GSVM: An SVM for handling imbalanced accuracy between classes in bi-classification problems

L. Gonzalez-Abril\textsuperscript{a}, H. Nuñez\textsuperscript{b}, C. Angulo\textsuperscript{c}, F. Velasco\textsuperscript{a}

\textsuperscript{a}COSDE Group. Applied Economics I Dept., Seville University, 41018 Seville, Spain
\textsuperscript{b}AI Lab. Sciences Faculty, Central University of Venezuela, 1020-A Caracas, Venezuela
\textsuperscript{c}GREC Research Group. Technical University of Catalonia, 08028 Barcelona, Spain

Abstract
A new Support Vector Machine, SVM, is introduced, called GSVM, which is specially designed for bi-classification problems where balanced accuracy between classes is the objective. Starting from a standard SVM, the GSVM is obtained from a low-cost post-processing strategy by modifying the initial bias. Thus, the bias for GSVM is calculated by moving the original bias in the SVM to improve the geometric mean between the true positive rate and the true negative rate. The proposed solution neither modifies the original optimization problem for SVM training, nor introduces new hyper-parameters. Experimentation carried out on a high number of databases (23) shows GSVM obtaining the desired balanced accuracy between classes. Furthermore, its performance improves well-known cost-sensitive schemes for SVM, without adding complexity or computational cost.

Key words: support vector machines, cost-sensitive schemes, classification, imbalanced databases, imbalanced accuracy

1. Introduction
Accuracy is the usual measure to be maximized when a bi-classification problem is solved. Nevertheless, there exist problems where this solution is obtained with imbalanced accuracy: one class reaches much higher accuracy than the other class. This situation is common when learning from imbalanced datasets [26, 34]. In bi-classification, it occurs when datasets provided for the learning algorithms contain many examples for one class (majority class), but very few for the other class (minority class). It is assumed that negative training instances are the majority and positive training instances
are the minority. In this context, a model of high and balanced accuracy is difficult to be generated using traditional classification techniques [24]. Objective functions employed for the determination of the classifiers typically tend to favour the larger class [20, 30], that is, the negative training examples overwhelmingly outnumber the positive ones making the classifier training to be imbalanced.

It is worth noting that the degree of imbalance between classes is not the only factor hindering learning. Other factors contributing to classification deterioration are the presence of small disjuncts, the lack of density in the training data, the overlapping between classes, the identification of noisy data, the significance of the border-line instances, and the dataset shift between the training and the test distributions [30]. As it turns out, dataset complexity is the primary determining factor of classification deterioration, which, in turn, is amplified by the addition of a relative imbalance [20].

Henceforth, let us consider that the problem of imbalance between the true positive rate and the true negative rate is addressed on imbalanced datasets, however keeping in mind that the problem of imbalanced accuracy is not exclusive of these datasets.

The use of SVMs is an attractive option for handling these situations since their learning mechanism usually considers a small subset of patterns to build the classification model. SVMs are learning machines which implement the structural risk-minimization inductive principle to obtain generalization on a limited number of learning patterns [36]. This theory was developed on the basis of a separable binary classification problem where the optimization criterion is the width of the margin between the positive and negative examples which provides a good generalization performance, as shown in a number of applications [9]. However, like other learning machines that build classification models, SVMs aim to minimize the error on the entire dataset, and hence they are inherently biased towards the majority class when classifying imbalanced datasets [35].

In order to improve the performance of SVMs on imbalanced datasets, a number of solutions have been proposed. Some of them, such as techniques for dataset sampling balance in a pre-processing stage, are of general application; other, more specific solutions, take into account particular features, for instance those based in cost-sensitive learning.

The proposed approach takes into account that it has been empirically shown [21, 29, 43] that the hyperplane learned by an SVM in the presence of imbalanced sets has approximately the same orientation as the ideal hyper-
plane. Hence, reduced generalization on the minority class would be indeed associated with the bias, as positive instances lie far from this ideal limit, i.e., the SVM learns a boundary that is much too close to this class. Modifications should be based on the distribution of classes in the dataset, which also do not directly affect standard SVM training.

Thus, a post-processing strategy, based on the consideration that the bias is a parameter to be tuned, is proposed in this paper. The function learned by a standard SVM will be adjusted through the bias to improve its performance, measured in terms of the geometric mean of recall and specificity. The proposed solution introduces no new parameters. Furthermore, it neither requires modifying the standard optimization problem for SVM training, nor additional steps, nor re-training.

The remainder of this paper is organized as follows: Section 2 introduces several metrics commonly employed in classification problems. The standard SVM approach is outlined and an overview of strategies to improve performance in problems with imbalanced datasets is provided in Section 3. In Section 4, a family of classifiers which depend on the bias is examined and certain results are obtained on the metrics considered. Furthermore, a new SVM, called GSVM, is introduced based on the geometric mean. Experimentation is carried out in Section 5 in order to compare the performance of the GSVM with that of the standard SVM. Its performance is also compared with well-known cost-sensitive schemes. Finally, conclusions are drawn.

2. The metrics

Let \( \mathcal{Z} = \{(x_1, y_1), \ldots, (x_N, y_N)\} \) be a training set, with \( x_i \in \mathcal{X} \) as the input space, \( y_i \in \mathcal{Y} = \{+1, -1\} \) the output space, and \( z_i = (x_i, y_i) \). Let \( f(x) \) be a binary classifier such that outputs are obtained as \( h(x) = \text{sign}(f(x)) \), where \( \text{sign}(\cdot) \) is the signed function.

Let \( \mathcal{Z}_{\text{pos}} = \{z_i \in \mathcal{Z} | y_i = +1\} \neq \emptyset \) and \( \mathcal{Z}_{\text{neg}} = \{z_i \in \mathcal{Z} | y_i = -1\} \neq \emptyset \) be the sets of training patterns for the positive and the negative class, respectively. Let \( N_{\text{pos}} = \#\mathcal{Z}_{\text{pos}} \) and \( N_{\text{neg}} = \#\mathcal{Z}_{\text{neg}} \) be the number of positive and negative instances, respectively, hence \( N = N_{\text{pos}} + N_{\text{neg}} \).

The most commonly used metric for the evaluation of the generalization performance of the classifier \( f(x) \) on a test set \( \mathcal{D} \subset \mathcal{X} \times \mathcal{Y} \) is the Accuracy, which is defined, from the confusion matrix (see Table 1), as follows:

\[
\text{Accuracy}(f, \mathcal{D}) = \frac{t_{\text{pos}} + t_{\text{neg}}}{N_{\text{pos}} + N_{\text{neg}}}
\]
Table 1: The confusion matrix for a classifier $f$ on a test set $\mathcal{D}$.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
<th>Positive ($N_{pos}$)</th>
<th>Negative ($N_{neg}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>True Positive ($t_{pos}$)</td>
<td>False Positive ($f_{pos}$)</td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td>False Negative ($f_{neg}$)</td>
<td>True Negative ($t_{neg}$)</td>
<td></td>
</tr>
</tbody>
</table>

The *Accuracy* is inappropriate when prior probabilities of classes differ greatly, since metrics fail to consider costs for wrong classifications, and thus remain very sensitive to the bias between classes [20].

Hence, other measures of assessment must be considered when the prior probabilities of classes differ so much. By considering accuracy rates on $Z_{pos}$ and $Z_{neg}$ separately, the *Recall* (or True Positive Rate) and *Specificity* (or True Negative Rate) metrics are defined as follows:

\[
Recall(f, \mathcal{D}) \equiv Rc(f, \mathcal{D}) = \frac{t_{pos}}{N_{pos}} \quad (1)
\]

\[
Specificity(f, \mathcal{D}) \equiv Sp(f, \mathcal{D}) = \frac{t_{neg}}{N_{neg}} \quad (2)
\]

From these metrics, an expression of *Accuracy* can be given:

\[
Accuracy(f, \mathcal{D}) = \frac{N_{pos} \cdot Rc(f, \mathcal{D}) + N_{neg} \cdot Sp(f, \mathcal{D})}{N_{pos} + N_{neg}} \quad (3)
\]

Hence, for imbalanced classes, $N_{pos} \ll N_{neg}$, Eq. 3 shows that *Accuracy* and *Specificity* are very similar, that is, the *Recall* metric barely alters the value of the *Accuracy* metric.

A different metric should be considered when a more balanced ratio between *Recall* and *Specificity* metrics is desired. A natural option would be to consider the *Arithmetic Mean*,

\[
Am(f, \mathcal{D}) = \frac{1}{2} (Recall(f, \mathcal{D}) + Specificity(f, \mathcal{D}))
\]

however it is well-known that the higher values in this measure compensate for the lower values.
A measure which is more sensitive to the low positive values and that combines \textit{Recall} and \textit{Specificity} is the \textit{Geometric Mean} [3], denoted by $Gm$, that is,

$$Gm(f, D) = \sqrt{Rc(f, D) \cdot Sp(f, D)} = \sqrt{\frac{t_{pos} \cdot t_{neg}}{N_{pos} \cdot N_{neg}}}$$

where both classes, positive and negative ones, have the same relevance for the purpose of classification.

3. SVMs on imbalanced datasets

Let $\mathcal{Z} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ be a training set, with $x_i \in \mathcal{X}$ as the input space, $y_i \in \mathcal{Y} = \{+1, -1\}$ the output space, and $z_i = (x_i, y_i)$. Let $\phi: \mathcal{X} \rightarrow \mathcal{F}$, $x = \phi(x)$, be a feature mapping with a dot product denoted by $\langle \cdot, \cdot \rangle$. A binary linear classifier $f: \mathcal{X} \rightarrow \mathbb{R}$, defined as $f(x) = \langle x, w \rangle - b$, is sought where $w \in \mathcal{F}$, $b \in \mathbb{R}$. Outputs are obtained as $h(x) = \text{sign}(f(x))$.

The standard 2-norm $C$-SVM primal formulation leads to the optimization problem [36]:

$$\min_{w \in \mathcal{F}, b \in \mathbb{R}} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \xi_i$$

s.t. \[
\begin{align*}
y_i \left( \langle x_i, w \rangle - b \right) + \xi_i & \geq 1, \\
\xi_i & \geq 0, \\
i & = 1 \ldots N
\end{align*}
\]

where $w$ is the vector perpendicular to the hyperplane, which defines the orientation of the classifier, and bias $b$ determines its position. Slack variables $\xi_i$ measure the error on the instances which violate constraints $y_i \cdot (\langle x_i, w \rangle - b) \geq 1$. The user-defined regularization parameter $C$ determines the tradeoff between maximizing the margin and minimizing error, i.e. the higher the value of $C$, the more the SVM is focused on minimizing errors.

In dual form, optimization problem (5) can be expressed as [17]:

$$\max_{\alpha_i \in \mathbb{R}} \sum_i \alpha_i - \frac{1}{2} \sum_{i,k} \alpha_i y_i \alpha_k y_k \langle x_i, x_k \rangle$$

s.t. \[
\begin{align*}
0 \leq \alpha_i & \leq C, \\
\sum_i \alpha_i y_i & = 0, \\
i & = 1 \ldots N
\end{align*}
\]
leading to a solution $w_0 = \sum_i \alpha_i y_i x_i$, where $\alpha_i$ are Lagrange multipliers for the dual formulation (6). Term $b$ is calculated a posteriori [18], which is denoted as $b_0$ (standard bias). Finally, the classifier is written as

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) - b_0 = \langle x, w_0 \rangle - b_0$$

[19] where $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, defined as $K(x, y) = \langle x, y \rangle = \langle \phi(x), \phi(y) \rangle$, is called the kernel function.

For moderately imbalanced datasets, empirical results show that, unlike other machine learning algorithms, SVMs can produce a good hypothesis without any modification [1, 43, 23]. One explanation for this phenomenon is that SVMs use only the set of support vectors to construct classification models, and hence the negative instances far from the decision border are not be taken into account, and SVMs remain unaffected.

However, SVMs cannot overcome the problem of imbalance between classes when the bias in the data distribution is significant. In these cases, it has been observed that the separation hyperplane learned by the SVM is very close to the minority class [43, 29, 21], resulting in a low performance or no generalization at all for examples from this class, in comparison with those from the majority class.

Several strategies have been proposed to improve the performance of SVMs on imbalanced datasets: the most promising of which are described in this section. These strategies are introduced in accordance with the stage at which they can be applied in the learning process. A deep study can be found in [12].

### 3.1. Pre-processing strategies

They are primarily based on resampling techniques to balance the dataset, in order to build a new learning set where the minority class is better represented. Since the objective of these algorithms is to balance the dataset before training, they are general purpose procedures, not targeted at a particular machine learning [13, 14].

SMOTE is a widely used pre-processing strategy [7], which is based on the technique of k-nearest neighbours for over-sampling the minority class [38]. Other strategies can also be used, for example, by applying clustering algorithms for sub-sampling the majority class [47, 28].
Taking into account that the function learned by an SVM is built from the support vectors, certain proposals seek to increase representation of the minority class by considering only the margin area between the two classes. The information on the border can be obtained by applying heuristics based on the k-neighbour technique, as proposed in [6]. Other studies are based on the use of SVMs for obtaining the positive support vectors, and over-sample from this data [31, 41]. This feature has also been exploited to build subsampling algorithms, as presented in [35], where an SVM is used to build a new dataset composed of only the negative support vectors and the positive data.

Other solutions have been proposed which use sampling methods with combined classifiers. In this case, several SVMs are trained with different sets of data that have been balanced as a result of applying sub-sampling and/or over-sampling [25, 29, 42, 45].

3.2. Training strategies

These strategies include those proposals which modify the standard optimization problem for SVM training in order to incorporate information related to the proportions of classes in the dataset. One approach is through cost-sensitive learning which, to determine the classification model, incorporates into the learning problem information related with the costs of penalties associated with wrong predictions for each class. For imbalanced scenarios, a similar strategy is used, but with greater weight assigned to errors in the minority class.

Hence, cost information about the two types of errors can be introduced into the formulation of the learning problem (5), using two regularization parameters [37, 8], as follows:

\[
\begin{align*}
\min_{w \in F, b \in \mathbb{R}} & \quad \frac{1}{2} \|w\|^2 + C^{(+)} \sum_{i \mid y_i = +1} \xi_i + C^{(-)} \sum_{i \mid y_i = -1} \xi_i \\
\text{s. t.} & \quad y_i \cdot (\langle x_i, w \rangle - b) + \xi_i \geq 1, \quad i = 1 \ldots N
\end{align*}
\]

(7)

where \(C^{(+)}\) and \(C^{(-)}\) are the costs associated with errors in the positive and negative classes, respectively.

Certain studies have also added new restrictions on the slack variables \(\xi_i\), in order to control the margin of separation between the two classes [44, 21].
A different approach is presented in [4], where only one regularization parameter $C$ is used, although information about the cost of errors is incorporated by assigning a different weight to each slack variable $\xi_i$. As a result, each instance in the data is associated with a cost according to its importance in the classification task.

Other proposed solutions combine cost-sensitive learning with other techniques. For example, the SMOTE algorithm is used in [1] to over-sample the minority class before training SVM with different costs. This strategy has been also used to build combinations of classifiers [39]. A last proposal, the KBA algorithm, is included in this strategic line as described in [43], which modifies the kernel matrix in accordance with the observed imbalance in the distribution of data.

3.3. Post-processing strategies

In general, work following this strategy is oriented either towards modifying the weight vector $w$ in the decision function or towards determining a new bias or threshold, in order to adjust the decision boundary learned by the SVM so that a good margin of separation for the positive class is provided.

Hence, the $z$-SVM method is proposed in [23], which determines the value of a new parameter $z$, by solving an added optimization problem. This optimal parameter weights the contribution of support vectors of the minority class in the vector $w$ of the decision function obtained after training.

In [27], the bias of the decision function is modified by calculating an offset $\theta$ from the average of the unsigned values generated by $f(x)$ for the support vectors. A similar strategy is used in [32], however the new threshold or offset $\theta_{opt}$ is calculated by applying the Beta-Gamma algorithm [46].

Other studies suggest replacing the standard classification scheme (based on the sign function), with other schemes which enable another interpretation of the outputs of the SVM in unbalanced scenarios. For example, a fuzzy decision function is employed in [27] whose parameters are estimated from the observed distribution in the data set. In [40], the decision process incorporates a post-processing module, whose construction is based on methods of information theory, to define a new threshold for classification.

3.4. Towards a new post-processing strategy

The strategies proposed to improve the performance of SVMs on imbalanced datasets generally require tuning new parameters, such as the sample
rate, and the number \( k \) of selected neighbours. Other methods can be computationally expensive since they consider the construction of several classifiers (such as methods based on ensembles), or they are based on iterative algorithms, (such as those which modify the kernel matrix, and certain sampling techniques that require several overtraining steps). In cost-sensitive approaches, the standard SVM optimization problem must be modified and costs of errors on the classes must be known. Furthermore, these approaches can produce over-fitted models [39].

On the other hand, as mentioned earlier, it has been also empirically shown [21, 29, 43] that the hyperplane learned by an SVM in the presence of imbalanced sets has approximately the same orientation as the ideal hyperplane. Reduced generalization on the minority class would be indeed associated with the bias, \( b \), as positive instances lie far from this ideal limit, i.e., the SVM learns a boundary that is much too close to this class. Other studies, such as those presented in [33] in the domain of text classification, suggest increasing research on strategies which determine new thresholds for the SVM decision function. Modifications should be based on the distribution of classes in the dataset, which also do not directly affect standard SVM training.

Thus, a post-processing strategy, based on the consideration that the bias is a parameter to be tuned, is introduced in the next section. The proposed approach inserts no new parameters, neither requires modifying the standard optimization problem for SVM training, nor additional steps, nor re-training.

4. The GSVM: The Geometric (mean) Support Vector Machine

Let us now consider the set of classifiers \( f_b(x) = \langle x, w_0 \rangle - b \), where \( w_0 \) has been obtained by solving (5), and \( b \in \mathbb{R} \) is a free parameter to be tuned. A map \( \Theta_b : \mathcal{X} \to \{-1, +1\} \) can be defined from classifiers \( f_b(x) \) such that, given an input vector \( x \), it assigns a label as follows:

\[
\Theta_b(x) = \text{sign}(f_b(x)) = \begin{cases} +1 & \text{if } \langle x, w_0 \rangle \geq b \\ -1 & \text{otherwise} \end{cases}
\]

**Proposition 1.** The Recall function, \( Rc : \mathbb{R} \rightarrow [0, 1] \), defined from (1) as \( Rc(b) = Rc(f_b, \mathcal{Z}) \) is a decreasing function. Furthermore, \( \max_{b \in \mathbb{R}} Rc(b) = Rc(\beta) = 1 \) and \( \min_{b \in \mathbb{R}} Rc(b) = 0, \forall \beta > \beta^* \), where \( \beta = \min_{z_i \in \mathcal{Z}_{pos}} \langle x_i, w_0 \rangle \), and \( \beta^* = \max_{z_i \in \mathcal{Z}_{pos}} \langle x_i, w_0 \rangle \).
Hence, $Rc$

Lemma 1. and therefore it has been proved that

Thus, if $\Theta_b(x_i) = +1$, then $\Theta_b(x_i) = +1$, and therefore $Rc(b) \geq Rc(b')$. Hence, $Rc(\cdot)$ is a decreasing function.

(i) If $\beta = \min_{z_i \in Z_{pos}} \langle x_i, w_0 \rangle$, then $\langle x_i, w_0 \rangle \geq \beta, \forall z_i \in Z_{pos}$. Hence $\Theta_\beta(x_i) = +1, \forall z_i \in Z_{pos}$, and therefore $t_{pos} = N_{pos}$ and $Rc(\beta) = 1$.

(ii) If $\beta^* = \max_{z_i \in Z_{pos}} \langle x_i, w_0 \rangle$, then $\langle x_i, w_0 \rangle \leq \beta^*, \forall z_i \in Z_{pos}$. Hence, if $b > \beta^*$, then $\Theta_b(x_i) = -1, \forall z_i \in Z_{pos}$, and therefore $t_{pos} = 0$ and $Rc(b) = 0$.

Therefore, it has been proved that

$$\beta = \max \{b \in \mathbb{R} \mid Rc(b) = 1\}, \quad \beta^* = \max \{b \in \mathbb{R} \mid Rc(b) > 0\} \quad \blacksquare$$

Proposition 2. The Specificity function, $Sp : \mathbb{R} \rightarrow [0, 1]$, defined from (2) as $Sp(b) = Sp(f_b, \mathcal{Z})$ is an increasing function. Furthermore, $\min_{b \in \mathbb{R}} Sp(b) = Sp(\alpha^*) = 0$, and $\max_{b \in \mathbb{R}} Sp(b) = 1, \forall b > \alpha$, where $\alpha^* = \min_{z_i \in Z_{neg}} \langle x_i, w_0 \rangle$ and $\alpha = \max_{z_i \in Z_{neg}} \langle x_i, w_0 \rangle$.

Proof: Let $b, b' \in \mathbb{R}$, with $b \leq b'$. Given $z_i \in Z_{neg}$, it holds $f_b(x_i) \geq f_{b'}(x_i)$. Hence, if $\Theta_b(x_i) = -1$, then $\Theta_{b'}(x_i) = -1$, and therefore $Sp(b) \leq Sp(b')$. Hence, $Sp(\cdot)$ is an increasing function.

(i) If $\alpha^* = \min_{z_i \in Z_{neg}} \langle x_i, w_0 \rangle$, then $\langle x_i, w_0 \rangle \geq \alpha^*, \forall z_i \in Z_{neg}$. Hence $\Theta_{\alpha^*}(x_i) = +1, \forall z_i \in Z_{neg}$, and therefore $t_{neg} = 0$ and $Sp(\alpha^*) = 0$.

(ii) If $\alpha = \max_{z_i \in Z_{neg}} \langle x_i, w_0 \rangle$, then $\langle x_i, w_0 \rangle \leq \alpha, \forall z_i \in Z_{neg}$. Hence, if $b > \alpha$, then $\Theta_b(x_i) = -1, \forall z_i \in Z_{neg}$, and therefore $t_{neg} = N_{neg}$ and $Sp(b) = 1$.

Hence, it has been proved that

$$\alpha = \max \{b \in \mathbb{R} \mid Sp(b) < 1\}, \quad \alpha^* = \max \{b \in \mathbb{R} \mid Sp(b) = 0\} \quad \blacksquare$$

The $\alpha^*$, $\alpha$, $\beta$, and $\beta^*$ values are key elements in the development of the SVM, since the optimum in the usual metrics based on the confusion matrix can be found in intervals determined from these elements. Clearly, $\alpha^* \leq \alpha$ and $\beta \leq \beta^*$, however there is also a relationship between $\alpha^*$ and $\beta^*$.

Lemma 1. Using definitions from Propositions 1 and 2, it can be concluded that $\alpha^* \leq \beta^*$.
Proof: For $\alpha \leq \beta$, the proof is trivial.

For the case $\beta < \alpha$, then $\mathcal{Z}$ is a non-linearly separable set [16]. Let us suppose $\alpha^* > \beta^*$, then $b^* = -\frac{\alpha^* + \beta^*}{2}$ is such that $\beta^* < -b^* < \alpha^*$, and $\forall z_i \in \mathcal{Z}_{pos}$ and $\forall z_j \in \mathcal{Z}_{neg}$:

\[
\langle x_i, w_0 \rangle \leq \max_{z_i \in \mathcal{Z}_{pos}} \langle x_i, w_0 \rangle = \beta^* < -b^* \rightarrow \langle x_i, -w_0 \rangle > b^*,
\]

\[
\langle x_j, w_0 \rangle \geq \min_{z_j \in \mathcal{Z}_{neg}} \langle x_j, w_0 \rangle = \alpha^* > -b^* \rightarrow \langle x_j, -w_0 \rangle < b^*.
\]

Thus, a vector ($-w_0$) and a value $b^*$ exist such that the positive and negative instances of $\mathcal{Z}$ are separated linearly, that is, $\mathcal{Z}$ is a linearly separable set. This is a contradiction, therefore $\alpha^* \leq \beta^*$ holds. ■

From the results above, the following Proposition can be enunciated.

**Proposition 3.** The geometric mean function, $Gm : \mathbb{R} \rightarrow [0,1]$, defined from (4) as $Gm(b) = Gm(f_b, \mathcal{Z})$ verifies that

\[
\max_{b \in \mathbb{R}} Gm(b) = \max_{b \in [\lambda, \beta^*]} Gm(b)
\]

where $\lambda = \max \{\alpha^*, \beta\}$.

Proof: If $\alpha < \beta$, then from (4), Propositions 1, and 2, it can be affirmed that $Gm(b) = 1$ for any $\alpha < b \leq \beta$. Therefore, the maximum of $Gm$ is attained in the interval $[\lambda, \beta^*]$, since $\lambda = \beta$.

For the case $\beta \leq \alpha$, then the maximum of $Gm$ is attained in the interval $[\lambda, \beta^*]$, since

- If $b \leq \lambda = \max \{\alpha^*, \beta\}$, then
  - if $\max \{\alpha^*, \beta\} = \alpha^*$, then $Sp(b) = 0$, hence $Gm(b) = 0 \leq Gm(\alpha^*) = Gm(\lambda)$.
  - if $\max \{\alpha^*, \beta\} = \beta$, then $Rc(b) = 1$. Since $Sp(b)$ is an increasing function, it is concluded that $Gm(b) \leq Gm(\beta) = Gm(\lambda)$.

- If $\beta^* < b$, then $Rc(b) = 0$ and $Gm(b) = 0 \leq Gm(\beta^*)$. ■

Let $\mathcal{Z}_{pos} = \{(x_1, +1), \cdots, (x_{N_{pos}}, +1)\} = \{(p_1, +1), \cdots, (p_{N_{pos}}, +1)\}$ be the set of training patterns with positive labels, with $p_i = x_{\sigma^*(i)}$, where $\sigma^*$ is a permutation of $N_{pos}$ elements such that $\beta = \langle p_1, w \rangle \leq \cdots \langle p_1, w \rangle \cdots \leq \langle p_{N_{pos}}, w \rangle = \beta^*$

\[
\beta = \langle p_1, w \rangle \leq \cdots \langle p_1, w \rangle \cdots \leq \langle p_{N_{pos}}, w \rangle = \beta^*
\]
In this form, biases \( b_i = \langle p_i, w \rangle \), \( i = 1, \cdots, N_{\text{pos}} \) are defined in increasing order. The following result shows a more explicit way to find the solution for the maximum of the geometric mean.

**Proposition 4.** The \( G_m \) function, \( G_m : \mathbb{R} \to [0, 1] \), verifies that

\[
\max_{b \in \mathbb{R}} G_m(b) = \max_{i = 1, \cdots, N_{\text{pos}}} G_m(b_i)
\]

**Proof:** From (9), \(-\infty < \beta = b_1 \leq \cdots \leq b_{N_{\text{pos}}} = \beta^* < +\infty\). Hence,

- If \( b \leq b_1 \), then \( Rc(b) = Rc(b_1) = 1 \), \( Sp(b) \leq Sp(b_1) \), and therefore \( G_m(b) \leq G_m(b_1) \).
- If \( b > b_{N_{\text{pos}}} \), then \( 0 = Rc(b) < Rc(b_{N_{\text{pos}}} \), and therefore \( 0 = G_m(b) \leq G_m(b_{N_{\text{pos}}} \).
- If \( b_1 < b \leq b_{N_{\text{pos}}} \), then an index \( i \) exists, \( 1 < i \leq N_{\text{pos}} \), such that \( b_{i-1} < b \leq b_i \). Hence, \( Rc(b) = Rc(b_i) \), \( Sp(b) \leq Sp(b_i) \), and therefore \( G_m(b) \leq G_m(b_i) \).

Furthermore, from the proof of Proposition 4, it is trivial to demonstrate the following result, which is stricter than (8).

**Proposition 5.** Let \( b_m = \min \{ b_i \mid \lambda \leq b_i \} \) and \( b_M = \min \{ b_i \mid \theta \leq b_i \} \), where \( \theta = \max \{ \alpha, \beta^* \} \). Therefore

\[
\max_{b \in \mathbb{R}} G_m(b) = \max_{b_m \leq b \leq b_M} G_m(b_i)
\]

One way to maximize \( G_m \) is given in Proposition 5 by using the properties of the \textit{Recall} and \textit{Specificity} metrics. This is a key result for the development of the GSVM.

Starting from \( w_0 \), the solution vector of the optimization problem (5), the GSVM classifier is defined by modifying the standard bias \( b_0 \),

\[
\Theta_{\text{GSVM}}(x) = \begin{cases} +1 & \text{if } \langle x, w_0 \rangle \geq b_{\text{GSVM}} \\ -1 & \text{if } \langle x, w_0 \rangle < b_{\text{GSVM}} \end{cases}
\]

where the bias \( b_{\text{GSVM}} \) is given as follows:

- If \( \alpha < \beta \), then \( b_{\text{GSVM}} = \frac{\alpha + \beta}{2} \). In this case, the problem (5) is linearly separable and the GSVM bias is equal to the standard bias \( b_0 \) [18]. Hence, if the problem is linearly separable, then the performance of the SVM and the GSVM is the same.
• If $\alpha \geq \beta$, then $b_{\text{GSVM}} = \arg \max_{m \leq i \leq M} G_m(b_i)$ (from Proposition 5).

According to the previous theoretical results, the classification performance will be always higher for GSVM than for SVM on the training set, when the $G_m$ metric is employed. This implies that a more balanced performance between the true positive rate and the true negative rate should be attained.

5. Experimentation

Experimentation is performed on the following standard UCI datasets [5]: Australian Credit Approval, Bupa liver disorders, Pima Indians diabetes, Iris Plants, Teaching assistant evaluation, Japanese Credit Screening, German Credit data, Wine Recognition, Lung Cancer, Haberman’s Survival, Blood Transfusion Service Center, Vehicle, Vertebral Column, Hepatitis, Thyroid Disease, Vowel Recognition, Balance Scale Weight & Distance Database, Dermatology, Glass Identification, and Protein Localization Sites (Ecoli).

Datasets are split into the form of one class, which is indicated in parentheses in the first column of Table 2, versus the rest of the classes. The class which contains the fewest instances has been chosen as the positive one for each dataset. Additionally, another class have been also selected as positive for the Glass, Dermatology and Ecoli datasets. Hence, up to 23 datasets have been used in the experimentation.

A summary of the characteristics for these datasets is shown in Table 2. The imbalance ratio ($IR$) [11], defined as the ratio between the number of instances of the majority and the minority class, is provided to rank the different datasets according to this measure [15].

There is no agreement in the research community on what threshold must be set up for a given dataset to suffer from the imbalance problem [12]. Nevertheless, the problem of imbalanced dataset is considered in our study only as a baseline in order to study the problem of imbalanced accuracy. Hence, let us define the imbalance accuracy ratio, denoted by $IAR$, as follows:

$$IAR = \frac{\max \{\text{Recall, Specificity}\}}{\min \{\text{Recall, Specificity}\}}$$

In this form, the higher the value of $IAR$, the more imbalanced is the Accuracy.
Table 2: UCI datasets for the experimentation.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Total</th>
<th>Positive</th>
<th>IR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australian(1)</td>
<td>690</td>
<td>307</td>
<td>01.25</td>
</tr>
<tr>
<td>Bupa(1)</td>
<td>345</td>
<td>145</td>
<td>01.38</td>
</tr>
<tr>
<td>Pima(2)</td>
<td>768</td>
<td>268</td>
<td>01.87</td>
</tr>
<tr>
<td>Iris(2)</td>
<td>150</td>
<td>050</td>
<td>02.00</td>
</tr>
<tr>
<td>Tae(1)</td>
<td>151</td>
<td>049</td>
<td>02.08</td>
</tr>
<tr>
<td>Japanese(1)</td>
<td>653</td>
<td>203</td>
<td>02.22</td>
</tr>
<tr>
<td>German(1)</td>
<td>1000</td>
<td>300</td>
<td>02.33</td>
</tr>
<tr>
<td>Wine(3)</td>
<td>178</td>
<td>048</td>
<td>02.71</td>
</tr>
<tr>
<td>Lung(1)</td>
<td>030</td>
<td>008</td>
<td>02.75</td>
</tr>
<tr>
<td>Haberman(1)</td>
<td>306</td>
<td>081</td>
<td>02.78</td>
</tr>
<tr>
<td>Transfusion(1)</td>
<td>748</td>
<td>178</td>
<td>03.20</td>
</tr>
<tr>
<td>Vehicle(4)</td>
<td>846</td>
<td>199</td>
<td>03.25</td>
</tr>
<tr>
<td>Column(1)</td>
<td>310</td>
<td>060</td>
<td>04.17</td>
</tr>
<tr>
<td>Hepatitis(1)</td>
<td>129</td>
<td>024</td>
<td>04.38</td>
</tr>
<tr>
<td>Dermat(2)</td>
<td>358</td>
<td>060</td>
<td>04.97</td>
</tr>
<tr>
<td>Ecoli(6)</td>
<td>331</td>
<td>052</td>
<td>05.37</td>
</tr>
<tr>
<td>Thyroid(3)</td>
<td>215</td>
<td>030</td>
<td>06.17</td>
</tr>
<tr>
<td>Vowel(10)</td>
<td>990</td>
<td>090</td>
<td>10.00</td>
</tr>
<tr>
<td>Balance(2)</td>
<td>625</td>
<td>049</td>
<td>11.76</td>
</tr>
<tr>
<td>Glass(4)</td>
<td>214</td>
<td>013</td>
<td>15.46</td>
</tr>
<tr>
<td>Dermat(6)</td>
<td>358</td>
<td>020</td>
<td>16.90</td>
</tr>
<tr>
<td>Glass(5)</td>
<td>214</td>
<td>009</td>
<td>22.78</td>
</tr>
<tr>
<td>Ecoli(5)</td>
<td>331</td>
<td>005</td>
<td>65.20</td>
</tr>
</tbody>
</table>

Experimentation is carried out by following a similar experimental framework to that suggested in [22, 2]. So, training data is normalized in order to prevent problems with outliers, and test data is normalized accordingly. Performance is evaluated on models using the Radial Basis Function (RBF) kernel, with \( \sigma \) (RBF width) and \( C \) (regularization term) explored on a two-dimensional grid: \( \sigma = [2^{-4}, 2^{-3}, \ldots, 2^{5}, 2^{6}] \) and \( C = [2^{-2}, 2^{-1}, \ldots, 2^{6}, 2^{7}] \). The criterion employed to estimate performance is the 3-fold cross-validation on the whole set of test data and this procedure is repeated 25 times in order to ensure good statistical behaviour. Performance is measured according to the \textit{Accuracy}, \textit{Recall}, \textit{Specificity} and \textit{Gm} metrics for the standard SVM, as well as for GSVM.

The results can be checked in Table 3, where \( \sigma \) and \( C \) are the values of the parameters that provide the best \textit{Accuracy} for the standard SVM and the best \textit{Gm} for the GSVM, respectively. The proposed \textit{IAR} measure is also provided.
Table 3: Results of the experiment with an RBF kernel.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>$\sigma$</th>
<th>$C$</th>
<th>Acc/Rec</th>
<th>Spec</th>
<th>Gm</th>
<th>IAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australian(1)</td>
<td>SVM</td>
<td>$2^3$</td>
<td>95.00</td>
<td>98.99</td>
<td>95.96</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>GSVM</td>
<td>$2^3$</td>
<td>95.00</td>
<td>98.99</td>
<td>95.96</td>
<td>1.06</td>
</tr>
<tr>
<td>Bupa(1)</td>
<td>SVM</td>
<td>$2^3$</td>
<td>96.95</td>
<td>96.95</td>
<td>96.95</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>GSVM</td>
<td>$2^3$</td>
<td>96.95</td>
<td>96.95</td>
<td>96.95</td>
<td>1.06</td>
</tr>
<tr>
<td>Pima(2)</td>
<td>SVM</td>
<td>$2^3$</td>
<td>97.96</td>
<td>97.96</td>
<td>97.96</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>GSVM</td>
<td>$2^3$</td>
<td>97.96</td>
<td>97.96</td>
<td>97.96</td>
<td>1.06</td>
</tr>
<tr>
<td>Iris(2)</td>
<td>SVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>GSVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td>Japanese(1)</td>
<td>SVM</td>
<td>$2^3$</td>
<td>99.97</td>
<td>99.97</td>
<td>99.97</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>GSVM</td>
<td>$2^3$</td>
<td>99.97</td>
<td>99.97</td>
<td>99.97</td>
<td>1.06</td>
</tr>
<tr>
<td>German(1)</td>
<td>SVM</td>
<td>$2^3$</td>
<td>97.96</td>
<td>97.96</td>
<td>97.96</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>GSVM</td>
<td>$2^3$</td>
<td>97.96</td>
<td>97.96</td>
<td>97.96</td>
<td>1.06</td>
</tr>
<tr>
<td>Wine(3)</td>
<td>SVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>GSVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td>Lung(1)</td>
<td>SVM</td>
<td>$2^3$</td>
<td>99.97</td>
<td>99.97</td>
<td>99.97</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>GSVM</td>
<td>$2^3$</td>
<td>99.97</td>
<td>99.97</td>
<td>99.97</td>
<td>1.06</td>
</tr>
<tr>
<td>Haberman(1)</td>
<td>SVM</td>
<td>$2^3$</td>
<td>97.96</td>
<td>97.96</td>
<td>97.96</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>GSVM</td>
<td>$2^3$</td>
<td>97.96</td>
<td>97.96</td>
<td>97.96</td>
<td>1.06</td>
</tr>
<tr>
<td>Transfusion(1)</td>
<td>SVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>GSVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td>Vehicle(4)</td>
<td>SVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>GSVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td>Column(1)</td>
<td>SVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>GSVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td>Hepatitis(1)</td>
<td>SVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>GSVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td>Dermat(2)</td>
<td>SVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>GSVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td>Ecoli(6)</td>
<td>SVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>GSVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td>Thyroid(3)</td>
<td>SVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>GSVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td>Vowel(10)</td>
<td>SVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>GSVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td>Balance(2)</td>
<td>SVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>GSVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td>Glass(4)</td>
<td>SVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>GSVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td>Dermat(6)</td>
<td>SVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>GSVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td>Glass(5)</td>
<td>SVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>GSVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td>Ecoli(5)</td>
<td>SVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>GSVM</td>
<td>$2^3$</td>
<td>98.97</td>
<td>98.97</td>
<td>98.97</td>
<td>1.06</td>
</tr>
</tbody>
</table>

Some conclusions can be drawn from the empirical experimentation carried out:

- The value of the $Gm$ in the GSVM is always greater than the value of
the $Gm$ in the Standard SVM for any dataset, except in the Dermat(6) and Vowel datasets where no difference exists. Therefore, the main aim of the GSVM is achieved, that is, a more balanced performance between the true positive rate and true negative rate is attained in the test set ($IAR_{GSVM} \leq IAR_{SVM}$).

- Dermat(6) and Vowel datasets can be considered as highly imbalanced datasets since $IR \geq 10$, however not imbalanced accuracy exists when a standard SVM is employed ($IAR \leq 1.03$). Therefore, no special strategies are needed in both datasets related with the imbalance between classes. The reason for this so good classification behaviour is that the SVM problem is linearly separable for both datasets.

- The value of the Accuracy in the GSVM is higher than the value of the Accuracy in the Standard SVM in six datasets (Australian, Bupa, Iris, Wine, Dermat(2), and Thyroid). Therefore, the GSVM also enables improvement in the Accuracy metric in several cases in addition to improvement in the $Gm$.

- In the Bupa, Tae, Japanese, German, Lung, Haberman, Transfusion, Hepatitis, Balance, Glass(4), Glass(5), and Ecoli(5) datasets, the value of Recall is low for SVM, although a significant improvement is attained with the GSVM. The value of Specificity is improved in the Australian, Pima, Wine, and Vehicle datasets.

- The Bupa dataset cannot be considered as an imbalanced dataset ($IR = 1.38$), however there exists a high imbalance accuracy ratio ($IAR_{SVM} = 1.81$), that is, Specificity is 81% higher than Recall. A similar conclusion can be drawn for Pima, Tae, Japanese, German, Lung, Haberman, Transfusion, and Hepatitis datasets.

In order to compare the performance of this post-processing strategy with other methods reported in the literature, the cost-sensitive scheme provided by the Matlab Bioinformatic’s toolbox is implemented to train SVM on the listed UCI datasets. Results obtained using the same evaluation metrics, as well as the same cross-validation structure are shown in Table 4, with $\sigma$ and $C$ being the values for the parameters providing the best Accuracy for the cost-sensitive SVM and the best $Gm$ for the GSVM, respectively. Let us indicate that given a $C$ value, the values for $C^{(+)}$ and $C^{(-)}$ in (7) used in the
Matlab toolbox are $\frac{C\cdot N}{N_{pos}}$ and $\frac{C\cdot N}{N_{neg}}$, respectively, that is, the ratio between them is the imbalance ratio,

$$\frac{C^+}{C^-} = \frac{N_{neg}}{N_{pos}} = IR$$

Table 4: Results of the experimentation with an RBF kernel for the cost-sensitive SVM. When the cost-sensitive SVM behave better than the GSVM, it is presented in bold.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>$\sigma$</th>
<th>$C$</th>
<th>Accuracy</th>
<th>Recall</th>
<th>Specificity</th>
<th>Gm</th>
<th>IAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australian(1)</td>
<td>2²</td>
<td>2²</td>
<td>66.39</td>
<td>70.91</td>
<td>62.50</td>
<td>66.54</td>
<td>1.13</td>
</tr>
<tr>
<td>Bupa(1)</td>
<td>2²</td>
<td>2²</td>
<td>68.51</td>
<td>63.22</td>
<td>72.30</td>
<td>67.48</td>
<td>1.14</td>
</tr>
<tr>
<td>Pima(2)</td>
<td>2²</td>
<td>2²</td>
<td>71.85</td>
<td>67.95</td>
<td>73.96</td>
<td>70.79</td>
<td>1.09</td>
</tr>
<tr>
<td>Iris(2)</td>
<td>2¹</td>
<td>2¹</td>
<td>95.96</td>
<td>96.58</td>
<td>95.71</td>
<td>96.09</td>
<td>1.01</td>
</tr>
<tr>
<td>Tae(1)</td>
<td>2¹</td>
<td>2¹</td>
<td>71.39</td>
<td>43.88</td>
<td>84.58</td>
<td>60.37</td>
<td>1.93</td>
</tr>
<tr>
<td>Japanese(1)</td>
<td>2²</td>
<td>2¹</td>
<td>62.13</td>
<td>28.15</td>
<td>78.82</td>
<td>44.38</td>
<td>3.13</td>
</tr>
<tr>
<td>German(1)</td>
<td>2¹</td>
<td>2¹</td>
<td>70.64</td>
<td>69.63</td>
<td>71.04</td>
<td>70.22</td>
<td>1.36</td>
</tr>
<tr>
<td>Wine(3)</td>
<td>2⁰</td>
<td>2⁰</td>
<td>85.78</td>
<td>73.88</td>
<td>90.18</td>
<td>81.27</td>
<td>1.22</td>
</tr>
<tr>
<td>Lung(1)</td>
<td>2¹</td>
<td>2¹</td>
<td>69.61</td>
<td>65.08</td>
<td>71.61</td>
<td>61.69</td>
<td>1.10</td>
</tr>
<tr>
<td>Haberman(1)</td>
<td>2¹</td>
<td>2¹</td>
<td>71.10</td>
<td>51.60</td>
<td>78.12</td>
<td>63.14</td>
<td>1.51</td>
</tr>
<tr>
<td>Transfusion(1)</td>
<td>2¹</td>
<td>2¹</td>
<td>68.03</td>
<td>54.35</td>
<td>72.29</td>
<td>62.49</td>
<td>1.33</td>
</tr>
<tr>
<td>Vehicle(4)</td>
<td>2²</td>
<td>2²</td>
<td>97.30</td>
<td>95.18</td>
<td>98.00</td>
<td>96.68</td>
<td>1.03</td>
</tr>
<tr>
<td>Column(1)</td>
<td>2¹</td>
<td>2¹</td>
<td>85.10</td>
<td>90.13</td>
<td>83.88</td>
<td>86.80</td>
<td>1.07</td>
</tr>
<tr>
<td>Hepatitis(1)</td>
<td>2²</td>
<td>2⁰</td>
<td>78.57</td>
<td>62.67</td>
<td>82.81</td>
<td>70.83</td>
<td>1.32</td>
</tr>
<tr>
<td>Dermat(2)</td>
<td>2¹</td>
<td>2¹</td>
<td>93.85</td>
<td>95.63</td>
<td>85.17</td>
<td>90.13</td>
<td>1.12</td>
</tr>
<tr>
<td>Ecoli(6)</td>
<td>2⁻²</td>
<td>2⁻²</td>
<td>94.51</td>
<td>91.24</td>
<td>95.26</td>
<td>93.10</td>
<td>1.04</td>
</tr>
<tr>
<td>Thyroid(3)</td>
<td>2⁻²</td>
<td>2⁻²</td>
<td>97.68</td>
<td>93.18</td>
<td>98.38</td>
<td>95.67</td>
<td>1.06</td>
</tr>
<tr>
<td>Vowel(10)</td>
<td>2¹</td>
<td>2²</td>
<td>99.83</td>
<td>98.31</td>
<td>99.98</td>
<td>99.12</td>
<td>1.02</td>
</tr>
<tr>
<td>Balance(2)</td>
<td>2¹</td>
<td>2¹</td>
<td>70.25</td>
<td>72.17</td>
<td>70.18</td>
<td>70.64</td>
<td>1.03</td>
</tr>
<tr>
<td>Glass(4)</td>
<td>2¹</td>
<td>2¹</td>
<td>97.96</td>
<td>96.93</td>
<td>98.01</td>
<td>97.63</td>
<td>1.26</td>
</tr>
<tr>
<td>Dermat(6)</td>
<td>2¹</td>
<td>2¹</td>
<td>98.11</td>
<td>94.98</td>
<td>98.30</td>
<td>96.56</td>
<td>1.03</td>
</tr>
<tr>
<td>Glass(5)</td>
<td>2¹</td>
<td>2¹</td>
<td>81.81</td>
<td>73.33</td>
<td>82.18</td>
<td>76.67</td>
<td>1.12</td>
</tr>
<tr>
<td>Ecoli(5)</td>
<td>2⁻²</td>
<td>2⁻²</td>
<td>89.44</td>
<td>96.00</td>
<td>89.35</td>
<td>92.27</td>
<td>1.07</td>
</tr>
</tbody>
</table>

Based on the empirical experimentation carried out, a list of conclusions can be drawn:

- GSVM obtains better results than cost-sensitive SVM in nine datasets (Australian, Bupa, Pima, Wine, Lung, Transfusion, Thyroid, Balance, and Dermatology) for all the metrics. However, cost-sensitive SVM approach behaves better than GSVM in two datasets (Ecoli(6) and Vowel) for all the metrics.
• With respect to the *Accuracy*, *Specificity* and *Gm* metrics, GSVM obtains better results than cost-sensitive SVM in 17 out of 23 datasets. The same result is obtained for the imbalance accuracy rate, *IAR*.

• With respect to the *Recall* metrics, GSVM is experimentally better than cost-sensitive SVM in 14 out of 23 datasets.

The Wilcoxon Signed-Rank Test [10], a non-parametric statistical procedure, has been employed to analyse whether differences are significant for the results obtained when comparing the GSVM and the standard SVM approaches, as well as the GSVM and the cost-sensitive SVM approaches on the different datasets for each considered metric. In this way, Table 5 has been obtained, where the p-value is displayed, which represents the lowest level of significance of a hypothesis resulting in a rejection.

<table>
<thead>
<tr>
<th>GSVM vs ...</th>
<th>Accuracy</th>
<th>Recall</th>
<th>Specificity</th>
<th>Gm</th>
<th>IAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.0071</td>
<td>0.0003</td>
<td>0.0064</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td>cost-sensitive SVM</td>
<td>0.0208</td>
<td>0.0885</td>
<td>0.1070</td>
<td>0.0097</td>
<td>0.0417</td>
</tr>
</tbody>
</table>

From this Table, for a confidence level fixed at 1%, it can be concluded that differences are significant for the results obtained between GSVM and standard SVM for all the considered metrics.

When comparing GSVM and cost-sensitive SVM, for a 5% confidence level, then it can be concluded that the performance of GSVM is better than that of cost-sensitive SVM for the *Gm* metric, *Accuracy* metric and *IAR*. In terms of the *Specificity* metric, both methods can be considered equivalent. Furthermore, for a confidence level fixed at 9%, it can be also concluded that the performance of GSVM for the *Recall* metric is better than that of cost-sensitive SVM.

6. Conclusions

A new SVM has been designed whose main aim is to improve the geometric mean metric in order to obtain a more balanced performance between the true positive rate and the true negative rate.

The proposed approach does not involve tuning new parameters and, furthermore, it neither requires modifying the standard optimization problem for training the SVM, nor does it require additional steps of re-training.
Experimentation on 23 standard UCI datasets shows this approach to be a good alternative to standard SVM since the property to attain a more balanced performance between the positive and the true negative ratio is generalized to the test set.

It is worth noting that the GSVM can be used with any variation for SVM; moreover this strategy can be applied to any classifier that incorporates a bias in its definition.

7. Acknowledgements

This research has been partly supported by the Spanish Ministry of Science and Innovation R&D projects TIN2009-14378-C02-01, TIN2012-38416-C03-01, and Simon (TIC-8052) of the Andalusian Regional Ministry of Economy, and the PG-03-7678-2009/1 project from the Central University of Venezuela.

References


