_i, \tag{1}$$

subject to

$$\begin{aligned} y_i\left(w^t \cdot x + b\right) &\geq 1 - \xi_i, \\ \xi_i &\geq 0, \quad i = 1, 2, \ldots, N, \end{aligned} \tag{2}$$

where $C > 0$ is the penalty parameter.

The standard SVM model for the design of classification algorithms minimizes the probability of error, assuming that all misclassifications have the same cost. To control the classification recall, the penalty-regularized model proposed by Veropoulos et al. [37] was closely inspected. The key idea of this model is to introduce uneven loss functions to reweight the penalties of the samples in the imbalanced classes [7] and reduce the bias of the classification boundary toward the negative class. By predetermining the class labels,

$$\begin{aligned} I^+ &= \{i | y_i = +1, i = 1, 2, \ldots, N\}, \\ I^- &= \{i | y_i = -1, i = 1, 2, \ldots, N\}, \end{aligned} \tag{3}$$

where I^+ and I^- denote the index set for the positive and negative classes, respectively. When I^+ and I^- are categorized, different costs are assigned to the two classes. The standard SVM can be expanded to

$$\arg\min_{w,b,\xi} \frac{1}{2}\|w^2\| + C^+\sum_{i\in I^+}\xi_i + C^-\sum_{i\in I^-}\xi_i, \tag{4}$$

subject to

$$\begin{aligned} y_i\left(w^t \cdot x + b\right) &\geq 1 - \xi_i, \\ \xi_i &\geq 0, \quad I = 1, 2, \ldots, N, \end{aligned} \tag{5}$$

where C^+ is the cost of a false negative and C^- is the cost of a false positive. ξ_i are slack variables.

In the regularization model proposed by Veropoulos et al. [37], the weight vector, w, is a d-dimensional transposed vector normal to the decision boundary; the bias, b, is a scalar for offsetting the decision boundary; and the slack variables, ξ_i, measuring the losses are used to urge samples to satisfy the boundary constraints in the optimization. Thus, the cost value for the positive class is typically higher than that for the negative class. The dual problem of the primal model is written using Lagrange multipliers as follows:

$$\arg\max_{\alpha} \sum_i \alpha_i - \frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j k\left(x_i, x_j\right), \tag{6}$$

subject to

$$\sum \alpha_i^+ = \sum \alpha_i^-, \quad 0 \leq \alpha_i^+ \leq C^+, 0 \leq \alpha_i^- \leq C^-. \tag{7}$$

4. Increasing Minority Recall Model

The key objective of this study was to identify a misclassification cost value with a special purpose using a specific method, assuming that the costs of all types of misclassifications are not equal and that the true costs of misclassification cannot be determined. The goals are to increase the recall rate of the positive class of all datasets to 100% for physical problems and to improve the accuracy of the negative class as much as possible. When positive recall is increased to 1, the accuracy of the negative class may be affected but does not decrease significantly and is within an acceptable range.

4.1. Strategies to Improve Recall Rates for Minority Samples. Presently, in cost-sensitive learning, the cost-sensitive factor is often determined by a random interval or by using the sample number ratio between categories as the misclassification cost [43]. However, we developed a class of imbalanced datasets whose data structure enables us to search for misclassification costs with a "special purpose," which is increasing the positive recall rate to 1 because misclassification of positive samples can cause massive losses in physical problems. Modifying the loss function forces the classification algorithms to be biased toward the positive classes, and the classification boundary leans toward the negative class. The key idea is to adjust the margin of the positive class to cause the classification boundary to shift. Because the theoretical threshold of 0 is used as the judgment threshold of the sign function, as long as the 0 point is on the left side of the classification boundary, as many positive classes as possible will be included. Using the grid search method, the theoretical threshold and classification boundary are adjusted, resulting in a recall rate of 1.

When the data distribution is as shown in Figure 1, the use of this method will be limited because there is too much overlap between classes. The interface adjusted by us can classify positive-class samples in the imbalanced dataset not only correctly but also simultaneously. In addition, it can divide the samples of the negative classes in the overlapping area into samples of positive classes.

4.2. Proposed Support Vector Machine Model. The SVM uses minimization hinge loss function $L(yf) = \lfloor 1 - yf \rfloor_+$, where $\lfloor x \rfloor_+ = \begin{cases} x & x > 0 \\ 0 & x \leq 0 \end{cases}$.

Veropoulos et al. [37] extended the loss function in the biased support vector machine (B-SVM) classification as follows:

$$L^+(yf) = C^+\lfloor 1 - yf \rfloor_+, \tag{8}$$

$$L^-(yf) = C^-\lfloor 1 - yf \rfloor_+. \tag{9}$$

Equations 8 and (9) assign different cost values to instances in the positive and negative classes, respectively. The misclassification costs from samples in the negative class are generally exploited to outweigh those in the positive class. As our objective was to increase the recall rate of the positive-class samples to 1, we put a constraint on the positive-class margin and extended the loss function as follows:

$$L^{+}(yf) = C^{+}\lfloor 1 - kyf \rfloor_{+}, \tag{10}$$

$$L^{-}(yf) = C^{-}\lfloor 1 - yf \rfloor_{+}. \tag{11}$$

In Figure 2, C^{+}·controls the slope of the positive class, k controls the intersection of the abscissa axis, and the intersection is $1/k$. C^{-} controls the negative slope, and the intersection is 1. If the loss is 0, the classification confidence of the loss function must be sufficiently high. For the positive class, the loss is 0 when the degree of confidence is greater than $1/k$.

By replacing the original loss with the loss functions shown in (10) and (11), the original SVM can be extended to

$$\arg\min_{w,b,\xi} \frac{1}{2}\left\|w^{2}\right\| + kC^{+}\sum_{i\in I^{+}}\xi_{i} + C^{-}\sum_{i\in I^{-}}\xi_{i}, \tag{12}$$

subject to

$$\begin{aligned} y_{i}\left(w^{t}\cdot x_{i}+b\right) &\geq \frac{1}{k} - \xi_{i}^{+}, \quad i \in I^{+}, \\ y_{i}\left(w^{t}\cdot x_{i}+b\right) &\geq 1 - \xi_{i}^{-}, \quad i \in I^{-}, \\ \xi_{i} &\geq 0, \quad i = 1, 2, \ldots, N. \end{aligned} \tag{13}$$

Here, C^{+} and C^{-} are two types of costs, and the positive margin can be changed by adjusting the value of k. In addition to the adjustable penalty, the motivation of this study is to provide the loss function of the imbalanced classes a different hinge point. A biased decision boundary caused by imbalanced classes can be recovered with the help of a scalable margin. Herein, we develop the cost-sensitive model for solving the SVM class-imbalanced problem that has both an adjustable penalty and a scalable margin.

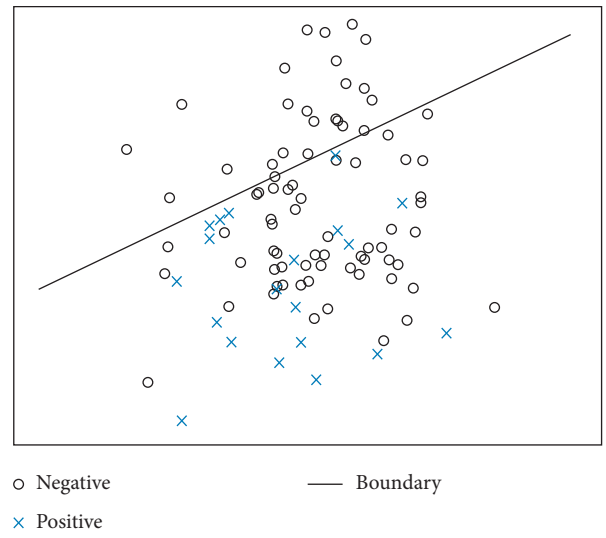

Figure 1: Limitations of the data distributions of various methods.

To solve the newly created problem, the Lagrangian function is introduced. The dual problem of the primal model can be written using Lagrange multipliers as follows:

$$L = \frac{1}{2}\left\|w^{2}\right\| + kC^{+}\sum_{i\in I^{+}}\xi_{i} + C^{-}\sum_{i\in I^{-}}\xi_{i} - \sum_{i\in I^{+}}\alpha_{i}\left[y_{i}\left(w^{t}\cdot x_{i}+b\right) - \frac{1}{k} + \xi_{i}\right] - \sum_{i\in I^{+}}\xi_{i}\mu_{i} - \sum_{i\in I^{-}}\alpha_{i}\left[y_{i}\left(w^{t}\cdot x_{i}+b\right) - 1 + \xi_{i}\right] - \sum_{i\in I^{-}}\xi_{i}\mu_{i}, \tag{14}$$

where $\alpha_{i} \geq 0$ and $\mu_{i} \geq 0$.

$$\frac{\partial L}{\partial w} \Longrightarrow w = \sum_{i}\alpha_{i}y_{i}x_{i}, \frac{\partial L}{\partial b} \Longrightarrow \sum_{I}\alpha_{i}y_{i} = 0, \frac{\partial L}{\partial \xi_{i}} \Longrightarrow \begin{cases} kC^{+} - \alpha_{i} - \mu_{i} = 0, & i \in I^{+} \Longrightarrow 0 \leq \alpha_{i} \leq kC^{+}, \\ C^{-} - \alpha_{i} - \mu_{i} = 0, & i \in I^{-} \Longrightarrow 0 \leq \alpha_{i} \leq C^{-}. \end{cases} \tag{15}$$

Therefore, the dual optimization model of (12) is defined as

$$\arg\max_{\alpha} -\frac{1}{2}\sum_{i}\sum_{j}\alpha_{i}\alpha_{j}y_{i}y_{j}K\left(x_{i}, x_{j}\right) + \sum_{i}\alpha_{i}\left(\frac{y_{i}+1}{2k} - \frac{y_{i}-1}{2}\right), \tag{16}$$

subject to

$$\sum \alpha_{i}^{+} = \sum \alpha_{i}^{-}, \tag{17}$$

$$0 \leq \alpha_{i} \leq kC^{+}, \quad i \in I^{+}, \tag{18}$$

$$0 \leq \alpha_{i} \leq C^{-}, \quad i \in I^{-}. \tag{19}$$

Our objective is to solve the dual problem (Algorithm 1).

4.3. Experiment

4.3.1. Performance Evaluation. In a classification problem, evaluation measures play a key role in assessing the performance of the classification model. The overall prediction accuracy is used to evaluate the classification of a balanced dataset; however, it is not an effective metric for an imbalanced dataset because it does not consider the prediction accuracy of either class. This lack of consideration is mainly due to the fact that the negative sample size is sometimes

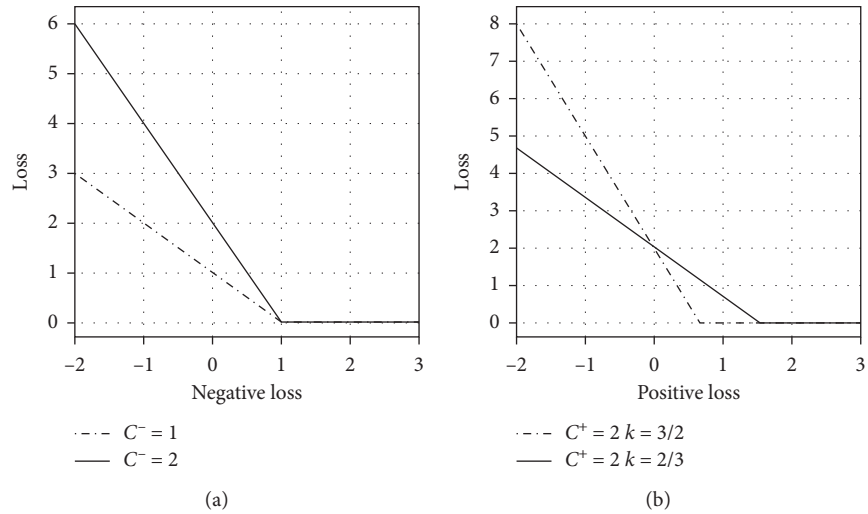

Figure 2: Loss function.

much larger than the positive sample size. In such cases, with an imbalance of 99 to 1, a classifier that classifies everything as negative will be 99% accurate, but it will be completely useless as a classifier. Therefore, more attention should be paid to the positive class. The current classification indicators are based on the confusion matrix presented in Table 1.

In the confusion matrix, true positive (TP) is the number of positive-class instances that have been correctly classified, true negative (TN) is the number of negative-class instances that have been correctly classified, false positive (FP) is the number of negative instances that have been incorrectly classified as positive, and false negative (FN) is the number of positive instances that have been incorrectly classified as negative.

To select the classification index of the positive class, we directly selected the recall (recall = TP/(TP + FN)) and specificity (specificity = TN/(TN + FP)), ensuring both a recall rate of 1 and high specificity.

To balance the effects of recall and specificity on the classification results, an evaluation index, g-means, can be constructed using the geometric mean of equilibrium recall and specificity:

$$g - \text{means} = \sqrt{\text{recall} * \text{specificity}}. \tag{20}$$

Although the number of positive samples may be very few, the omission and misjudgment of positive samples will be considered sufficient. Even if the classification accuracy rates for all the samples are excellent, the g-means value may be very low. The g-means indicator is effective for the classification evaluation of imbalanced data; however, it attaches equal importance to recall and specificity. As more attention was paid to recall in the present study, g-means was extended to recall $* g$-means, which increases the emphasis on recall. Ultimately, the proposed method performed the best with more emphasis on recall.

4.3.2. Experiments. We used ten class-imbalanced datasets with various positive ratios and compared our algorithm with the B-SVM [37], cost-sensitive support vector machine (CS-SVM) [44], and BP neural network algorithms [45], as well as two special classification costs. These experiments proved that it is feasible to adjust the positive-class margin to achieve a positive recall rate of 1.

4.3.3. Open Dataset. As shown in Table 2, we compared the classification differences between the proposed method and other methods under different classification standards. We selected ten real-world imbalanced datasets from the UCI machine learning data repository [46]. Among them, ecoli1, ecoli2, glass6, car1v3, car1v4, glass5, segment1, and glass6 were constructed using a multiclassification problem.

4.3.4. Experimental Design. We compared the proposed method with several SVM-based methods and neural network, including the B-SVM, CS-SVM, and BP approaches, as well as two special classification costs (no cost and prorated cost). For the prorated cost, we set the misclassification cost based on the following equation:

$$\frac{\text{misclassification cost (positive class)}}{\text{misclassification cost (negative class)}} = \frac{\text{total number of negative instances}}{\text{total number of positive instances}}. \tag{21}$$

Algorithm:
Positive margin adjustment using SVM.
Given: a sequence of N examples X_{Train} and $X_{\text{Validation}}$
Output: G #Output combination classifier
Variables:
α #Karush–Kuhn–Tucker conditions (KKT) initial alpha
ρ #G-means value
C_p, C_n, k #Positive cost, Negative cost, Positive margin calibration variable
T #the selected running iterations
Function:
S #classifier model
$R * G\text{-}means\ (G)$ #Obtain the Recall $*$ G-means values from G
Begin
Initialize
$\alpha^{(0)} = 0$
$\rho^{(0)} = 0$
$T = 1$
Set the 3D grid search range of C_p, C_n, k:
(a) Select optimization variables $\alpha_1^{(n)}$ and $\alpha_2^{(n)}$ and solve the optimization problem using the sequential minimal optimization (SMO) algorithm to obtain $\alpha_1^{(n+1)}$ and $\alpha_2^{(n+1)}$, and update α to $\alpha_1^{(n+1)}$.
(b) If the KKT conditions in (16)–(19) are satisfied within the allowable range of precision, ε, the KKT condition can be used for the next step; otherwise, continue with process (b).
(c) Get $\alpha^* = \alpha^{(n+1)}$.
(d) Finally, $\alpha^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_N^*)^T$ is obtained, and w^* and b^* are calculated as follows:
$w^* = \sum_{i=1}^{N} \alpha_i^* y_i x_i$, $b^* = y_j - \sum_{i=1}^{N} \alpha_i^* y_i (x_i, x_j)$.
nstruct a classifier model $S = S = w^* x + b$
(e) $G = sign\ (S)$
(f) $\rho^{(n)} = R * G\text{-}means\ (G(X_{Validation}))$ at a condition recall of 1
(g) If $\rho^{(n)} > \rho_{\text{best}}$, then $\rho_{\text{best}} = \rho^{(n)}$
Return G_t
End

Algorithm 1: Algorithm to increase the recall rate to 1.

Table 1: Confusion matrix.

	Predicted as the positive class	Predicted as the negative class
True positive class	TP	FN
True negative class	FP	TN

Table 2: Descriptions of the experimental datasets.

Name	Example	Dimension	Rate of imbalance	No. of classes
Breast cancer o	699	10	1.9	2
Breast cancer d	569	32	1.933	2
Ecoli1	336	7	3.36	2
Ecoli2	336	7	5.46	2
Glass6	214	9	6.38	2
Car1v3	1279	6	17.54	2
Car1v4	1275	6	18.62	2
Glass5	214	9	22.78	2
Yeast5	1484	8	32.73	2
Segment1	2310	19	6	2

The RBF Gaussian kernel $K(x, z) = \exp(-\|x - z^2\|/2\sigma^2)$ is used for all SVM algorithms, where σ is the bandwidth parameter that must be predetermined. We attempted to select the best σ value through a grid search of the σ range between 0.01 and 20. For different datasets, while searching for parameters such as C and k, different grid widths were selected to accelerate parameter selection. For both positive and negative class costs, the range of [0.01, 100] was selected, and the search range of K was fixed at [0.01, 10]. Logistic was selected as the activation function. The learning rate was selected within the range of [0.000001, 1], the number of hidden layers was selected from [1–3], and the search range of the number of neurons in each hidden layer was [2, 1000]. Five-fold cross-validation was used in each set of parameter experiments to determine the set of parameters with the best generalization performance. We chose the recall, specificity, g-means, and area under the receiver operating characteristic curve (AuROC) instead of global evaluation indicators to evaluate the performance of the classification method for imbalanced datasets. No cost and prorated cost were classified using a judgment threshold of 0, whereas the B-SVM, CS-SVM, and BP

approaches used AuROC as the classification standard for the recall, specificity, and g-means.

5. Results and Discussion

To demonstrate the ability of the proposed method to adjust the classification boundary visually, we select two sets of data (yeast5 and breast cancer d) and projected them into a two-dimensional space for visualization. Firstly, to demonstrate the influence of the value of k on the change in the decision boundary under the two sets of data, we used different k values, as shown in Figure 3. In Figure 3, from left to right, the values of k are as follows: $k = 1$, $k = 3$, $k = 2.416$, $k = 1$, $k = 3$, and $k = 2$. The values were selected after a broad survey, and the recall of the adjusted k value of both datasets is 1. For the breast cancer d dataset, it is obvious that the classification effect of the adjusted k value is better than the k value that is randomly selected by the other two groups; hence, we will not explain this dataset in detail. However, for the yeast5 dataset, the other two kinds of k will also reach a recall of 1. The preliminary observation is that the classification effect when $k = 1$ is similar to that when $k = 2$; therefore, we compared their respective g-means. When $k = 1$, g-means = 0.936; when $k = 3$, g-means = 0.900; and when $k = 2$, g-means = 0.964. The comparison revealed that the adjusted k value not only increased the recall to 1 but also did not reduce the g-means value. Therefore, changing the margin of the positive class can adjust the position of the classification boundary so that the positive classes can be properly classified by the boundary.

In addition, the receiver operating characteristic (ROC), through which the threshold is determined, is a widely used evaluation index for classification problems. For the classification problem, we obtained a set of predicted values and classified the data by traversing the predicted values and used the predicted values as thresholds. Predicted values less than the threshold are classified as negative, and predicted values greater than the threshold are classified as positive. Therefore, for each set of predicted values, we can determine a unique set of TPR and FPR. The threshold determination criterion of the ROC curve is maximum (true positive rate (TPR)-false positive rate (FPR)) when the TPR-FPR is the largest, and the corresponding threshold is the threshold of the ROC curve.

Frequently, the classification threshold under the ROC criterion cannot increase the recall to 1, such as in the experiments on four datasets whose results are presented in Figure 4. The experiments prove that the threshold converges to a certain value as k increases. As shown in Figure 4, the classification ROC threshold of the four datasets decreases gradually with increasing k and exhibits an obvious trend of asymptotic stability. Based on the experimental results, an appropriate threshold can be chosen to ensure that the requirement of 1 can be achieved again by adjusting k. We completed five cross-validation experiments to reduce randomness, and the results indicate that our scheme is universal and widely applicable.

Table 3 summarizes the recall and specificity obtained with different methods. In Tables 4–6, the best results are highlighted (bold font). To examine the significance of the positive examples in the imbalanced datasets, we studied the recall of the positive examples, as presented in Table 4. In the datasets shown, the average recall rate of the proposed method is the largest, effectively increasing the recall rate to 1 in each case. The g-means index is shown in Table 5. It can be observed from Table 5 that although the average g-means value of our method is not the highest, this value is not reduced much compared to that in the other three ideal classification cases. Moreover, the g-means value of our proposed method is higher than that of the other five cases for some datasets. It can be observed that when the recall rate of the positive class reaches 1, the accuracy of the negative class does not decrease significantly and is within the acceptable range. These results demonstrate that the proposed method outperforms the other methods for the positive class of imbalanced datasets and that there may be a certain degree of improvement in the g-means index.

Figure 5 shows the combined g-means indicators obtained using several methods. Because we consider the recall rate to be extremely important, g-means multiplied by recall is n, used as the new indicator, which improves the weight of recall in g-means, so that the new indicator weighted emphasis on the recall. The g-means indices that attach more importance to recall are presented in Table 6. When the new indicator is adopted, our proposed method has the highest average value.

Figure 6 clearly depicts the recall $*g$-means evaluation indicator under the comprehensive influence of the classification effects of the positive and negative classes. These results verify that the proposed method performs well on all the datasets.

For the statistical analysis, we implement Student's t-test to verify whether significant differences exist between the proposed method and other methods in the experiment. The t-value in Student's t-test is calculated as follows:

$$t = \frac{\overline{x}}{\sqrt{\left(\sigma^2/n\right)}}, \tag{23}$$

where $\overline{x}$represents the example mean of the data; σ is the standard variance of the data; and n is the sample size. In this case, the sample size is set to 10. As a case study, we compare the proposed method with other methods. We calculate the t-value using recall and g-means data listed in Tables 4 and 5. The null hypothesis should be H_0: $\overline{x}_1 = \overline{x}_2$, and the alternative hypothesis should be H_1: $\overline{x}_1 \neq \overline{x}_2$. Let x_1 be the sample mean obtained by the proposed method and x_2 be the sample mean of the other three methods considered for the comparison. The same is true for the g-means test. Three t-tests were conducted for the three models listed in Table 4, and the results were compared. For the recall of BP, CS-SVM, and B-SVM models, the t-value obtained is 2.258, 2.631, and 2.671, respectively. We found that the t-value is 1.813 with the probability threshold of 0.05 using Student's t-distribution table. The calculated t-values 2.258, 2.631, and 2.671 are all greater than the t-value 1.813; thus, at the 0.05 level of significance, the null hypothesis is rejected in favor of the alternative hypothesis. The recall value obtained by the proposed method is greater than that of other methods. For the t-test of the g-means of BP, CS-SVM, and B-SVM

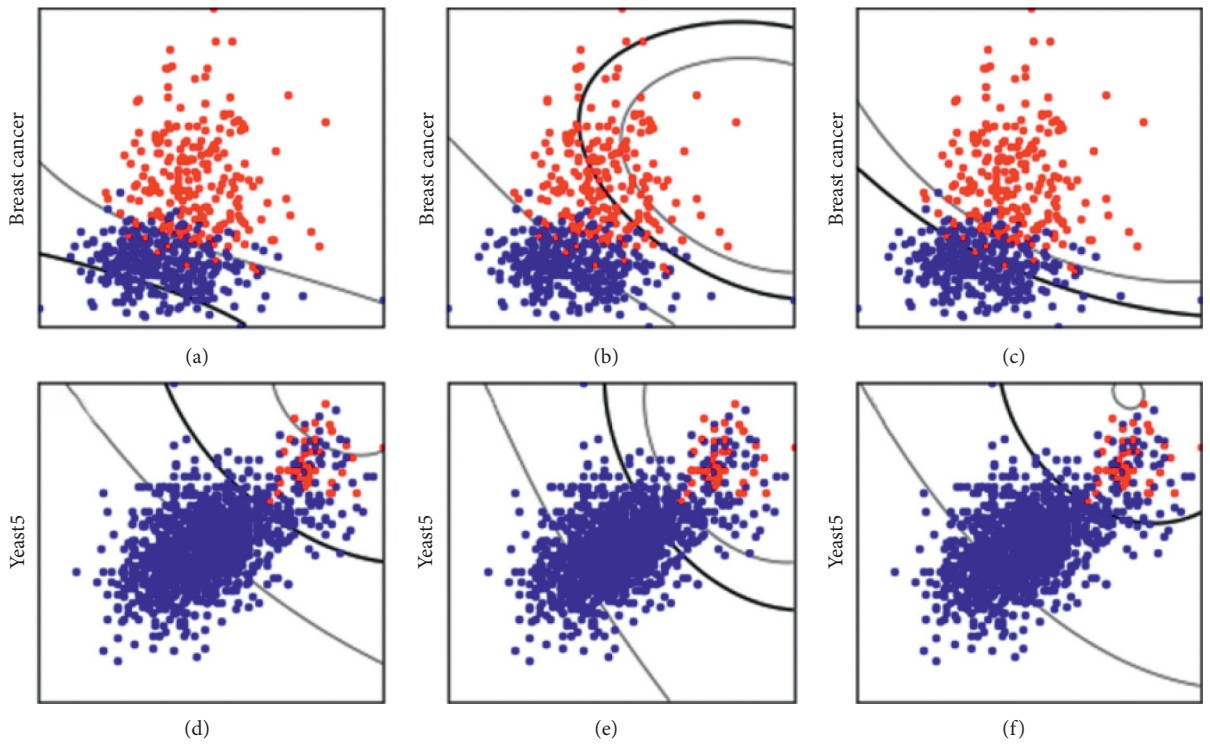

Figure 3: Comparative observation of the decision boundaries with various k values. The decision boundaries are represented by the black lines and the margins by the gray lines. (a) $k = 1$. (b) $k = 3$. (c) $k = 2.416$. (d) $k = 1$. (e) $k = 3$. (f) $k = 2$.

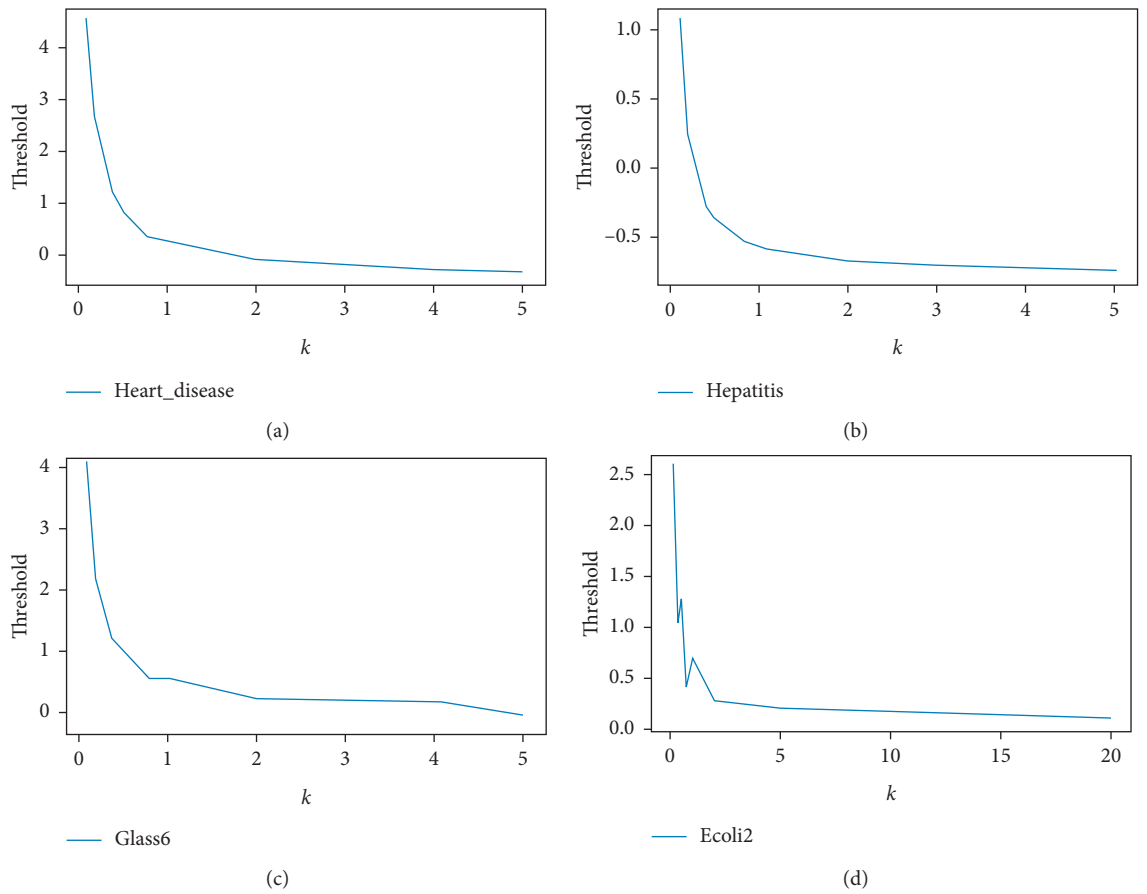

Figure 4: Influence of k on the ROC threshold.

Table 3: Recall and specificity (represented by r and s, respectively).

Dataset	No cost		Prorated cost		B-SVM		CS-SVM		BP		Proposed method	
	r	s	r	s	r	s	r	s	r	s	r	s
Breast cancer o	0.983	0.921	0.749	0.891	0.975	0.958	0.988	0.958	0.988	0.969	1	0.930
Breast cancer d	0.390	0.939	0.872	0.861	0.920	0.936	0.920	0.927	0.925	0.964	1	0.701
Ecoli1	0.811	0.796	0.775	0.250	0.898	0.927	0.910	0.873	0.975	0.829	1	0.807
Ecoli2	0.982	0.230	0.742	0.732	0.865	0.932	0.942	0.937	0.793	0.835	1	0.823
Glass6	0.867	0.816	0.333	0.795	0.900	0.978	0.933	0.962	0.893	0.984	1	0.919
Car1v3	0.832	0.830	1.000	0.956	1.000	0.990	1.000	0.997	1.000	0.935	1	0.997
Car1v4	0.908	1.000	0.954	1.000	1.000	1.000	1.000	1.000	1.000	0.955	1	1.000
Glass5	0.100	1.000	0.400	0.990	1.000	0.937	1.000	0.932	1.000	0.624	1	0.922
Yeast5	0.881	0.903	1.000	0.283	1.000	0.938	1.000	0.956	0.728	0.667	1	0.905
Segment1	0.188	1.000	0.188	1.000	0.992	0.997	0.994	0.997	1.000	0.995	1	0.994
Average	0.694	0.844	0.701	0.776	0.955	0.959	0.969	0.954	0.930	0.876	1	0.900

Table 4: Recall.

Dataset	No cost	Prorated cost	B-SVM	CS-SVM	BP	Proposed method
Breast cancer o	0.983	0.749	0.975	0.988	0.988	**1.000**
Breast cancer d	0.390	0.872	0.920	0.920	0.925	**1.000**
Ecoli1	0.811	0.775	0.898	0.910	0.975	**1.000**
Ecoli2	0.982	0.742	0.865	0.942	0.793	**1.000**
Glass6	0.867	0.333	0.900	0.933	0.893	**1.000**
Car1v3	0.832	1.000	1.000	1.000	1.000	**1.000**
Car1v4	0.908	0.954	1.000	1.000	1.000	**1.000**
Glass5	0.100	0.400	1.000	1.000	1.000	**1.000**
Yeast5	0.881	1.000	1.000	1.000	0.728	**1.000**
Segment1	0.188	0.188	0.992	0.994	1.000	**1.000**
Average	0.694	0.701	0.955	0.969	0.930	**1.000**

Table 5: g-means index.

Dataset	No cost	Prorated cost	B-SVM	CS-SVM	BP	Proposed method
Breast cancer o	0.951	0.817	0.966	0.973	**0.978**	0.964
Breast cancer d	0.605	0.866	0.928	0.923	**0.944**	0.837
Ecoli1	0.803	0.440	**0.912**	0.891	0.899	0.898
Ecoli2	0.475	0.737	0.898	**0.939**	0.814	0.907
Glass6	0.841	0.515	0.938	0.947	0.937	**0.959**
Car1v3	0.831	0.978	0.995	**0.998**	0.967	**0.998**
Car1v4	0.953	0.977	**1.000**	**1.000**	0.977	**1.000**
Glass5	0.316	0.629	**0.968**	0.965	0.790	0.960
Yeast5	0.892	0.532	0.969	**0.978**	0.697	0.951
Segment1	0.434	0.434	0.994	0.995	**0.997**	**0.997**
Average	0.710	0.693	**0.957**	**0.957**	0.900	0.947

Table 6: Recall $^*g$-means.

Dataset	No cost	Prorated cost	B-SVM	CS-SVM	BP	Proposed method
Breast cancer o	0.935	0.612	0.942	0.961	**0.966**	0.964
Breast cancer d	0.236	0.755	0.854	0.849	**0.873**	0.837
Ecoli1	0.651	0.341	0.819	0.811	0.877	**0.898**
Ecoli2	0.466	0.547	0.777	0.885	0.646	**0.907**
Glass6	0.729	0.171	0.844	0.884	0.837	**0.959**
Car1v3	0.691	0.978	0.995	**0.998**	0.967	**0.998**
Car1v4	0.865	0.932	**1.000**	**1.000**	0.977	**1.000**
Glass5	0.032	0.252	**0.968**	0.965	0.790	0.960
Yeast5	0.786	0.532	0.969	**0.978**	0.507	0.951
Segment1	0.082	0.082	0.986	0.989	**0.997**	**0.997**
Average	0.547	0.520	0.915	0.932	0.840	**0.952**

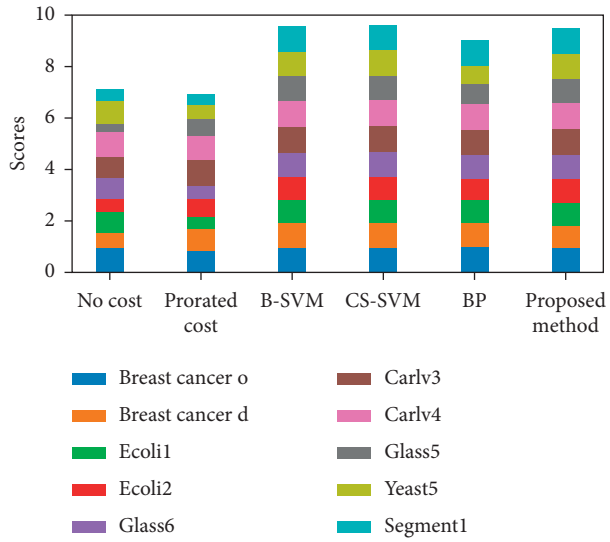

Figure 5: Comparisons of the classification results' g-means indices.

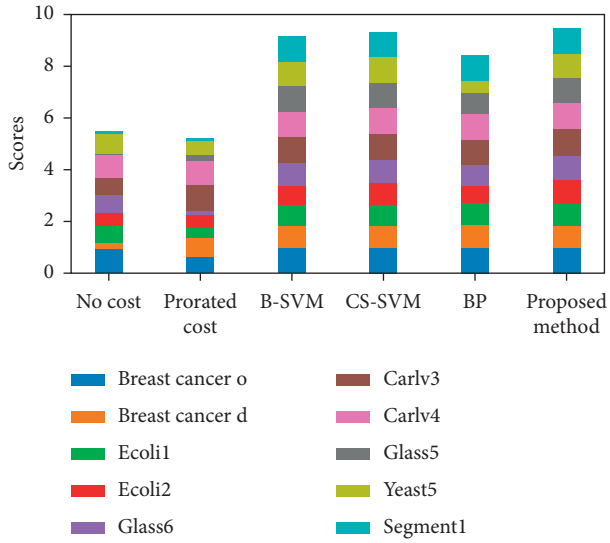

Figure 6: Comparisons of the recall $*g$-means indices of the classification results.

models, the t-value obtained is 1.320, −0.689, and −0.483, respectively. The calculated t-values 1.320, −0.689, and −0.483 are all lower than the t-value 1.813; thus, the null hypothesis cannot be rejected. Therefore, at the 0.05 level of significance, we believe that no significant difference exists in the g-means indicators obtained by several models.

Based on the above conclusions, at the 0.05 level of significance, the recall of the proposed method is significantly greater than that of other methods, and no significant difference exists in the g-means.

6. Conclusions

In this paper, a cost-sensitive SVM algorithm based on an imbalanced margin was proposed for the classification of imbalanced data. This method was based on the theory proposed by Veropoulos et al. [37], and its feasibility was verified using both theoretical and experimental results. The recall rate of small classes was improved by adjusting the SVM positive classification margin. The proposed method was also compared with other traditional methods. The experimental results demonstrate that a small-class recall rate of 1 can be achieved using the proposed method. However, the proposed approach still has some disadvantages. Specifically, the accuracy of positive classes is lost in some datasets, but in many cases, the performance improves compared with that of the traditional methods. When the classification evaluation criteria are changed (i.e., more emphasis is placed on the positive classes), the average evaluation index of the proposed method is the highest. Such classification results are of great significance in the fields of finance, medicine, engineering, and astronomy, to name some. In future work, we will test the experimental setup employed in this study using different machine learning models and attempt to apply this method to practical problems. Additionally, we will extend the proposed method to multiclass classification problems by adopting a one-versus-all approach.

Data Availability

The datasets used to support the findings of this study have been deposited in the UCI machine learning repository (http://archive.ics.uci.edu/ml/datasets).

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors would like to thank Editage (http://www.editage.com) for English language editing. This work was supported by Heilongjiang Province Statistical Science Project (no. 2020B06).

References

[1] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: a review," *GESTS International Transactions on Computing in Science and Engineering*, vol. 30, no. 1, pp. 25–36, 2006.

[2] G. Y. Wong, F. H. F. Leung, and S.-H. Ling, "A hybrid evolutionary preprocessing method for imbalanced datasets," *Information Sciences*, vol. 454, pp. 161–177, 2018.

[3] Q. Yang and X. Wu, "10 challenging problems in data mining research," *International Journal of Information Technology & Decision Making*, vol. 5, no. 4, pp. 597–604, 2006.

[4] F. Provost, "Machine learning from imbalanced data sets," in *Proceedings of the AAAI'2000 Workshop on Imbalanced Data Sets*, pp. 1–3, Menlo Park, CA, USA, April 2000.

[5] G. M. Weiss, "Mining with rarity," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 7–19, 2004.

[6] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," in *Proceedings of the 14th*

International Conference on Machine Learning, pp. 179–186, Nashville, TN, USA, July 1997.

[7] M. J. Siers and M. Z. Islam, "Novel algorithms for cost-sensitive classification and knowledge discovery in class imbalanced datasets with an application to NASA software defects," *Information Sciences*, vol. 459, pp. 53–70, 2018.

[8] S. Barua, M. M. Islam, X. Yao, and K. Murase, "MWMOTE-Majority weighted minority oversampling Technique for imbalanced data set learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 2, pp. 405–425, 2014.

[9] G. Wu and E. Chang, "Class-boundary alignment for imbalanced dataset learning," in *Proceedings of the ICML 2003 Workshop on Learning from Imbalanced Data Sets II*, pp. 49–56, Washington, DC, USA, 2003.

[10] B. W. Yap, K. A. Rani, H. A. A. Rahman, S. Fong, Z. Khairudin, and N. N. 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)*, pp. 13–22, Singapore, 2014.

[11] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.

[12] I. Martin-Diaz, D. Morinigo-Sotelo, O. Duque-Perez, and R. de J. Romero-Troncoso, "Early fault detection in induction motors using Adaboost with imbalanced small data and optimized sampling," *IEEE Transactions on Industry Applications*, vol. 53, no. 3, pp. 3066–3075, 2017.

[13] A. Zakaryazad and E. Duman, "A profit-driven Artificial Neural Network (ANN) with applications to fraud detection and direct marketing," *Neurocomputing*, vol. 175, pp. 121–213, 2016.

[14] Y.-H. Liu, Y.-C. Liu, and Y.-Z. Chen, "High-speed inline defect detection for TFT-LCD array process using a novel support vector data description," *Expert Systems with Applications*, vol. 38, no. 5, pp. 6222–6231, 2011.

[15] D. M. J. Tax and R. P. W. Duin, "Support vector data description," *Machine Learning*, vol. 54, no. 1, pp. 45–66, 2004.

[16] M. Wu and J. Ye, "A small sphere and large margin approach for novelty detection using training data with outliers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 2088–2092, 2009.

[17] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[18] C.-C. Chang and C.-J. Lin, *LIBSVM: A Library for Support Vector Machines*, National Taiwan University, Taipei, Taiwan, 2000, http://www.csie.ntu.edu.tw/?cjlin/libsvmS.%20Department%20of%20Computer%20Science.

[19] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, UK, 2000.

[20] B. Schölkopf and A. J. Smola, *Learning with Kernels*, MIT Press, Cambridge, MA, USA, 2002.

[21] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.

[22] D. Ramyachitra and P. Manikandan, "Imbalanced dataset classification and solutions: a review," *International Journal of Computing and Business Research*, vol. 5, no. 4, 2014.

[23] A. Rocha and S. Klein Goldenstein, "Multiclass from binary: expanding one-versus-all, one-versus-one and ecoc-based approaches," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 2, pp. 289–302, 2014.

[24] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: special issue on learning from imbalanced data sets," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 1–6, 2004.

[25] B. Scholkopf and A. J. Smola, "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond," *MIT Press*, Cambridge UK, 2002.

[26] A. Onan, "Consensus clustering-based undersampling approach to imbalanced learning," *Scientific Programming*, vol. 2019, Article ID 5901087, 2019.

[27] A. Ha and R. Ezzahir, "Sampling techniques for Arabic sentiment classification: a comparative study," in *Proceedings of the 3rd International Conference on Networking, Information Systems & Security*, Rabat, Morocco, 2020.

[28] J. Błaszczy and J. Stefanowski, "Neighbourhood sampling in bagging for imbalanced data," *Neurocomputing*, vol. 150, pp. 529–542, 2015.

[29] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on *k*-means and SMOTE," *Information Sciences*, vol. 465, pp. 1–20, 2018.

[30] H. Haibo and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[31] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, 2002.

[32] W. Gao, R. Jin, S. Zhu, and Z. Zhou, "One-pass AUC optimization," *Proceedings of the 30th International Conference on Machine Learning*, pp. 906–914, Atlanta, GA, USA, June 2013.

[33] G. J. Karakoulas and J. Shawe-Taylor, "Optimizing classifiers for imbalanced training sets," in *Advances in Neural Information Processing Systems*, M. S. Kearns, S. A. Solla, and D. A. Cohn, Eds., pp. 253–259, MIT Press, Cambridge, MA, USA, 1999.

[34] P. Domingos, "MetaCost: a general method for making classifiers cost sensitive," in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 155–164, San Diego, CA, USA, August 1999.

[35] C. Elkan, "The foundations of cost-sensitive learning," in *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01)*, pp. 973–978, New York, NY, USA, August 2001.

[36] K. M. Ting, "An instance-weighting method to induce cost-sensitive trees," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 3, pp. 659–665, 2002.

[37] K. Veropoulos, C. Campbell, and N. Cristianini, "Controlling the sensitivity of support vector machines," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Montreal, Quebec, Canada, August 1999.

[38] A. Onan, "A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer," *Expert Systems with Applications*, vol. 42, no. 20, pp. 6844–6852, 2015.

[39] B. Huang, Y. Zhu, Z. Wang et al., "Imbalanced data classification algorithm based on clustering and SVM," *Journal of Circuits, Systems and Computers*, vol. 30, Article ID 2150036, 2020.

[40] H. Duan, X. Shao, W. Hou, G. He, and Q. Zeng, "An incremental learning algorithm for Lagrangian support vector machines," *Pattern Recognition Letters*, vol. 30, no. 15, pp. 1384–1391, 2009.

[41] C. X. Jian, J. Gao, and Y. H. Ao, "A new sampling method for classifying imbalanced data based on support vector machine ensemble," *Neurocomputing*, vol. 193, pp. 15–122, 2016.

[42] W. Lee, C.-H. Jun, and J.-S. Lee, "Instance categorization by support vector machines to adjust weights in AdaBoost for imbalanced data classification," *Information Sciences*, vol. 381, pp. 92–103, 2017.

[43] H. G. Chew, R. E. Bogner, and C. C. Lim, "Dual/spl nu/-support vector machine with error rate and training size biasing," in *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings (Cat. No.01CH37221)*, pp. 1269–1272, Salt Lake City, UT, USA, May 2001.

[44] A. Iranmehr, H. Masnadi-Shirazi, and N. Vasconcelos, "Cost-sensitive support vector machines," *Neurocomputing*, vol. 343, pp. 50–64, 2019.

[45] Y. Wang and N. Wang, "An unbalanced data classification model to increase minority recall for medical application," *Basic and Clinical Pharmacology and Toxicology*, vol. 127, p. 203, 2020.

[46] K. Bache and M. Lichman, *UCI Machine Learning Repository*, University of California, School of Information and Computer Science, Irvine, CA, USA, 2013, http://archive.ics.uci.edu/ml.