Handling Imbalanced Classification Problems With Support Vector Machines via Evolutionary Bilevel Optimization

Alejandro Rosales-Pérez, Salvador García, and Francisco Herrera, Senior Member, IEEE

Abstract—Support vector machines (SVMs) are popular learning algorithms to deal with binary classification problems. They traditionally assume equal misclassification costs for each class; however, real-world problems may have an uneven class distribution. This article introduces EBCS-SVM: evolutionary bilevel cost-sensitive SVMs. EBCS-SVM handles imbalanced classification problems by simultaneously learning the support vectors and optimizing the SVM hyperparameters, which comprise the kernel parameter and misclassification costs. The resulting optimization problem is a bilevel problem, where the lower level determines the support vectors and the upper level the hyperparameters. This optimization problem is solved using an evolutionary algorithm (EA) at the upper level and sequential minimal optimization (SMO) at the lower level. These two methods work in a nested fashion, that is, the optimal support vectors help guide the search of the hyperparameters, and the lower level is initialized based on previous successful solutions. The proposed method is assessed using 70 datasets of imbalanced classification and compared with several state-of-the-art methods. The experimental results, supported by a Bayesian test, provided evidence of the effectiveness of EBCS-SVM when working with highly imbalanced datasets.

Index Terms—Bilevel optimization (BLO), data preprocessing, evolutionary algorithms (EAs), imbalanced classification, support vector machines (SVMs).

I. INTRODUCTION

Support vector machines (SVMs) [1] are among the most popular supervised learning algorithms, with strong theoretical foundations and high effectiveness in real-world problems. The idea behind SVMs is to find the hyperplane that maximizes the separation margin between two categories. In their canonical form, SVMs assume an equal cost for each class. This assumption works well when the number of instances for each class is roughly similar. However, real-world problems seldom have a balanced class distribution.

The imbalanced classification problem refers to an uneven class distribution [2]–[4], that is, there is an overrepresented class, known as the majority class, and an underrepresented class, known as the minority class. Imbalanced classification problems further have the characteristic that the minority class is the one of interest. Therefore, accurately recognizing the minority class becomes crucial in several applications, such as medical diagnosis, fraud detection, and face recognition.

There are two main approaches to deal with imbalanced classification problems with SVMs: 1) data-level (DL) and 2) algorithm-level (AL) methods [5]. The first approach aims to balance the dataset by oversampling the minority class [3], [6]–[8] or undersampling the majority class [3], [9]–[11]. Then, SVM learns from the edited dataset [12]. Although DL methods are flexible, they ignore the particularities of learning algorithms. Conversely, AL methods modify the learning algorithm to be robust to uneven class distributions. For SVMs, common modifications comprise hyperplane shifting [13], [14], kernel adaptation [15], and cost sensitive [16]. AL methods often offer better performance than DL ones [16]; however, they need to define a set of hyperparameters, such as the extent of shifting compensation or the correct costs for each class. Therefore, methods for imbalanced classification must not only learn when class distributions are unequal but their hyperparameters must also be tuned to get peak performance.

Bilevel optimization (BLO) arises as an alternative for hyperparameter optimization. BLO differs from traditional optimization in that the optimization problem has as part of its constraints a second optimization problem. In our context, the principal optimization or upper level problem is the hyperparameter optimization, and the second or lower level problem is the learning of support vectors. These two problems interplay, that is, the definition of the hyperparameters influences the optimal set of support vectors, and this set of support vectors defines the model to predict unseen cases. However, BLO

1The misclassification costs are the weights applied to errors incurred by classifying positive or negative samples.
problems are computationally challenging because of their nonconvexity and nonlinearity [17], [18].

Evolutionary algorithms (EAs) are powerful search tools capable of solving complex optimization problems, such as BLO problems. Recently, the interest in using EAs to address machine learning problems is growing fastly [19]–[29]. For imbalanced learning, EAs have been used for data sampling [30], [31] and cost-sensitive learning [32]. Although recent studies address the problem of determining the optimal misclassification costs [32], [33], they have paid little attention to considering the hyperparameters of the learning algorithm, along with exploiting the hierarchical nature of parameter and hyperparameter learning to guide search. Furthermore, taking advantage of the properties of the learning algorithm to estimate the classification performance efficiently in imbalanced problems is almost unexplored.

In the light of the above mentioned, this article introduces EBSCS-SVM: evolutionary bilevel cost-sensitive SVMs. EBSCS-SVM combines an EA and the sequential minimal optimization (SMO) algorithm in a nested manner. The EA optimizes the cost of hyperparameters, which are the costs of each class and the kernel parameters, and the SMO learns the optimal support vectors. These two optimizers interact such that information from one level is used by the other to improve search and convergence capabilities. We summarize the main contributions of this article as follows.

1) We propose EBSCS-SVM, which allows learning SVMs in imbalanced classification problems and automatically sets the misclassification costs and kernel parameter.
2) EBSCS-SVM uses the information from the lower level to guide the search to the upper level and takes advantage of the previous successful hyperparameters to initialize the set of support vectors.
3) The bilevel formulation that jointly learns parameters and hyperparameters.
4) The definition of the upper level objective function that allows estimating the performance of the SVM without performing cross-validation.

The performance of EBSCS-SVM was assessed using a suite of 70 benchmark datasets of imbalanced classification and compared with the state-of-the-art methods. The experimental evaluation revealed an outstanding efficacy of EBSCS-SVM when faced with problems with a high disproportion of classes. The hierarchical Bayesian test supported the main findings.

The remainder of this article is organized as follows. Section II introduces the related works on imbalanced classification problems, hyperparameter optimization, and BLO. Section III describes the general bilevel formulation for cost-sensitive SVM. Next, Section IV describes the proposed EBSCS-SVM. Section V details the datasets, reference methods, and performance measures, while Section VI presents the experimental results. Finally, Section VII discusses the main conclusions.

II. RELATED WORK

This section presents the preliminaries. Section II-A describes the main approaches for imbalanced classification. Then, Section II-B describes the hyperparameter optimization problem, and Section II-C presents the main concepts related to BLO.

A. Methods for Imbalance Classification

Most learning algorithms may face difficulties when dealing with imbalanced classification problems, as they can favor the majority class, leading to an ineffective classification model. Two main approaches to dealing with imbalanced datasets are: 1) sampling strategies and 2) algorithm adaptation [2]–[4]. The former works with training data by modifying its class distribution, while the latter adjusts the training algorithm or inference process to consider the imbalance. We describe these two approaches as follows.

1) Data-Level Preprocessing Methods: This approach aims to reduce the effect of class imbalance by adding or removing samples from training data to balance the class distribution [34]. There are two primary sampling strategies, which are as follows.

1) Oversampling methods attempt to balance the dataset by replicating or creating samples from the minority class. Random oversampling (ROS) [3] is the simplest method for data balancing that replicates samples from the minority class. SMOTE [6] is a popular method for generating artificial samples through a linear interpolation of samples from the minority class. Variants of SMOTE include SVMSMOTE [7] and ADASYN [8]. The major criticism of oversampling methods is that the synthetic samples can cause overfitting of the classification model.

2) Undersampling methods balance the dataset by removing instances from the majority class. Random undersampling (RUS) [3] is an uninformed method that removes instances from the majority class at random. The condensed nearest neighbor (CNN) [9] is an informed method that eliminates examples distant from the boundary decision. The major criticism of undersampling methods is that they can discard meaningful instances and lead to loss of information.

It is unclear whether oversampling is better than undersampling or vice versa [35], [36]. Both methods are effective in handling imbalanced classification problems.

2) Algorithm-Level Methods: This approach aims to adapt the way a particular classifier learns in such a manner that it can deal with imbalanced problems. For SVMs, there are three main approaches of adaptation.

1) Cost-sensitive methods consider different costs to each class during learning, such that minority class errors have a higher penalization than those of the majority class. SVMs can work in a cost-sensitive framework by using different regularization parameters for positive and negative samples [37]. Also, an instance can be weighted based on the density of its neighborhood [38]. However, the proper setting of the costs is unknown.

2) Kernel adaptation methods adapt the kernel function or kernel matrix to reduce the bias toward the majority class. For example, WK-SMOTE [15] expands the kernel matrix by incorporating the dot products of
artificial samples generated in the feature space. Then, the SVM training algorithm uses the modified kernel matrix to learn a model.

3) **Hyperplane shifting** methods shift the separating hyperplane to enlarge the margin around minority class [14]. The major criticism of the AL methods is that they are algorithm-specific and require in-depth knowledge of the classifier. However, these methods are more accurate than DL methods [2].

**B. Hyperparameter Optimization**

Hyperparameter optimization refers to the problem of automatically setting the hyperparameter configuration of a learning algorithm to optimize the performance. Hyperparameter optimization is a complicated problem with several challenges. The challenges include computationally expensive evaluations of the objective function, a complex and nonconvex search space, hyperparameters that cannot be differentiable, and a finite amount of data that may limit the estimation of the generalization performance [39]. Formally, the hyperparameter optimization problem can be stated as follows [39]:

$$\theta^* = \arg \min_{\theta \in \Theta} \mathbb{E}_{D_{\text{train}}, D_{\text{val}}} \mathcal{L}(\mathcal{A}_\theta, D_{\text{train}}, D_{\text{val}})$$

(1)

where $\mathcal{L}(\mathcal{A}_\theta, D_{\text{train}}, D_{\text{val}})$ is a loss function that measures the loss of the model learned by algorithm $\mathcal{A}$ with hyperparameters $\theta$ that is trained with $D_{\text{train}}$ and is validated with $D_{\text{val}}$.

Global optimization techniques are commonly adopted to face the nonconvex nature of the hyperparameter optimization problem. For SVMs, the works on hyperparameter optimization can be categorized as follows.

1) **Model-free optimization** includes classical techniques, such as Grid Search [40]; Random Search [40]; EAs [41], [42]; and Particle Swarm Optimization [21].

2) **Model-based optimization** includes Bayesian optimization [43], [44] and surrogate-assisted optimization [27].

The works on SVMs hyperparameter optimization have also considered the optimization of the model pipeline (data preprocessing + learning algorithm) [42], [45], the training set selection problem [28], [46], [47], or more recently a combination of feature and training set selection together with hyperparameter optimization [48]. However, these studies neglected the imbalanced classification problem.

**C. Evolutionary Bilevel Optimization**

BLO is a hierarchical optimization problem with two levels: 1) the upper level, also known as the leader and 2) the lower level, also called the follower. Formally, a BLO problem is stated as follows [17], [18]:

$$\min f_u(v_u, v_l)$$

$$\text{s.t. } v_l \in \{ \arg \min_{v_l} \min f_l(v_u, v_l) : g_l(v_u, v_l) \leq 0 \}$$

$$g_u(v_u, v_l) \leq 0$$

(2)

where $v_u$ and $v_l$ are the upper level and the lower level variables, respectively, $f_u$ and $f_l$ represent the objective functions for the upper level and lower level, and $g_u$ and $g_l$ are the set of constraints for upper level and lower level, respectively.

The lower level is a constraint to the upper level optimization problem. Therefore, the lower level solution partially determines the upper level solution. The nested structure leads to several difficulties, such as nonlinearity, nonconvexity, and disconnectedness. These difficulties can be present even for the simplest bilevel problems [17], [18], [49].

EAs have shown success when dealing with complex optimization problems. Thus, EAs emerge as an alternative to deal with BLO problems. Most of the current EAs proposed to handle these problems are nested in nature. These approaches have two optimization algorithms, where one algorithm runs within the other. Overviews of evolutionary bilevel algorithms can be found in [17] and [18].

Supervised learning can be treated as a bilevel problem, in which the upper level optimizes the hyperparameters that minimize the expected generalization error, and the lower level learns the parameters [50]. Next, we explain the bilevel formulation for learning parameters and hyperparameters for an SVM with cost sensitive.

**III. BILEVEL COST-SENSITIVE SUPPORT VECTOR MACHINE: OPTIMIZATION PROBLEM**

The bilevel formulation breaks the problem down into two levels: 1) the upper level concerned with the hyperparameters configuration and 2) the lower level with the SVM training. In this section, we explain the optimization objectives at each level. First, Section III-A defines the lower level that finds the optimal separating hyperplane when training the SVM. Next, Section III-B explains the objective function for the upper level, which optimizes the classification performance considering the uneven class distributions for a given hyperparameter configuration.

**A. Lower Level—Optimizing Parameters**

The lower level problem focuses on finding the support vectors that define the hyperplane for an imbalanced classification problem. A cost-sensitive SVM penalizes the errors differently for positive and negative classes. This is formulated as follows [37]:

$$\min \frac{1}{2} \| \mathbf{w} \|^2 + C^+ \sum_{i,y_i=1} \xi_i + C^- \sum_{i,y_i=-1} \xi_i$$

subject to $y_i((\mathbf{w}, \mathbf{x}) + b) \geq 1 - \xi_i$

$$\xi_i \geq 0$$

(3)

where $\langle \cdot, \cdot \rangle$ represents the dot product, $\mathbf{w}$ is the separating hyperplane, and $C^+$ and $C^-$ are the costs for the positive and negative samples, respectively.

The dual problem obtained through Lagrange multipliers is as follows:

$$\max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

subject to $\sum_{i=1}^{n} y_i \alpha_i = 0$
For learning nonlinear functions, the dot product \(\langle x_i, x_j \rangle\) is replaced by a kernel function \(K(x_i, x_j)\). The radial basis function (RBF) is a kernel function that is highly effective and theoretically supported [51], [52]. The RBF kernel is defined as
\[
K(x_i, x_j) = e^{-\gamma \|x_i-x_j\|^2}
\]
where \(\gamma\) is an adjustable parameter given by the upper-level.

Section III-B explains the upper level optimization problem that optimizes the hyper parameters \(C^+, C^-,\) and the \(\gamma\) value for the RBF kernel.

B. Upper Level—Optimizing Hyperparameters

The upper level optimizes the hyperparameters used in the lower level to learn the support vectors. Let \(\lambda\) be a vector that encodes the hyperparameters \(C^+, C^-,\) and \(\gamma\). The goal of the upper level is to find the set of hyperparameters that gets the minimum generalization error on imbalanced classification problems, which is estimated using the balanced error rate (BER) score. The BER is defined as
\[
\text{BER}(\lambda, \alpha^*, b) = \frac{1}{2} \left( \frac{FN}{TP+FN} + \frac{FP}{FP+TN} \right)
\]
where \(TP\) and \(TN\) are, respectively, the number of positive and negative samples correctly classified; and \(FN\) and \(FP\) are, respectively, the number of positive and negative samples incorrectly classified.

Computing BER using the training set can lead to overfitting. \(K\)-fold cross-validation is commonly used to assess the expected performance and to reduce the risk of overfitting. However, this procedure can become computationally inefficient, as it implies solving the lower level problem \(K\) times for each configuration of hyperparameters \(\lambda\). We face that disadvantage by approximating the bound on the leave-one-out cross-validation for the upper level. Based on the lower level solution \(\alpha^*\), the predicted value for the \(j\)th training sample is given by
\[
f(x_j) = \sum_{i=1}^{n} \alpha_i^* y_i K(x_j, x_i) + b
\]
where \(b\) is the bias term and is set to satisfy the Karush–Kuhn–Tucker condition.

Eliminating a nonsupport vector from the training set does not affect the model; therefore, we focus on support vectors, as they can contribute to the error. Assuming that the set of support vectors remains the same during the leave-one-out procedure, we can approximate the output when removing the \(j\)th support vector from the training set as
\[
\tilde{f}(x_j) = \sum_{i=1}^{n} \alpha_i^* y_i K(x_j, x_i) + b - \alpha_j^* y_j K(x_j, x_j).
\]

In the case of the RBF kernel, the term \(K(x_j, x_j)\) equals one. Simplifying (8) and multiplying it by \(y_j\), an instance \(x_j\) is incorrectly classified if
\[
y_j \left( \sum_{i=1}^{n} \alpha_i^* y_i K(x_j, x_i) + b - \alpha_j^* y_j \right) < 0.
\]

After determining the incorrectly classified instances with (9), the BER of each individual at the upper level is determined using (6).

After solving the BLO problem, the optimal support vectors are used to classify a new instance using (7).

IV. EBCS-SVM: EVOLUTIONARY BILEVEL COST-SENSITIVE SUPPORT VECTOR MACHINES

EBCS-SVM aims to build an optimal SVM model for handling imbalanced classification problems through BLO. EBCS-SVM determines the hyperparameters values, that is, the costs of each class and the kernel parameters that minimize BER at the upper level, and the lower level finds the optimal separating hyperplane. Fig. 1 graphically depicts the bilevel interaction between the lower level and upper level in EBCS-SVM.

Algorithm 1 describes EBCS-SVM, and it works as follows:
1) In line 1, a population for the upper level is randomly created with \(n_u\) individuals and three variables representing the costs for the positive and the negative class, and the \(\gamma\) value for RBF kernel, based on the bounds given in Section IV-B.
2) In lines 2–5, each upper level solution is evaluated. To this end, the following steps are carried out.
   a) The lower level problem described in Section III-A is solved using the SMO algorithm and the given hyperparameters.
   b) The resulting SVM model is evaluated by computing the BER, as described in Section III-B.
3) In lines 6–14, the evolutionary process takes place.
   a) In line 7, a new population of hyperparameters is created by applying evolutionary operators over $P_u$ using the adaptation of DE operators proposed by SHADE [53].
   b) In line 9, the nearest neighbors in $P_u$ are found for each individual in $O_u$. In line 10, the set of support vectors for the lower level is warm started by considering the nearest neighbors of the upper level to determine how probable an instance is a support vector, as described in Section IV-A.
   c) Line 11 solves the lower level optimization problem with the SMO algorithm, and line 12 computes the BER score for the given hyperparameters.

A. Lower Level Initialization

When solving the lower level using the hyperparameter of the initial upper level population, all training instances are initially assumed to be a support vector, and the SMO solves the lower level problem. The solutions obtained by the initial configuration of hyperparameters are stored to perform a warm starting of the lower level.

The warm starting takes place after the first generation of the upper level and works as follows. First, for a new given hyperparameters configuration, its $m$ nearest neighbors are found among the hyperparameters from the current population. After, the set of support vectors is retrieved and used to determine the chance of an instance becoming a support vector, based on the relative frequency of its $m$ nearest neighbors. Finally, the SMO algorithm solves the lower level based on previous initialization. The premise for using this initialization is that similar configurations of hyperparameters lead to a similar set of support vectors.

The value of $m$ is determined during the search. For doing so, the number of neighbors is randomly selected between $[1, u_s]$ with uniform probability. Then, the probability of each value of $m$ is updated based on the normalized frequency of success of such value, that is, an improvement in the BER score.

B. Representation and Evolutionary Operators

At the upper level, individuals encode the hyperparameters as a real-value vector. Initially, $C^+$ and $C^-$ are randomly generated in the range $[2^{-5}, 2^{10}]$ and $\gamma$ between $[2^{-10}, 2^5]$. Individuals at this level are then evolved using DE [54], [55]. However, the classical DE requires the definition of two parameters: 1) the differential weights ($F$) and 2) the crossover rate ($CR$). For this reason, we adopt a self-adaptative variant called SHADE [53], which uses history information to adapt the values of $F$ and $CR$ during the search and a greedy current-to-best mutation strategy. Therefore, these parameters are learned during the search.

After a child solution of hyperparameters is created and its BER value is computed, it competes with the current individual to determine which one is better and therefore, survives. A child solution is better if its BER value is lower than the BER value of the current individual, or if the BER values of both solutions are equivalent, but the child has fewer support vectors.

V. EXPERIMENTAL SETTINGs

In this section, we describe the configuration setup for our experimental study. In Section V-A, we present the set of benchmark datasets considered for experimentation. Section V-B provides the metrics used to assess the performance of all methods and the test to analyze them. Finally, Section V-C details the state-of-the-art techniques used
to compare the performance and their respective tuning of hyperparameters.

A. Datasets

The performance of EBCS-SVM is assessed using a benchmark of 70 datasets from the KEEL repository [56]. Table I details the characteristics of the datasets, including the number of instances (Inst.), the number of features (Feat.), and the imbalance ratio (IR). These datasets are diverse in the IR, number of samples, and number of features. Thus, the performance is assessed using problems with different characteristics. We divided the datasets into three groups based on the degree of IR: 1) ten datasets with small IR, that is, the IR is less than or equal to three; 2) 35 datasets with medium IR, that is an IR higher than three and less than or equal to 20; and 3) 25 datasets with high IR, that is, the IR is higher than 20.

Datasets were partitioned using the 10 × 5-fold cross-validation. In the 5-fold cross-validation, the dataset is randomly split into five disjoint subsets. In each fold, a subset is used as the test set and the remaining as the training set. Then, 5-fold cross-validation is repeated ten times, with a different split each time. Thus, each dataset is tested 50 times.

B. Performance Metrics and Statistical Tests

We assessed the performance of the methods using a set of metrics well suited for imbalanced classification problems [57]. Let sensitivity (\(sen\)) and specificity (\(spe\)) be defined as

\[
\begin{align*}
\text{sen} & = \frac{TP}{TP + FN} \\
\text{spe} & = \frac{TN}{TN + FP}
\end{align*}
\]

where \(TP\) and \(TN\) are, respectively, the number of positive and negative samples correctly classified; and \(FN\) and \(FP\) are, respectively, the number of positive and negative samples incorrectly classified.

The metrics described in [57] can be defined in terms of specificity and sensitivity. These metrics are listed as follows.

1) \(BAR\), the balanced accuracy rate is equivalent to 1-BER and is defined as the average between sensitivity and specificity, that is

\[
\text{BAR} = \frac{1}{2}(sen + spe).
\]

2) \(BMI\), the bookmarker informedness is defined as the sum of sensitivity and specificity minus one, that is

\[
\text{BMI} = sen + spe - 1.
\]

3) \(GM\) indicates the geometric mean between sensitivity and specificity, that is

\[
\text{GM} = \sqrt{sen \cdot spe}.
\]

\[\text{IR} \text{ is defined as the ratio between the number of samples in the majority class and the number of samples in the minority. Therefore, the higher the IR, the greater the imbalance.}\]

4) \(uF1\) [57] is the unbiased version of the F1 score and is defined as two times the sensitivity between the sum of two plus sensitivity minus specificity.
that is,

\[ uF1 = \frac{2 \cdot sen}{2 + sen - spe}. \] (14)

5) \(uMCC\) [57] is the unbiased version of the Matthews correlation coefficient, and is defined as follows:

\[ uMCC = \frac{sen \cdot spe - 1}{\sqrt{1 - (sen - spe)^2}}. \] (15)

The Bayesian hypothesis tests are used to analyze EBCS-SVM regarding reference methods. They allow comparing the difference in the results achieved by two algorithms, estimating the posterior probabilities that one algorithm is better than the other and that both are practically equivalent. These tests are not affected by the number of datasets. Moreover, they can provide more information than the null hypothesis significance test, even when the latter does not reject the null hypothesis [58], [59]. We used the hierarchical Bayesian test to analyze the results, as it considers both the mean and the variance through the cross-validation partitions for each dataset. In the analysis, we considered that two methods are equivalent if the difference is below 0.01.

C. Reference Methods

We compared EBCS-SVM with several state-of-the-art techniques for imbalanced classification. To this end, we considered methods for both DL preprocessing and AL. Specifically, the comparative study considers the following methods.

1) SVM: The standard SVM (BL) is used on the dataset without preprocessing or modification to weight the class distributions.

2) DL Methods: The methods in this group are used to preprocess the data. Then, the edited dataset is used to train an SVM. This group consists of ROS [3], SMOTE [6], SVM-SMOTE [7], RUS [3], and CNN [9].

3) AL Methods: This group encompasses SVMDC [37], uNBSVM [14], WK-SMOTE [15], CSSVM [60], and RBI-LP-SVM [16].

Reference methods require defining a set of hyperparameters to use in training. Properly selecting hyperparameters is a crucial step to compare classification algorithms [51], [61]. We adopted the RBF kernel because of its effectiveness with SVM. For the sake of a fair comparison, the hyperparameters of each method were optimized for each dataset independently by optimizing the BER, computed through an internal stratified 5-fold cross-validation on the training set. The set of hyperparameters includes: the \(\gamma\) value optimized in the range of \([2^{-10}, 2^5]\) for all methods; the regularization parameter \(C\) in the range of \([2^{-5}, 2^{10}]\) for SVM, DL methods, and CSSVM; the regularization parameter for positive \((C^+)\) and negative \((C^-)\) class optimized between \([2^{-5}, 2^{10}]\) for both WK-SMOTE and SVMDC; the regularization parameter for synthetic samples \((C^s)\) optimized between \([2^{-5}, 2^{10}]\) for WK-SMOTE; the weight factor for positive \((\omega_p)\) and negative \((\omega_n)\) class in the range of \([0, 1]\) for uNBSVM; the cost-sensitive parameter \((\kappa)\) ranges from zero to one and the margin violation weight \((C_1)\) is the range of \([1, 10]\) for CSSVM; the amount of sampling is searched in the range of \([(l_m + 1)/n_M], 1]\), with \(n_m\) and \(n_M\) as the number of samples in the minority and majority class, respectively, for DL methods and WK-SMOTE, and the number of neighbors is between \([1, (n_m/2)]\) for SMOTE, SVM-SMOTE, and WK-SMOTE. SHADE was also used to optimize the hyperparameters, which ran with a population size that is equal to 30 and the stopping criteria considered performing 1000 fitness function evaluations or that the standard deviation of fitness values of the population is below 0.001.

We provide the implementation of EBCS-SVM, datasets, splits, and detailed results in each partition as the supplementary material. The supplementary material can be downloaded at www.cimat.mx/alejandro.rosales/resources/EBCSSVM.tar.gz.

VI. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the experimental results reported by EBCS-SVM and reference methods. Section VI-A shows the results obtained with datasets with small IR. Next, Section VI-B considers the 35 datasets with medium IR, and Section VI-C presents the results using the 25 datasets with high IR. Finally, Section VI-D compares the training time required by each method.

A. Experiments on Small Imbalance Ratio

In this section, we focused on comparing the performance of EBCS-SVM against reference methods. Table II shows the results on the ten datasets with a small IR. The reported results correspond to the average and standard deviations for BAR, BMI, GM, uF1, and uMCC scores. Fig. 2 graphically depicts...
barycenter plots for the posterior probabilities reported by the hierarchical Bayesian test for the BAR score. Each point on the barycenter plot represents an estimate of the probability that it belongs to each region. Based on the results and the analysis carried out by the hierarchical Bayesian test, the following observations are highlighted.

1) Most methods showed a competitive performance when dealing with small IR.

2) For metrics BAR, GM, and uF1, most methods reported scores above 0.820. The exceptions were WK-SMOTE and RBI-LP-SVM, which obtained performances below 0.800. On the other hand, for metrics BMI and MCC, most methods reported scores above 0.710, except for WK-SMOTE, uNBSVM, and RBI-LP-SVM.

3) CSSVM obtained the highest performance in all metrics. SVMDC was the second-best position for BAR, BMI, GM, and uMCC, and the third-best for uF1. On the other hand, EBCS-SVM ranked second-best for uF1; it was in the seventh position for BAR, BMI, and uMCC; and in the eighth position for GM.

4) The hierarchical Bayesian tests provided strong evidence on the practical equivalence between CSSVM and EBC-SVM. We can observe similar behavior when EBCS-SVM is compared with SVM, ROS, SMOTE, SVMSMOTE, RUS, CNN, and SVMDC. Thus, EBCS-SVM exhibited a performance practically similar to that of CSSVM, the best-ranked method.

5) Among DL methods, ROS had the best average performance over the ten datasets with a small IR for BAR, BMI, GM, and uMCC, while SVMSMOTE was the worst for uF1.

6) RBI-LP-SVM showed the lowest performance among AL methods. The hierarchical Bayesian test reported probabilities above 0.999 in the region of EBCS-SVM for all metrics.

7) AL methods generally reported better performance than DL methods.

In the next section, we delved into our analysis when the IR increases.

B. Experiments With Medium Imbalance Ratio

In this section, we analyzed the performance of EBCS-SVM and reference methods when using the 35 benchmark datasets with medium IR. Table III presents the average results obtained by each method, and Fig. 3 shows the posterior plots for the BAR score when EBCS-SVM is compared with reference methods. From these, the following is pointed out.

1) Most methods reported results above 0.800 for BAR, GM, and uF1 scores, except for SVM for GM and uF1; WK-SMOTE for BAR, GM, and uF1; uNBSVM for GM and uF1; and RBI-LP-SVM for BAR, GM, and uF1.

2) Regarding BMI and uMCC, most methods reported performances above 0.700. The exceptions were SVM for BMI and WK-SMOTE, uNBSVM, and RBI-LP-SVM for BMI and uMCC.

3) ROS achieved the highest performance for BAR, BMI, GM, and uMCC and the second-best performance for uF1. RUS achieved the best performance for uF1 and
TABLE III

OBTAINED RESULTS ON MEDIUM IR DATASETS FOR BAR, BMI, GM, UF1, AND uMCC S CORES

| Fam. | Method   | BAR       | BMI       | GM         | UF1        | uMCC        |
|------|----------|-----------|-----------|------------|------------|-------------|
| DL   | SVM      | 0.8420 ± 0.1226 | 0.6841 ± 0.2453 | 0.7931 ± 0.1929 | 0.7724 ± 0.2107 | 0.7090 ± 0.2314 |
| DL   | ROS      | 0.8832 ± 0.0728 | 0.7664 ± 0.1457 | 0.8693 ± 0.0808 | 0.8579 ± 0.0892 | 0.7813 ± 0.1395 |
| DL   | SMOTE    | 0.8738 ± 0.0974 | 0.7475 ± 0.1948 | 0.8455 ± 0.1656 | 0.8345 ± 0.1687 | 0.7613 ± 0.1921 |
| DL   | SVMSMOTE | 0.8678 ± 0.0942 | 0.7356 ± 0.1884 | 0.8329 ± 0.1549 | 0.8216 ± 0.1599 | 0.7496 ± 0.1845 |
| DL   | RUS      | 0.8793 ± 0.0858 | 0.7586 ± 0.1716 | 0.8652 ± 0.1025 | 0.8603 ± 0.1073 | 0.7690 ± 0.1672 |
| DL   | CNN      | 0.8626 ± 0.0974 | 0.7252 ± 0.1948 | 0.8388 ± 0.1262 | 0.8254 ± 0.1393 | 0.7495 ± 0.1894 |
| AL   | SVMDMC   | 0.8635 ± 0.0808 | 0.7271 ± 0.1616 | 0.8334 ± 0.1181 | 0.8198 ± 0.1244 | 0.7554 ± 0.1558 |
| AL   | CSSVM    | 0.8689 ± 0.0795 | 0.7378 ± 0.1590 | 0.8435 ± 0.1141 | 0.8307 ± 0.1207 | 0.7548 ± 0.1536 |
| AL   | WK-SMOTE | 0.7651 ± 0.1131 | 0.5302 ± 0.2263 | 0.6523 ± 0.2231 | 0.6250 ± 0.2274 | 0.5646 ± 0.2269 |
| AL   | uNBSVM   | 0.8315 ± 0.0856 | 0.6629 ± 0.1712 | 0.7927 ± 0.1216 | 0.7884 ± 0.1253 | 0.6852 ± 0.1664 |
| AL   | RBI-LP-SVM | 0.5427 ± 0.1139 | 0.0855 ± 0.2278 | 0.0923 ± 0.2294 | 0.6926 ± 0.0752 | 0.0870 ± 0.2279 |
| AL   | EBCS-SVM | 0.8784 ± 0.1051 | 0.7569 ± 0.2102 | 0.8581 ± 0.1440 | 0.8722 ± 0.1502 | 0.7671 ± 0.2052 |

![Fig. 3. Posterior distribution for BAR when comparing EBCS-SVM with the reference methods on medium IR datasets. The region of the bottom-left represents EBCS-SVM, the region at the top is for ROS, and the region in the bottom-right represents the reference method.](image)

Fig. 3. Posterior distribution for BAR when comparing EBCS-SVM with the reference methods on medium IR datasets. The region of the bottom-left represents EBCS-SVM, the region at the top is for ROS, and the region in the bottom-right represents the reference method. (a) SVM–EBCS-SVM. (b) ROS–EBCS-SVM. (c) SMOTE–EBCS-SVM. (d) SVMSMOTE–EBCS-SVM. (e) RUS–EBCS-SVM. (f) CNN–EBCS-SVM. (g) SVMDC–EBCS-SVM. (h) CSSVM–EBCS-SVM. (i) uNBSVM. (j) WK-SMOTE–EBCS-SVM. (k) RBI-LP-SVM–EBCS-SVM.

the second-best performance for BAR, BMI, GM, and MCC. EBCS-SVM ranked the third-best for all metrics.

4) The hierarchical Bayesian analysis revealed a probability above 0.900 in the region of practical equivalence when EBCS-SVM was compared with ROS, SMOTE, SVMSMOTE, and RUS for all metrics. Thus, there is evidence in favor of the competitiveness of these methods for handling medium IR problems. We can observe this behavior in Fig. 3, where we can note that the center mass is in the region of practical equivalence. Although for CNN and SVMDC, the center mass fell in the practically equivalent region, the posterior plot distribution was spread throughout the area of EBCS-SVM, providing at some extent evidence in favor of EBCS-SVM.

5) ROS stood out as the most effective DL method.

6) Among AL methods, EBCS-SVM obtained the best performance, and CSSVM was the second best. Conversely, RBI-LP-SVM ranked in the last position.

For datasets with a medium IR, EBCS-SVM, ROS, and RUS were the most superior methods. ROS is highlighted as a prominent method for problems with medium IR. CSSVM remained a competitive method. In the next section, methods are evaluated when handling datasets with a high IR.

C. Experiments With High Imbalance Ratio

In this section, we considered the 25 datasets with an IR above 20 to analyze the performance of EBCS-SVM and reference methods. Table IV reports the average results for all metrics, and Fig. 4 shows the posterior probabilities obtained with the hierarchical Bayesian test. Based on these results, we remark the following.

1) EBCS-SVM excelled in dealing effectively with highly imbalanced datasets in all metrics. Ergo, the superiority of EBCS-SVM, is stressed when the IR is increased.

2) Among DL methods, RUS had the best score in all metrics and ranked second-best position globally. ROS
was the second-best DL method and ranked third-best position globally.

3) Among AL methods, EBCS-SVM was the best one, followed by SVMDC. However, when observing the posterior probabilities reported by the hierarchical Bayesian test, we noted that the center mass is in the region of EBCS-SVM. Furthermore, the posterior odds revealed strong evidence in favor of EBCS-SVM. We observed this behavior for all metrics.

4) Posterior probabilities plots also showed that for most reference methods, the center mass fell in the region of EBCS-SVM. The exception was RUS, whose center mass is in the region of practical equivalence; however, its distribution spreads in the regions of both EBCS-SVM and RUS.

As the IR increased, the performance of EBCS-SVM stood out over the reference methods, regardless of the adopted metric. The hierarchical Bayesian test reported high probabilities in favor of EBCS-SVM that supported these observations.

D. Computational Time

In this section, we analyze the training time for each method. We used the performance profile [62], which represents the cumulative distribution on a performance metric. The performance profile is constructed for the necessary training time required by each method to optimize the hyperparameters and learn the classification model. Fig. 5 depicts the performance profile for all methods. The $y$-axis represents the probability ($\rho(\tau)$) that a method can learn a model within a factor $\tau$ times the fastest method, and the $x$-axis represents the $\tau$ factor. Thus, $\rho(1)$ indicates the probability where a given method achieves the lowest training time among all methods.

From Fig. 5, we can observe that both RUS and EBCS-SVM exhibited the best training times. From the value of $\rho(1)$, it is observed that EBCS-SVM had the highest performance time.
probability (0.58) of being the fastest one, while RUS had the second-highest probability (0.30). Furthermore, AL methods generally required less training time than DL methods, which can be observed in the near-zero probability of oversampling techniques with a value of τ equals one. The outstanding performance of RUS in training time is because by removing training instances, the SVM algorithm works with a reduced number of samples and reduces the computational time. Although CNN is also an undersampling method, the way in how instances to remove are selected slows down the training time. On the other hand, EBCS-SVM exploits the information of both levels to feedback with the information of previous support vectors of similar solutions and improving the convergence in solving the lower level. Finally, RBI-LP-SVM showed the worst performance profile.

VII. CONCLUSION

This article introduced EBCS-SVM for learning an SVM in imbalanced scenarios. EBCS-SVM formulated the optimization of hyperparameters and support vectors as a BLO problem and showed to be able to handle imbalanced classification problems effectively and freed practitioners from defining the optimal cost to each class. To this end, an EA at the upper level and the SMO at the lower level interplay, such that lower level solutions impact the BER of the upper level solutions, and previous hyperparameters help initialize the set of support vectors. Thus, there is a dual enrichment in the two levels.

The efficacy of EBCS-SVM was assessed using 70 benchmark datasets and compared with those of state-of-the-art techniques. Experimental results revealed a leading performance of EBCS-SVM. The traditional SVM was unable to deal effectively with imbalanced datasets. On the other hand, DL methods showed excellent performance, although in most cases required larger training times than AL methods. SVMSMOTE and CNN were the worst among DL methods. The low performance of both can be because the implicit mapping of the kernel function is not taken into account to select boundary instances when sampling. EBCS-SVM considered the particularities of SVM to learn a model.

The most competitive method was RUS, a DL technique that randomly subsamples the majority class. Since it does not require further information, RUS is fast. However, this method was outperformed by EBCS-SVM as the imbalanced ratio increased and EBCS-SVM required lower training time. Moreover, EBCS-SVM employed a self-adapted EA to adjust the evolutionary parameters during the learning. Therefore, EBCS-SVM is accurate and does not require fine-tuning of the SVM hyperparameters.

REFERENCES

[1] C. Cortes and V. Vapnik, “Support-vector networks,” Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.
[2] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, Learning From Imbalanced Data Sets. Cham, Switzerland: Springer, 2018.
[3] H. He and E. A. Garcia, “Learning from imbalanced data,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[4] B. Krawczyk, “Learning from imbalanced data: Open challenges and future directions,” Prog. Artif. Int., vol. 5, no. 4, pp. 221–232, 2016.
[5] Z. Zhu, Z. Wang, D. Li, Y. Zhu, and W. Du, “Geometric structural ensemble learning for imbalanced problems,” IEEE Trans. Cybern., vol. 50, no. 4, pp. 1617–1629, Apr. 2020.
[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” J. Artif. Int. Res., vol. 16, no. 1, pp. 321–357, 2002.
[7] H. M. Nguyen, E. W. Cooper, and K. Kamei, “Borderline over-sampling for imbalanced data classification,” Int. J. Knowl. Eng. Soft Data Paradigm., vol. 3, no. 1, pp. 4–21, 2011.
[8] H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” in Proc. IEEE IJCNN, 2008, pp. 1322–1328.
[9] P. Hart, “The condensed nearest neighbor rule (Corresp.),” IEEE Trans. Inf. Theory, vol. 14, no. 3, pp. 515–516, May 1968.
[10] Q. Kang, X. Chen, S. Li, and M. Zhou, “A noise-filtered under-sampling scheme for imbalanced classification,” IEEE Trans. Cybern., vol. 47, no. 12, pp. 4263–4274, Dec. 2017.
[11] S. Xia, Y. Zheng, G. Wang, P. He, H. Li, and Z. Chen, “Random space division sampling for label-noisy classification or imbalanced classification,” IEEE Trans. Cybern., early access, Apr. 28, 2021, doi: 10.1109/TCYB.2021.3070005.
[59] G. Corani, A. Benavoli, J. Demšar, F. Mangili, and M. Zaffalon, “Statistical comparison of classifiers through bayesian hierarchical modelling,” Mach. Learn., vol. 106, no. 11, pp. 1817–1837, 2017.

[60] A. Iranmehr, H. Masnadi-Shirazi, and N. Vasconcelos, “Cost-sensitive support vector machines,” Neurocomputing, vol. 343, pp. 50–64, May 2019.

[61] J. Wainer and G. Cawley, “Empirical evaluation of resampling procedures for optimising SVM hyperparameters,” J. Mach. Learn. Res., vol. 18, no. 15, pp. 1–35, 2017.

[62] E. D. Dolan and J. J. Moré, “Benchmarking optimization software with performance profiles,” Math. Program., vol. 91, no. 2, pp. 201–213, Jan. 2002.

Alejandro Rosales-Pérez received the B.S. degree in electronic engineering from the Instituto Tecnológico de Tuxtla Gutiérrez, Chiapas, Mexico, in 2008, and the M.Sc. and Ph.D. degrees in computer science from INAOE, Puebla, Mexico, in 2011 and 2016, respectively.

He is currently with the Centro de Investigación en Matemáticas, Monterrey, Mexico. His doctoral thesis was awarded by the National Association of Education Institutions in Information Technology in 2016 and also by the Mexican Society on Artificial Intelligence in 2016. He is a member of the Mexican National System of Researchers. His current research interests mainly include evolutionary computation and machine learning.

Salvador García received the B.S. and Ph.D. degrees in computer science from the University of Granada, Granada, Spain, in 2004 and 2008, respectively.

He is currently a Full Professor with the Department of Computer Science and Artificial Intelligence, University of Granada. He has published more than 100 papers in international journals (more than 70 in Q1), H-index 54. His research interests include data science, data preprocessing, big data, evolutionary learning, deep learning, and metaheuristics.

Dr. García is an Associate Editor-in-Chief of Information Fusion (Elsevier) and an Associate Editor of Swarm and Evolutionary Computation (Elsevier) and AI Communications (IOS Press). He belongs to the list of the Highly Cited Researchers in the area of Computer Sciences from 2014 to 2020: http://highlycited.com/ (Clarivate Analytics).

Francisco Herrera (Senior Member, IEEE) received the M.Sc. degree in mathematics and the Ph.D. degree in mathematics from the University of Granada and the Director of DaSCI Institute (Andalusian Research Institute in Data Science and Computational Intelligence). He has been the Supervisor of more than 50 Ph.D. students. He has published more than 500 journal papers, receiving more than 98 000 citations (Google Scholar, H-index 152). He is the coauthor of several books.

His current research interests include computational intelligence (including fuzzy modeling, computing with words, evolutionary algorithms, and deep learning), information fusion and decision making, and data science (including data preprocessing, prediction, and big data).

Prof. Herrera is the Editor-in-Chief of Information Fusion (Elsevier). He has been selected as a Highly Cited Researcher (in the fields of Computer Science and Engineering, since 2014, Clarivate Analytics).