Applied Soft Computing 61 (2017) 661–669

Contents lists available at ScienceDirect

Applied Soft Computing

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

Handling binary classification problems with a priority class by using Support Vector Machines

L. Gonzalez-Abril[a,*], C. Angulo[b], H. Nuñez[c], Y. Leal[d]

[a] Dpto. Economía Aplicada I, Universidad de Sevilla, 41018 Sevilla, Spain
[b] Automatic Control Dept. – Univ. Politècnica de Catalunya, 08028 Barcelona, Spain
[c] AI Lab. Facultad de Ciencias, Univ. Central Venezuela, 1020-A Caracas, Venezuela
[d] Unitat of Diabetes, Endocrinology and Nutrition (UDEN), Institut d'Investigació Biomédica de Girona (IdIBGi), 17007 Girona, Spain

ARTICLE INFO

Article history:
Received 28 October 2016
Received in revised form 29 May 2017
Accepted 8 August 2017
Available online 18 August 2017

Keywords:
Support Vector Machines
Post-processing strategies
Pattern recognition
Cost-sensitive SVM

ABSTRACT 这里的priority class就是一种类别比另一种类别更重要

A post-processing technique for Support Vector Machine (SVM) algorithms for binary classification problems is introduced in order to obtain adequate accuracy on a priority class (labelled as a positive class). That is, the true positive rate (or recall or sensitivity) is prioritized over the accuracy of the overall classifier. Hence, false negative (or Type I) errors receive greater consideration than false positive (Type II) errors during the construction of the model.

This post-processing technique tunes the initial bias term once a solution vector is learned by using standard SVM algorithms in two steps: First, a fixed threshold is given as a lower bound for the recall measure; second, the true negative rate (or specificity) is maximized.

Experiments, carried out on eleven standard UCI datasets, show that the modified SVM satisfies the aims for which it has been designed. Furthermore, results are comparable or better than those obtained when other state-of-the-art SVM algorithms and other usual metrics are considered.

1. Introduction 一种类别的发现比另一种类别更加的重要优先

There exist situations where the correct detection of instances of one class (the positive class) is considered to be of greater importance or priority than the other class (the negative class) in binary classification problems. Problems which involve this situation arise: in the medical diagnosis of certain disorders, such as the detection of breast cancer [16,30] or cardiac care [23]; in certain financial problems, such as credit-card fraud detection [29], financial crisis [18], detection of financial statement fraud [24,6], bankruptcy prediction [27,21], prediction of liquefaction potential [28] and bank marketing [20]; and in criminological investigations, among other applications.

For these kinds of problems, the attainment of a high percentage for the true positive rate (also named recall or sensitivity) is more important than the overall accuracy measure. That is, Type I errors should receive serious consideration during the model construction. For these cases, models should therefore exhibit a high performance in the priority class, while simultaneously striving to maintain a low error performance in the non-priority class. For instance, when classifying whether a firm is bankrupt, Type I error occurs when the classifier incorrectly classifies a bankrupt firm into the non-bankrupt class. Type II errors are based on the classifier incorrectly classifying a non-bankrupt firm into the bankrupt class. A higher Type I error rate incurs greater costs on financial institutions, which can enhance business risk.

In order to handle this kind of problem, the application of Support Vector Machines (SVMs) is a popular choice in the machine-learning research area since these are learning machines that implement the structural-risk-minimization inductive principle to obtain good generalization on a limited number of learning patterns [25,26,22]. The theory of SVMs was developed on the basis of a separable binary classification problem where the optimization criterion is the width of the margin with ℓ_2-norm,[1] between the positive and negative examples. An SVM with a large margin separating two classes has a small Vapnik-Chervonenkis dimension, which provides good generalization performance [5].

Furthermore, SVMs present an attractive option for binary problems with a priority class, since they can be modified in order to incorporate information on the penalties associated with erroneous 错误的预测 predictions for each class into the learning problem. Therefore, SVMs can afford to prioritize one class over the other. 合并信息

* Corresponding author.
E-mail address: luisgon@us.es (L. Gonzalez-Abril).

[1] A generalization is given in [13].

http://dx.doi.org/10.1016/j.asoc.2017.08.023
1568-4946/

The main contribution in this paper is that once the solution vector for the standard SVM problem is obtained, the bias is considered as a parameter to be tuned in order to improve the generalization performance on the non-priority class, whereas a generalization performance with a threshold is maintained on the priority class.

It must be emphasized that the proposed solution neither modifies the original optimization problem for SVM training, nor introduces new hyper-parameters. Thus, no increase in the computational cost is incurred.

The remainder of this paper is organized as follows: Section 2 introduces the metrics commonly employed in classification problems. In Section 3, the standard SVM approach is outlined and a family of classifiers that depends on the bias is considered. Furthermore, certain results on the metrics are obtained. An SVM, called Biased SVM (BSVM), is introduced in Section 4, based on both recall and specificity metrics. Experiments are carried out in Section 5 which illustrate the performance of the BSVM in comparison with the standard SVM and the well-known cost-sensitive SVM approach. Finally, conclusions are drawn.

2. Metrics

Let $\mathcal{Z} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ be a training set, with $x_i \in \mathcal{X} \subset \mathbb{R}^d$, $y_i \in \mathcal{Y} = \{+1, -1\}$, and $z_i = (x_i, y_i)$. Let $f(x)$ be a binary classifier such that outputs are obtained as $h(x) = sign(f(x))$, where $sign(\cdot)$ is the sign function. Let $\mathcal{Z}_{pos} = \{z_i \in \mathcal{Z} | y_i = +1\} \neq \emptyset$ and $\mathcal{Z}_{neg} = \{z_i \in \mathcal{Z} | y_i = -1\} \neq \emptyset$ be the sets of training patterns for the positive class and the negative class, respectively. Let $N_{pos} = \#\mathcal{Z}_{pos}$ and $N_{neg} = \#\mathcal{Z}_{neg}$ be the number of positive and negative instances, respectively. Hence $N = N_{pos} + N_{neg}$.

Based on the confusion matrix or contingency table (see Table 1), 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*, denoted by *Ac*, which computes the proportion of instances that are correctly classified by the model, that is,

$$Ac(f, \mathcal{D}) = \frac{t_{pos} + t_{neg}}{N_{pos} + N_{neg}}.$$

Using this metric, both classes, positive and negative, have the same priority for the purpose of classification. Hence, accuracy can be deceiving when:

1. one class is considered of greater significance than the other class. Thus, the cost for wrong classifications in the priority class (Type I error) is greater than the cost for wrong classifications in the other class (Type II error). For instance, a false negatives, which may deny medical treatment to a patient, is more critical than a false positive that leads to carrying out a medical check on a healthy person; and
2. prior probabilities of classes differ greatly (imbalanced datasets), since the metrics fail to consider costs for wrong classifications, and thus remain very sensitive to the bias between classes [14]. For instance, when detecting fraud in credit-card transactions, the ratio between fraudulent and healthy transaction instances is around 1/6000; a naive approach of classifying every example to be a negative instance would provide a high accuracy, however this description fails to reflect the fact that none of the fraudulent transactions is detected.

Therefore, other measures of assessment must be considered. By considering accuracy rates on $\mathcal{Z}_{pos}$ and $\mathcal{Z}_{neg}$ separately, the *recall*[2] and *specificity* metrics, denoted by *Re* and *Sp*, respectively, are defined as follows:

Table 1
The confusion matrix for a classifier f on a test set $\mathcal{D}$.

	Actual	
Predicted	Positive	Negative
Positive	True positive (t_{pos})	False positive (f_{pos})
Negative	False negative (f_{neg})	True negative (t_{neg})
Number of instances	N_{pos}	N_{neg}

$$Re(f, \mathcal{D}) = \frac{t_{pos}}{N_{pos}}, \text{ and } Sp(f, \mathcal{D}) = \frac{t_{neg}}{N_{neg}}. \quad (1)$$

The recall measure is the proportion of positive cases that are correctly identified (the true positive rate), that is, one minus the proportion of Type I error. On the other hand, specificity is the fraction of correctly identified examples among all instances that are negative (the true negative rate), that is, one minus the proportion of Type II error. Both metrics are measures of completeness.

Hence, a new expression for accuracy can be derived as follows:

$$Ac(f, \mathcal{D}) = \frac{N_{pos} \cdot Re(f, \mathcal{D}) + N_{neg} \cdot Sp(f, \mathcal{D})}{N_{pos} + N_{neg}} \quad (2)$$

as a weighted arithmetic mean of recall and specificity, where the associated weights are the number of positive instances and negative instances, respectively.

In the case that the set $\mathcal{Z}_{pos}$ is considered a "priority" for the purpose of classification, then the $Re(f, \mathcal{D})$ measure is more representative than the $Sp(f, \mathcal{D})$ measure, and $Ac(f, \mathcal{D})$ is no longer an adequate metric [8].

Therefore, a new approach to obtain a classifier when one class is considered a "priority" is introduced: For a fixed test set $\mathcal{D} \subset \mathcal{X} \times \mathcal{Y}$, when seeking for a good classifier, the search is restricted to classifiers that hold recall measures at a specified level. Within this set of classifiers, a classifier is sought in order to maximize specificity. It is worth noting that, in general, it is not possible to arbitrarily obtain great recall and specificity values, that is, the capacity of any classifier cannot increase the number of the true positives without also increasing the number of false positives.

According to the proposed approach, the following well-known metrics will be also considered in the measurement of the performance of the classifiers:

- The *geometric mean* (*g*-mean), defined as

$$G_{mean}(f, \mathcal{D}) = \sqrt{Re(f, \mathcal{D}) \cdot Sp(f, \mathcal{D})}.$$

- The *precision*, given by

$$Pr(f, \mathcal{D}) = \frac{t_{pos}}{t_{pos} + f_{pos}}.$$

- The F_{value} metric (*f*-value), defined as

$$F_{value}(f, \mathcal{D}) = (1 + \beta^2) \cdot \frac{Pr(f, \mathcal{D}) \cdot Re(f, \mathcal{D})}{\beta^2 Re(f, \mathcal{D}) + Pr(f, \mathcal{D})},$$

where β is a tuning parameter (usually, $\beta = 1$).

[2] Also called sensitivity in certain fields.

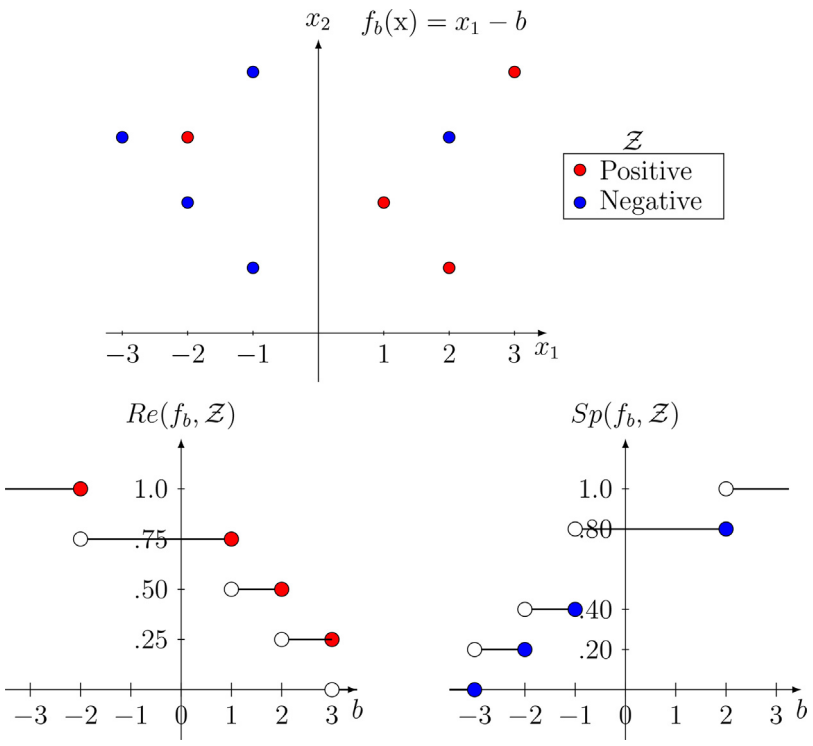

Fig. 1. Graphical representation on a toy dataset of the recall and the specificity metrics with bias b taken as a parameter.

3. Support Vector Machine and bias

赋予

Let $\boldsymbol{\phi} : \mathcal{X} \to \mathcal{H}, \mathrm{x} = \phi(x)$ be a mapping to a feature space endowed with a dot product denoted by $\langle \cdot, \cdot \rangle$. A binary linear classifier $f : \mathcal{X} \to \mathbb{R}$, defined as[3] $f(x) = \langle \mathrm{x}, \mathrm{w} \rangle - b$, is sought, where $\mathrm{w} \in \mathcal{H}, b \in \mathbb{R}$. Outputs are obtained as $h(x) = sign(f(x))$.

3.1. The standard primal SVM

原始的

The standard SVM formulation leads to the optimization problem [10]:

$$\min_{\mathrm{w}\in\mathcal{H},\ b\in\mathbb{R}} \ \frac{1}{2}\|\mathrm{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$
$$\text{s.t.} \ \begin{cases} y_i\left(\langle \mathrm{x_i}, \mathrm{w}\rangle - b\right) + \xi_i \geq 1, \\ \xi_i \geq 0, \ z_i \in \mathcal{Z}\ i = 1, \ldots, N \end{cases} \tag{3}$$

拉格朗日乘数

where ξ_i are slack variables and $C > 0$ acts as a term of regularization.

The solution of this problem can be written as $\mathrm{w}_0 = \sum_i \gamma_i y_i \mathrm{x_i}$, where $\mathrm{x_i} = \phi(x_i)$ and γ_i are Lagrange multipliers for the dual formulation of (3), with $\sum_i \gamma_i y_i = 0$. Term b is calculated a posteriori [11], and is denoted by b_0 (bias standard). Hence, the classifier can be written as

$$f(x) = \sum_{i=1}^{N} \gamma_i y_i K(x_i, x) - b_0 = \langle \mathrm{x}, \mathrm{w}_0 \rangle - b_0$$

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

3.2. The cost-sensitive SVM

合并信息

The cost-sensitive SVM modifies the standard optimization problem (3) in order to incorporate information on the penalties associated with erroneous predictions for each class into the learning problem [7]. Thus, the two types of errors can be introduced into the formulation of the learning problem, using two regularization parameters, as follows [4]:

考虑到俩种类别错误的预测，使用俩种正则化分别管理俩种类别的错误预测

$$\min_{\mathrm{w}\in\mathcal{H},\ b\in\mathbb{R}} \ \frac{1}{2}\|\mathbf{w}\|^2 + C^+\sum_{i|y_i=+1}\xi_i + C^-\sum_{i|y_i=-1}\xi_i$$
$$\text{s.t.} \ \begin{cases} y_i\left(\langle \mathrm{x_i}, \mathrm{w}\rangle - b\right) + \xi_i \geq 1, \\ \xi_i \geq 0, \ z_i \in \mathcal{Z}\ i = 1, \ldots, N \end{cases} \tag{4}$$

where C^+ and C^- are the costs associated with errors in the positive class and negative class, respectively.

Furthermore, the solution vector can be written in the same way as the solution vector of the optimization problem (3). It is worth noting that the cost-sensitive SVM usually provides a better recall measure than does the standard SVM. Nevertheless, this approach fails to provide a way to regulate the level of recall. The SVM proposed in this paper regulates via the bias tuning.

对B进行微调得出的　　规范

3.3. The recall and specificity metrics

依据

A key point of the proposed approach is that once the solution vector w_0 is obtained using the standard SVM formulation, the bias b can be considered as a parameter. In accordance with to the notation in [12], the set of classifiers $\mathcal{F}(\mathrm{w}_0)$ defined as

$$\left\{ f_b : \mathcal{X} \to \mathbb{R}, f_b(\,\cdot\,) = \langle \phi(\,\cdot\,), \mathrm{w}_0 \rangle - b, b \in \mathbb{R} \right\}$$

[3] For notation consistency, bias b is incorporated with a negative sign.

is considered. A map $\Theta_b : \mathcal{X} \to \{-1, +1\}$ is also defined, and is associated to the classifier $f_b(x) \in \mathcal{F}(w_0)$ such that, given an input vector x, it assigns a label as follows:

$$\Theta_b(x) = sign(f_b(x)) = \begin{cases} +1 & \text{if } \langle x, w_0\rangle \geq b, \\ -1 & \text{if } \langle x, w_0\rangle < b. \end{cases} \quad (5)$$

Let us define the recall function, $Re : \mathbb{R} \to [0, 1]$, as $Re(b) = Re(f_b, \mathcal{Z})$ and the specificity function, $Sp : \mathbb{R} \to [0, 1]$, as $Sp(b) = Sp(f_b, \mathcal{Z})$ from (1), where $f_b \in \mathcal{F}(w_0)$.

Furthermore, let order

$$\mathcal{Z}_{pos} = \{(x_1, +1), \ldots, (x_{N_{pos}}, +1)\} = \{(p_1, +1), \ldots, (p_{N_{pos}}, +1)\}$$

with $p_i = x_{\sigma^*(i)}$, where σ^* is a permutation of N_{pos} such that 排列

$$\langle p_1, w_0\rangle \leq \cdots \leq \langle p_i, w_0\rangle \leq \cdots \leq \langle p_{N_{pos}}, w_0\rangle$$

for $i = 1, \ldots, N_{pos}$.

Let us consider the values

$$\beta = \min_{z_i \in \mathcal{Z}_{pos}} \langle x_i, w_0\rangle = \langle p_1, w_0\rangle \text{ and } \beta^* = \max_{z_i \in \mathcal{Z}_{pos}} \langle x_i, w_0\rangle = \langle p_{N_{pos}}, w_0\rangle.$$

Hence, by defining the bias $b_i = \langle p_i, w_0\rangle$ for $i = 1, \ldots, N_{pos}$, the following results hold:

- $Re(b)$ is a decreasing function of b.
- $Re(b_i) \geq \frac{N_{pos}-(i-1)}{N_{pos}}$ $\forall i$. Furthermore, if $b_i \neq b_j$ for $i \neq j$, then $Re(b_i) = \frac{N_{pos}-(i-1)}{N_{pos}}$ and $Re(b_i) = Re(b_{i+1}) + \frac{1}{N_{pos}}$.
- If $b_i < b_{i+1}$, then $Re(b_i) > Re(b) = Re(b_{i+1})$ for any b such that $b_i < b \leq b_{i+1}$.
- If $b \leq \beta < b' \leq \beta^* < b^*$, then $Re(b) = 1 > Re(b') > 0 = Re(b^*)$.

Analogously, let 类似的

$$\mathcal{Z}_{neg} = \{(x_1, -1), \ldots, (x_{N_{neg}}, -1)\} = \{(q_1, -1), \ldots, (q_{N_{pos}}, -1)\}$$

with $q_j = x_{\sigma'(j)}$ where σ' is a permutation of N_{neg} such that 排列

$$\langle q_1, w_0\rangle \leq \cdots \leq \langle q_j, w_0\rangle \leq \cdots \leq \langle q_{N_{neg}}, w_0\rangle$$

for $j = 1, \ldots, N_{neg}$.

Let us consider the values

$$\alpha^* = \min_{z_i \in \mathcal{Z}_{neg}} \langle x_i, w_0\rangle = \langle q_1, w_0\rangle \text{ and } \alpha = \max_{z_i \in \mathcal{Z}_{neg}} \langle x_i, w_0\rangle = \langle q_{N_{neg}}, w_0\rangle.$$

Hence, by considering the bias $bq_j = \langle q_j, w_0\rangle$ for $j = 1, \ldots, N_{neg}$, it follows that:

- $Sp(b)$ is an increasing function of b.
- $Sp(bq_j) \leq \frac{j-1}{N_{neg}}$. Furthermore, if $bq_i \neq bq_j$ for $i \neq j$, then $Sp(bq_j) = \frac{j-1}{N_{neg}}$ and $Sp(bq_j) + \frac{1}{N_{neg}} = Sp(bq_{j+1})$.
- If $bq_j < bq_{j+1}$, then $Sp(bq_j) < Sp(b) \leq Sp(bq_{j+1})$ for any b such that $bq_j < b \leq bq_{j+1}$.
- If $b \leq \alpha^* < b' \leq \alpha < b^*$, then $Sp(b) = 0 < Sp(b') < 1 = Sp(b^*)$.

By applying a toy dataset, an example for both metrics is depicted in Fig. 1, where $x = (x_1, x_2)$, $w_0 = (1, 0)^t$, and $f_b(x) = x_1 - b$. It can be verified that $N_{pos} = 4$, $N_{neg} = 5$, $\alpha = 2$, $\alpha^* = -3$, $\beta = -2$, and $\beta^* = 3$.

As expected, recall and specificity are inverse measures in the sense that an improvement of one of them implies a deterioration of the other. 恶化

4. BSVM: biased support vector machine 权衡

Within the set of classifiers $\mathcal{F}(w_0)$, a tradeoff must be found between $Re(b)$, a decreasing function of b, and $Sp(b)$, an increasing function of b, in order to maximize generalization.

Table 2
UCI datasets used in the experimentation. The number in parentheses indicates the priority class. These datasets are ordered in terms of the number of instances.

Datasets	Total	Priority	Brief description
Iris (2)	150	50	The best known dataset for classification
Tae (1)	151	49	Evaluations of teaching performance
Glass (1)	214	70	Motivated by criminological investigation
Thyroid (3)	215	30	Thyroid disease
Ecoli (6)	331	52	Localization site of protein
Bupa (1)	345	145	Medical diagnosis of livers
Japanese (1)	653	203	Relative to financial problems
Australian (1)	690	307	Relative to financial problems
Pima (2)	768	268	Medical diagnosis of diabetes
German (2)	1000	300	Relative to financial problems
Bank (1)	4521	521	Relative to financial problems

By assuming that the positive class is of a higher priority than the negative class, then $\beta = \max\{b \in \mathbb{R},\ Re(b) = 1\}$ is the best bias for the training set $\mathcal{Z}$, which leads to the classifier $f_\beta(x) = \langle x, w_0\rangle - \beta$. Nevertheless, $Sp(\beta)$ can be arbitrarily small if the instance p_1 is an outlier. It can be seen in Fig. 1 that if $\beta = -2$, then $Re(-2) = 1$, $Sp(-2) = 0.20$ and $Ac(-2) = \frac{5}{9} = 0.5556$. 绕过这个问题

One way to circumvent this problem, given the set of classifiers $\mathcal{F}(w_0)$, is to fix a threshold $0 \leq r \leq 1$, such that $Re(b) \geq r$ in order to guarantee a minimum true positive rate in the positive class and to maximize $Sp(b)$. For example, from Fig. 1, if $r = 0.70$ then it can be seen that by considering $b = 1$, one obtains $Re(1) = 0.75 > 0.70$, $Sp(1) = 0.80$, and $Ac(1) = \frac{7}{9} = 0.7778$.

Hence, the following problem is considered:

$$\max_{b \in \mathbb{R}} \ Sp(b) \quad \text{s.t. } Re(b) \geq r,\ 0 \leq r \leq 1,\ f_b \in \mathcal{F}(w_0) \quad (6)$$

Proposition 4.1. *A classifier $f_b(x) \in \mathcal{F}(w_0)$ exists such that it is a solution of the problem (6).*

Proof. Let us consider the two possible cases: 线性可分

- If $\alpha < \beta$, then $\mathcal{Z}$ is a linearly separable training set [9], and hence, for any b such that $\alpha < b < \beta$, $Re(b) = 1$ and $Sp(b) = 1$ from the results given in Section 3, whereby both metrics obtain the maximum value. Therefore, b is a solution of the problem (6) for any $0 \leq r \leq 1$.

  In this case, the bias $b_r^* = (1 - r)\beta + r\alpha$ is a solution of (6) such that $100r\%$ of the space between α and β is given for the positive instances. Hence, if r is near to 1, more space is given to the positive class (priority class) than to the negative class in order to carry out a better generalization with the positive class than with the negative class.
- If $\alpha \geq \beta$, then by taking into account that $Re(b_i) \geq \frac{N_{pos}-(i-1)}{N_{pos}}$ and by imposing $r \leq \frac{N_{pos}-(i-1)}{N_{pos}}$, it is obtained that $i \leq 1 + N_{pos}(1 - r)$. Hence, the b_{ir} bias is considered, where $ir = \max\{i, i \leq 1 + N_{pos}(1 - r)\}$, that is, b_{ir} is the $(1 - r)$th q-quantile of $\mathcal{Z}_{pos}$.

  Clearly, b_{ir} exists since $\mathcal{Z}_{pos}$ is a non-empty and finite set. Furthermore, if $b > b_{ir}$, then $Re(b) \leq Re(b_{ir+1}) = \frac{N_{pos}-(ir+1-1)}{N_{pos}} < r$. Therefore, since the specificity function is an increasing function, the b_{ir} bias is a solution of the problem (6) for $0 \leq r \leq 1$.

The proof is completed. □

From the optimization problem (6), a mixed SVM is proposed, which is called Biased SVM (BSVM), where the solution of the problem (3) is tuned in order to obtain $Re(b_r) \geq r$, and specificity is maximized. 推理统计中的假设检验

It should be indicated that the proposed approach can be viewed as hypothesis testing from Inference Statistics [3], that is, the value

权衡的方法就是设置一个阈值，召回率大于阈值的时候，当精确率最大的时候，就是解

Table 3
Average figures of the results of the experiment with a linear kernel.

Datasets	Approach	100r	Recall	Specificity	Accuracy	g-Mean	Precision	f-Value
Iris (2)	SVM	***	79.35	66.07	70.35	72.41	53.90	64.20
	CS-SVM	***	68.21	81.49	72.64	**74.14**	**88.51**	**76.63**
		85	82.02	62.86	69.31	71.80	52.48	64.00
	BSVM	90	86.91	57.20	67.16	70.51	50.38	63.78
		95	92.37	49.11	63.49	67.35	47.58	62.80
Tae (1)	SVM	***	55.14	76.77	68.85	65.06	53.28	54.19
	CS-SVM	***	61.90	75.04	66.17	**67.76**	**84.05**	**70.93**
		85	82.78	39.98	53.63	57.53	39.85	53.80
	BSVM	90	88.24	30.51	49.23	51.89	37.89	53.01
		95	93.28	18.28	42.45	41.29	35.42	51.34
Glass (1)	SVM	***	67.91	77.75	74.33	72.66	**59.74**	63.56
	CS-SVM	***	81.90	65.47	70.84	72.93	53.92	64.75
		85	80.40	68.40	72.28	74.16	55.29	65.52
	BSVM	90	86.53	63.63	71.06	**74.20**	53.63	66.22
		95	92.42	58.18	69.34	73.33	51.79	**66.38**
Thyroid (3)	SVM	***	88.98	99.37	97.90	94.03	**95.82**	92.27
	CS-SVM	***	93.33	98.37	95.88	95.01	90.14	90.10
		85	93.70	98.64	97.89	96.14	91.78	92.73
	BSVM	90	94.25	98.93	98.29	96.56	93.46	**93.85**
		95	95.23	98.11	97.69	**96.66**	89.10	92.06
Ecoli (6)	SVM	***	81.73	91.42	89.76	86.44	63.97	71.77
	CS-SVM	***	91.61	87.35	88.01	**89.39**	57.70	70.53
		85	82.63	91.80	90.29	87.09	**65.26**	**72.92**
	BSVM	90	86.28	88.58	88.36	87.42	58.47	69.71
		95	94.49	64.90	69.59	78.31	33.41	49.37
Bupa (1)	SVM	***	45.46	83.81	67.70	61.73	**67.06**	54.19
	CS-SVM	***	65.01	66.08	65.63	**65.31**	58.38	61.31
		85	83.15	36.53	56.05	55.11	48.71	**61.43**
	BSVM	90	87.59	26.11	52.02	47.82	46.22	60.51
		95	92.93	16.24	48.52	38.85	44.58	60.25
Japanese (1)	SVM	***	44.37	68.71	60.94	55.21	**67.29**	53.47
	CS-SVM	***	54.33	58.70	57.34	**56.25**	37.27	44.10
		85	80.59	28.24	44.49	47.70	61.97	70.07
	BSVM	90	86.00	20.44	40.63	41.86	61.07	71.42
		95	91.78	11.93	36.67	32.83	60.19	**72.71**
Australian (1)	SVM	***	82.90	87.12	85.30	84.98	83.76	83.33
	CS-SVM	***	92.42	79.91	85.48	85.92	78.74	**85.00**
		85	81.71	88.91	85.71	85.23	**85.52**	83.57
	BSVM	90	87.43	84.62	85.87	**86.01**	82.00	84.63
		95	94.26	70.49	81.07	81.51	71.91	81.58
Pima (2)	SVM	***	54.13	84.83	73.93	67.76	**65.67**	59.34
	CS-SVM	***	70.65	77.78	75.29	**74.04**	63.16	**66.58**
		85	83.93	63.37	70.60	72.93	55.12	66.54
	BSVM	90	89.35	55.55	67.32	70.45	51.86	65.63
		95	94.12	43.47	61.13	63.96	47.16	62.83
German (2)	SVM	***	74.87	64.93	67.86	69.73	47.78	58.33
	CS-SVM	***	71.96	71.86	71.89	**71.86**	**52.36**	**60.56**
		85	82.22	58.06	65.28	69.09	45.65	58.37
	BSVM	90	87.75	48.34	60.16	65.13	42.13	56.92
		95	93.46	34.37	52.11	56.68	37.90	53.93
Bank (1)	SVM	***	42.43	90.87	85.25	62.09	**84.01**	56.38
	CS-SVM	***	66.92	82.97	81.12	74.49	33.89	44.97
		85	84.50	66.56	68.63	**75.00**	74.07	**78.94**
	BSVM	90	89.25	57.53	61.19	71.66	70.37	78.69
		95	94.42	40.87	47.05	62.13	64.35	76.54

*** Means white space.

$(1-r)$ in the BSVM approach is similar to the significance level α in hypothesis testing.

Therefore, given an r value such that $0 \leq r \leq 1$, then the classifier $f \in \mathcal{F}(\mathrm{w}_0)$ considered is as follows:

$$f(x) = f_{b_r}(x) = \sum_{i=1}^{N} \gamma_i y_i K(x_i, x) - b_r = \langle \mathrm{x}, \mathrm{w}_0 \rangle - b_r$$

where

$$b_r = \begin{cases} b_r^* & \text{if } \alpha < \beta, \\ b_{ir} & \text{otherwise.} \end{cases} \tag{7}$$

Given a set of classifiers $\mathcal{F}(\mathrm{w}_0) = \{f_b(x) = \langle \mathrm{x}, \mathrm{w}_0 \rangle - b,\ b \in \mathbb{R}\}$ on a training set $\mathcal{Z}$, the influence of the threshold on the performance of the BSVM can be analysed: for $0 \leq r \leq r' \leq 1$, the $b_{r'}$ bias is a solution of (6) for r' and verifies that $Re(b_{r'}) \geq r' \geq r$, and therefore $Sp(b_{r'}) \leq Sp(b_r)$, where b_r is a solution of (6) for r. Furthermore, since $Sp(b)$

Table 4
Average figures of the results of the experiment for $r = 0.85$, 0.90, and 0.95 with a linear kernel for the Iris, Tae, Glass, Japanese, and German datasets by changing the priority between classes.

Datasets	Approach	100r	Recall	Specificity	Accuracy	g-Mean	Precision	f-Value
Iris (2)*	SVM	***	66.33	79.36	72.85	72.55	86.54	75.10
	CS-SVM	***	80.85	68.81	72.76	**74.07**	**88.17**	76.88
		85	82.70	55.76	69.23	67.91	78.90	80.75
	BSVM	90	88.28	41.26	64.77	60.35	75.04	**81.12**
		95	93.25	25.62	59.44	48.88	71.49	80.93
Tae (1)*	SVM	***	78.00	52.49	65.25	63.99	83.24	64.38
	CS-SVM	***	75.42	61.68	66.25	**67.89**	**84.29**	70.96
		85	83.42	45.17	64.30	61.38	76.00	79.54
	BSVM	90	88.36	34.40	61.38	55.13	73.71	80.37
		95	93.25	21.53	57.39	44.81	71.21	**80.75**
Glass (1)*	SVM	***	77.76	67.90	72.83	72.66	86.26	75.99
	CS-SVM	***	65.32	81.67	70.66	72.74	**88.30**	74.79
		85	81.49	65.14	73.32	**72.86**	82.78	82.13
	BSVM	90	86.54	50.77	68.66	66.28	78.34	**82.23**
		95	92.30	29.06	60.68	51.79	72.80	81.40
Japanese (2)*	SVM	***	68.17	45.19	60.76	55.51	64.35	66.20
	CS-SVM	***	58.48	54.50	57.25	**56.26**	**74.02**	65.34
		85	82.53	30.26	66.28	49.97	63.20	71.58
	BSVM	90	88.27	20.39	67.19	42.43	61.67	72.61
		95	93.86	10.34	67.90	31.15	60.30	**73.43**
German (2)*	SVM	***	67.72	71.67	69.00	69.79	50.60	57.92
	CS-SVM	***	71.87	72.01	71.81	**71.88**	**85.73**	**78.15**
		85	82.76	58.96	75.62	69.85	46.36	59.43
	BSVM	90	88.23	48.58	76.32	65.47	42.37	57.25
		95	93.84	31.01	74.98	53.95	36.83	52.90

*** Means white space.

is an increasing function of b, then $b_r \geq b_{r'}$ and, since $Re(b)$ is a decreasing function of b, then $Re(b_r) \leq Re(b_{r'})$.

Therefore, if $0 \leq r \leq 1$, then $Re(b_r)$ is an increasing function of r, and $Sp(b_r)$ is a decreasing function of r on the training set $\mathcal{Z}$. This result can be seen clearly in the experimentation given in the following section.

5. Experimentation

To the best our knowledge, there is no machine-learning approach for classification problems with an associated optimization problem similar to the problem proposed (6). Learning machines tend to strive to maximize the recall and specificity metrics simultaneously. No lower bound is determined for the recall measure on the training set. Hence, a comparative is carried out with the standard SVM and the cost-sensitive SVM in order to analyse the behaviour of the BSVM.

Experimentation is conducted on the following standard UCI datasets [2]: **Iris** plants, **T**eaching **a**ssistant **e**valuation, **Glass** identification database, **Thyroid** disease, protein localization sites (**Ecoli**), **Bupa** liver disorders, **Japanese** credit screening, **Australian** credit approval, **Pima** Indians diabetes, **German** credit data, and **Bank** marketing.[4] A summary of the characteristics and a brief description of these data sets is shown in Table 2.

In order to apply the BSVM, datasets are split in the form of one class (the priority class), indicated in the first column, versus the rest of the classes. The priority class has been chosen with a lower number of examples than the non-priority class, that is, there is an imbalanced ratio between positive and negative instances [17,19]. This selection is considered in order to better illustrate the performance of the BSVM.

Three experiments are considered:

- In the first experiment, the performance of the BSVM with a lineal kernel is analyzed.
- Next, a similar approach is carried out with a lineal kernel where the priority is changed with respect to the former experiment in several datasets.
- In the last experiment, the rbf kernel is used instead of the linear kernel.

The first experiment is carried out by following a similar experimental framework to that used in [1] as suggested in [15]. Performance is evaluated on SVM classifiers using the linear kernel, which is chosen as a baseline for the empirical evaluation, and the regularization term C is explored on a one-dimensional grid with the following values: $C = [2^{-4}, 2^{-3}, \ldots, 2^{9}, 2^{10}]$. A cost-sensitive SVM, denoted by CS-SVM, is employed to train an SVM on the listed UCI datasets. The Matlab Bioinformatic toolbox is used where the values for C^+ and C^- in (4) are calculated from C as,

$$C^+ = \frac{C \cdot N}{2N_{pos}}, \quad C^- = \frac{C \cdot N}{2N_{neg}}.$$

The criterion employed for the estimation of the generalized accuracy in the standard SVM and the cost-sensitive SVM, and the specificity in the BSVM is the three-fold ($N_{fold} = 3$) cross-validation on the whole set of training data. This procedure is repeated 100 times in order to ensure good statistical behaviour. The values considered for r, the minimum true positive rate in the positive class for the training set, are set to 0.85, 0.90, and 0.95, following a similar experimental framework for hypothesis testing. It is worth noting that optimization problems (3), (4) and (6) are carried out on the training set, whereas results are given on the test set, hence recall values on the test set are not always greater or equal to the selected r bound when training the BSVM.

The results obtained, given as percentages, are displayed in Table 3. Several conclusions can be drawn from this empirical experimentation:

[4] For this dataset only the quantitative features have been considered.

Table 5
Average figures of the results of the experiment with an rbf kernel.

Datasets	Approach	100*r*	Recall	Specificity	Accuracy	*g*-Mean	Precision	*f*-Value
Iris (1)	SVM	***	95.20	96.01	95.69	95.60	92.27	93.71
	CS-SVM	***	96.06	96.37	96.27	**96.15**	93.40	**94.51**
		85	85.29	97.44	93.35	91.16	94.34	89.59
	BSVM	90	89.48	97.69	94.90	93.50	**95.10**	92.20
		95	95.08	87.58	90.03	91.26	79.29	86.47
Tae (1)	SVM	***	45.19	85.81	72.16	**62.16**	60.48	51.73
	CS-SVM	***	38.62	97.45	78.36	60.82	**88.58**	53.00
		85	79.90	47.49	57.97	61.60	42.23	**55.25**
	BSVM	90	84.35	32.20	49.09	52.12	37.41	51.83
		95	89.28	22.60	44.17	44.93	35.66	50.96
Glass (1)	SVM	***	70.69	90.06	83.60	**79.79**	**77.56**	**73.96**
	CS-SVM	***	82.01	75.73	79.97	78.61	67.71	71.14
		85	86.73	59.81	68.34	72.02	51.20	64.39
	BSVM	90	89.36	44.82	59.19	63.28	44.05	59.01
		95	93.18	52.37	65.64	69.86	48.74	64.00
Thyroid (3)	SVM	***	89.22	98.89	97.53	93.93	92.89	91.02
	CS-SVM	***	91.73	98.75	97.77	**94.96**	**93.05**	91.77
		85	86.96	98.80	97.17	92.69	92.17	89.49
	BSVM	90	89.67	98.76	97.53	94.11	92.17	90.90
		95	94.78	98.35	97.87	96.55	90.31	**92.49**
Ecoli (6)	SVM	***	85.76	97.66	95.75	**91.52**	**87.23**	**86.49**
	CS-SVM	***	83.31	96.94	94.80	89.75	84.06	83.36
		85	85.18	90.32	89.49	87.71	62.11	71.84
	BSVM	90	89.51	75.67	77.85	82.30	40.68	55.94
		95	94.28	73.87	77.10	83.46	40.21	56.38
Bupa (1)	SVM	***	48.57	85.17	69.61	64.32	**70.36**	57.47
	CS-SVM	***	67.17	74.06	71.17	**70.39**	65.47	**66.14**
		85	83.88	28.92	51.93	49.25	46.11	59.50
	BSVM	90	87.30	35.07	56.98	55.33	49.36	63.07
		95	96.75	7.88	45.26	27.61	43.23	59.76
Japanese (1)	SVM	***	02.56	98.04	68.34	15.85	**37.14**	04.80
	CS-SVM	***	00.33	98.48	67.95	02.26	07.73	00.55
		85	72.55	27.04	41.11	44.29	30.97	43.41
	BSVM	90	76.77	29.31	44.00	**47.44**	32.88	46.04
		95	84.79	19.98	40.09	41.16	32.34	**46.82**
Australian (1)	SVM	***	92.26	79.81	85.33	85.81	78.55	84.86
	CS-SVM	***	92.45	80.55	85.84	86.27	79.28	**85.33**
		85	81.13	89.39	85.72	85.16	**85.97**	83.48
	BSVM	90	88.29	84.58	86.24	**86.41**	82.11	85.08
		95	94.27	72.70	82.31	82.79	73.46	82.58
Pima (2)	SVM	***	61.22	81.57	74.45	70.67	**64.03**	62.60
	CS-SVM	***	68.94	76.83	74.08	**72.67**	61.59	64.92
		85	83.56	63.18	70.31	72.66	54.88	**66.25**
	BSVM	90	88.90	56.24	67.65	70.71	52.13	65.72
		95	94.12	45.70	62.60	65.58	48.16	63.72
German (2)	SVM	***	28.55	93.05	73.66	51.55	**63.77**	39.45
	CS-SVM	***	65.90	75.03	72.29	**70.25**	53.20	**58.79**
		85	85.14	40.68	54.03	58.85	38.09	52.63
	BSVM	90	89.89	33.40	50.32	54.79	36.65	52.07
		95	96.48	07.49	34.19	26.88	30.89	46.80
Bank (1)	SVM	***	00.66	99.61	88.21	08.14	18.07	01.28
	CS-SVM	***	00.23	99.68	88.22	03.71	11.95	00.45
		85	79.27	54.77	57.61	**65.89**	**18.58**	**30.11**
	BSVM	90	87.07	43.30	48.32	61.40	16.67	27.98
		95	93.40	25.61	33.42	48.91	14.06	24.43

*** Means white space.

- It is observed that specificity values are greater than recall values in the standard SVM (except for the Iris, Australian and German datasets). As mentioned earlier, the standard SVM is inherently biased towards the majority class when classifying imbalanced datasets. This is the main difference between SVM and BSVM, since BSVM focuses on the priority class regardless of whether it is the majority class.
- Due to the generalization capacity of BSVM, recall values on the test set are about 85%, 90% and 95% for all datasets, respectively. Hence, empirical results show that BSVM produces a good hypothesis in terms of recall, by using the optimization problem of the same SVM. It should be noted that recall value is greater (near to 95% of *r*) for the Thyroid dataset, since this optimization problem is linearly separable and the chosen bias is b_r^* (7).
- If the recall value is low for the initial SVM, then the 'price to pay' in order to improve this metric in the BSVM is a significant reduction in the specificity. This reduction can be observed in the Tae, Bupa, Japanese, Pima, and Bank datasets. Nevertheless, this does not constitute a drawback since BSVM is specifically designed for problems in which this balance is not a critical point.

- The values for recall using BSVM are always greater than those using standard SVM, for all datasets except in the Australian dataset for BSVM with $r = 0.85$. In this case, it can be seen that the value of specificity is greater with BSVM than SVM.
- It is clear that cost-sensitive SVM is better than SVM for the increase in the recall metric (except for the Iris and German datasets). Nevertheless, the value for recall in certain datasets (Iris, Tae, Bupa, Japanese, Pima, German, and Bank) can be considered poor in comparison with BSVM.
- The best result for *g*-mean, precision and *f*-value metrics in each dataset is indicated in bold in Table 3. As expected, BSVM produces the optimal hypothesis in terms of recall without penalizing too much other criteria, so that there is no model which is the best with respect to all metrics. Hence, when the objective is the recall metric, then BSVM with a high value of r is the best option in all cases.

Let us now turn our attention to the case where the priority between classes is changed. Thus, in order to observe how priority affects metrics, a similar experiment to that reported previously is carried out with the Iris, Tae, Glass, Japanese and German datasets.[5] The results can be observed in Table 4, where symbol "*" denotes that the considered class is a non-priority class and the rest of the classes provide the priority class.

An initial observation, the priority class corresponds to the class with the highest number of examples, that is, the experiment presented is not specially addressed to manage imbalanced datasets. A number of conclusions can be drawn:

- Results obtained with standard SVM and cost-sensitive SVM differ from those in Table 3 since the training sets provided by the cross-validation procedure are different. However, the performance values are very similar, with no significant differences between them, that is, the change of priority between classes exerts no noticeable effect.
- Values for specificity when using BSVM are smaller than or equal to those provided in Table 3 (except for the Japanese dataset) since the majority class is now the positive class.
- In certain cases (Glass, Japanese, and German datasets), accuracy using BSVM is better than for standard SVM because the size of the region in the feature space $\mathcal{H}$ has been increased for the majority class (see Section 3) and therefore the number of successes increases.
- With respect to the *g*-mean, precision and *f*-value metrics, results are qualitatively similar to those depicted in Table 3.

Finally, a third experiment is carried out on SVM classifiers using the rbf kernel with σ (the width of rbf) and C being explored on a two-dimensional grid: $\sigma = [2^{-4}, 2^{-3}, \ldots, 2^3, 2^4]$ and $C = [2^{-4}, 2^{-3}, \ldots, 2^9, 2^{10}]$.

The criterion employed for the estimation of the generalized accuracy in the standard SVM and the cost-sensitive SVM, and the proximity[6] between the recall and the threshold r in the BSVM is that of the three-fold ($N_{fold} = 3$) cross-validation on the whole set of training data. This procedure is repeated 100 times in order to ensure good statistical behaviour.

The results obtained, given as percentages, are displayed in Table 5. Certain conclusions can be drawn from this empirical experimentation:

- As it is well-known, the accuracy with an rbf kernel is usually better than that with a linear kernel in standard and cost-sensitive SVM. With respect to the recall metric, sometimes it is better with a linear kernel (Tae, Japanese, German, and Bank datasets). With respect to the specificity metric, except for German datasets, the rbf SVM is better than the linear SVM.
- It can be seen that, with the Iris, Thyroid Ecoli and Australian datasets, the value of the recall metric is high for all the approaches considered. Nevertheless, BSVM is currently improving the other approaches for the Japanese and Bank datasets.
- It can be observed that using the rbf kernel, the *g*-mean, precision and *f*-value performance metrics for standard SVM have improved with respect to the linear kernel.
- With respect to the BSVM, it can be seen that the same conclusions are obtained with the rbf kernel and with the linear kernel.

6. Conclusions

A post-processing technique for Support Vector Machines, called BSVM, has been introduced to deal with datasets where one class, the positive class, is considered more relevant than another class in binary classification problems. Hence, BSVM is designed for the case when it is non-critical to increase the true positive ratio in exchange for an increase in the false positive rate.

By modifying the bias on the standard SVM, the aim of BSVM is, given a fixed minimum value for recall, to maximize specificity on the training set.

Empirical results show that due to the generalization capacity of SVMs, recall on the test set in the BSVM is close to the fixed level r, that is, BSVM produces a good hypothesis in terms of recall.

Furthermore, the algorithm presented can easily be applied in practice in order to address to real-world complexities, such as those faced by practitioners with an interest in benefitting management.

Acknowledgements

This research is partially supported by the projects of the Spanish Ministry of Economy and Competitiveness HERMES (TIN2013-46801-C4-1-R), PATRICIA (TIN2012-38416-C03-01) and Simon (TIC-8052) of the Andalusian Regional Ministry of Economy, Innovation and Science. Haydemar Nuñez acknowledges the I3 grant has been partly supported by the PG-03-7678-2009/1 project from the Central University of Venezuela. Yenny Leal was supported by the Agency for Management of University and Research Grants of the Government of Catalonia, Spain (Beatriu de Pinós [BP-DGR-2013]).

References

[1] C. Angulo, F. Ruiz, L. González, J.A. Ortega, Multi-classification by using tri-class SVM, Neural Proc. Lett. 23 (1) (2006) 89–101.
[2] C.L. Blake, C.J. Merz, UCI Repository of Machine Learning Databases, 1998.
[3] G. Casella, R. Berger, Statistical Inference Duxbury Resource Center, 2001.
[4] G. Cohen, M. Hilario, H. Sax, S. Hugonnet, A. Geissbuhler, Learning from imbalanced data in surveillance of nosocomial infection, Artif. Intell. Med. 37 (May) (2006) 7–18.
[5] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University press, 2000.
[6] R.S. Debreceny, G.L. Gray, Data mining journal entries for fraud detection: an exploratory study, Int. J. Acc. Inf. Syst. 11 (3) (2010) 157–181.
[7] C. Elkan, The foundations of cost-sensitive learning, in: Proceedings of the 17th International Joint Conference on Artificial Intelligence – vol. 2, IJCAI'01, Morgan Kaufmann Publishers Inc., 2001, pp. 973–978.
[8] T. Fawcett, An introduction to ROC analysis, Pattern Recognit. Lett. 27 (2006) 861–874.
[9] L. González, C. Angulo, F. Velasco, A. Català, Unified dual for bi-class SVM approaches, Pattern Recognit. 38 (10) (2005) 1772–1774.
[10] L. González, C. Angulo, F. Velasco, A. Català, Dual unification of bi-class support vector machine formulations, Pattern Recognit. 39 (7) (2006) 1325–1332.

[5] These datasets are chosen since their specificity values in Table 3 are the lowest for the standard SVM.

[6] This criterion has been used since it is well-known that the rbf kernel in certain cases leads to overfitting on the training set.

[11] L. Gonzalez-Abril, C. Angulo, F. Velasco, J.A. Ortega, A note on the bias in SVMs for multiclassification, IEEE Trans. Neural Netw. 19 (4) (2008) 723–725.
[12] L. Gonzalez-Abril, H. Nu nez, C. Angulo, F. Velasco, GSVM: an SVM for handling imbalanced accuracy between classes in bi-classification problems, Appl. Soft Comput. 17 (2014) 23–31.
[13] L. Gonzalez-Abril, F. Velasco, J.A. Ortega, L. Franco, Support vector machines for classification of input vectors with different metrics, Comput. Math. Appl. 61 (9) (2011) 2874–2878.
[14] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (September (9)) (2009) 1263–1284.
[15] C.-W. Hsu, C.-J. Lin, A comparison of methods for multiclass support vector machine, IEEE Trans. Neural Netw. 13 (March (2)) (2002) 415–425.
[16] M. Karabatak, M. Cevdet, Ince An expert system for detection of breast cancer based on association rules and neural network, Expert Syst. Appl. 36 (2, Part 2) (2009) 3465–3469.
[17] B. Krawczyk, Learning from imbalanced data: open challenges and future directions, Progr. Artif. Intell. 5 (4) (2016) 221–232.
[18] W.Y. Lin, Y.H. Hu, C.F. Tsai, Machine learning in financial crisis prediction: a survey, IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) 42 (4) (2012) 421–436.
[19] V. López, A. Fernández, S. García, V. Palade, F. Herrera, An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics, Inf. Sci. 250 (2013) 113–141.
[20] S. Moro, P. Cortez, P. Rita, A data-driven approach to predict the success of bank telemarketing, Decis. Support Syst. 62 (2014) 22–31.
[21] D.L. Olson, D. Delen, Y. Meng, Comparative analysis of data mining models for bankruptcy prediction, Decis. Support Syst. 52 (2) (2012) 464–473.
[22] L. Oneto, S. Ridella, D. Anguita, Tikhonov, Ivanov and Morozov regularization for support vector machine learning, Mach. Learn. 103 (1) (2016) 103–136.
[23] R.B. Rao, S. Krishnan, R.S. Niculescu, Data mining for improved cardiac care, ACM SIGKDD Explor. Newsl. 8 (1) (2006) 3–10.
[24] P. Ravisankar, V. Ravi, G. Raghava Rao, I. Bose, Detection of financial statement fraud and feature selection using data mining techniques, Decis. Support Syst. 50 (2) (2011) 491–500.
[25] V. Vapnik, Statistical Learning Theory, John Wiley & Sons, New York, 1998.
[26] V. Vapnik, The Nature of Statistical Learning Theory (Information Science and Statistics), Springer, 1999, November.
[27] Z. Yang, W. You, G. Ji, Using partial least squares and support vector machines for bankruptcy prediction, Expert Syst. Appl. 38 (7) (2011) 8336–8342.
[28] J.S. Yazdi, F. Kalantary, H.S. Yazdi, Prediction of liquefaction potential based on cpt up-sampling, Comput. Geosci. 44 (2012) 10–23.
[29] J.S. Yoon, Y.S. Kwon, A practical approach to bankruptcy prediction for small businesses: substituting the unavailable financial data for credit card sales information, Expert Syst. Appl. 37 (5) (2010) 3624–3629.
[30] M. Zieba, J.M. Tomczak, M. Lubicz, J. Swiatek, Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients, Appl. Soft Comput. 14 (Part A) (2014) 99–108.