Pattern Recognition Letters 34 (2013) 344–348

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

A study on output normalization in multiclass SVMs

L. Gonzalez-Abril [a,*], F. Velasco [a], C. Angulo [b], J.A. Ortega [c]

[a] Applied Economics I Dept., Seville University, 41018 Seville, Spain
[b] Grup de Recerca en Enginyeria del Coneixement, Universitat Politècnica de Catalunya, 08800 Vilanova i la Geltrú, Spain
[c] Computer Languages and Systems Dept., Seville University, 41012 Seville, Spain

ARTICLE INFO

Article history:
Received 11 May 2012
Available online 22 November 2012

Communicated by S. Sarkar

Keywords:
1-v-r SVM
Convex hull
Kernel methods
Multiclassification

ABSTRACT

The use of binary support vector machines (SVMs) in multi-classification is addressed in this paper. Margins associated to the bi-classifiers, since they depend on the geometrical disposition of the classes being separated, are, in general, of various magnitudes. In order to overcome this scaling problem, a normalization process should be applied on the SVMs' outputs. Thus, a new normalization approach is presented based on the convex hulls that contain the classes to be separated. Furthermore, a theoretical study is developed which justifies the proposed approach, and an interpretation is provided. An empirical study is also carried out to compare this normalization with others found in the literature.

1. Introduction

SVMs are learning machines which implement the structural risk minimization inductive principle to obtain good generalization on a limited number of learning patterns (Vapnik, 1998). This theory 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 VC dimension, which provides good generalization performance, as it has been demonstrated in several applications (Cristianini and Shawe-Taylor, 2000).

The extension of binary classification to multi-classification is an on-going research issue (Mayoraz and Alpaydin, 1999; Angulo et al., 2006; Wang et al., 2008). Although some joint SVM methods exist, the binary ad hoc methods of K one-versus-rest (1-v-r) or $K(K-1)/2$ one-versus-one SVMs for the solution of the multi-class problem still prevail due, in general, to their good performance and manageable optimization.

In standard SVM formulation, the output scale is determined such that outputs for the support vectors are ±1. Therefore, a direct comparison of the output of different SVMs working on a multi-classification problem is inadequate because scaling varies for each machine considered. Some kind of normalization is therefore crucial for the comparison of outputs of different SVMs.

The usual procedure to circumvent this scaling problem is the direct comparison of the real-valued outputs for each machine. It is argued in this work, however, that the scaling problem cannot be circumvented but must be tackled by considering some kind of normalization. Those available in the literature are analyzed, and a new output normalization method is proposed based on the convex hulls that contain the classes to be separated. More precisely, the proposed normalization is based on the problem of finding the nearest point between reduced convex hulls. Furthermore, the reliability of the standard SVM is taken into account in order to carry out the normalization.

The remainder of this paper is arranged as follows: Section 2 presents the standard SVM approach. Section 3 puts forward two output-normalization schemes based on different criteria, and the proposed normalization is developed. An experiment is carried out in Section 4 in order to show the accuracy rate of several normalization processes. Finally, conclusions are drawn.

2. Standard SVM approach

Let $\mathcal{Z} = \{z_i\}_{i=1}^n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a training set, with $x_i \in \mathcal{X}$ as the input space and $y_i \in \mathcal{Y} = \{\theta_1, \theta_2\} = \{+1, -1\}$ the output space. Let $\phi : \mathcal{X} \to \mathcal{F}$ be a feature mapping with a dot product denoted by $\langle \cdot, \cdot \rangle$. A linear classifier $f_w(x) = \langle x, w \rangle + b$ is sought in $\mathcal{F}$, with $b \in \mathbb{R}$. Outputs are obtained in the form $h_w(x) = \text{sign}(f_w(x))$.

For the standard primal SVM 2-norm formulation (Vapnik, 1998), the optimization problem becomes

* Corresponding author.
E-mail addresses: luisgon@us.es (L. Gonzalez-Abril), velasco@us.es (F. Velasco), cecilio.angulo@upc.edu (C. Angulo), ortega@lsi.us.es (J.A. Ortega).

[1] A generalization is given in (Gonzalez-Abril et al., 2011).

0167-8655/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2012.11.003

$$\min_{w\in\mathcal{F},b\in\mathbb{R}} \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$
$$\text{s.t. } y_i(\langle x_i, w\rangle + b) + \xi_i \geqslant 1, \quad \xi_i \geqslant 0,\ z_i \in \mathcal{Z} \tag{1}$$

where C is the regularization term and ξ_i are slack variables. The solution can be written as

$$w = \sum_i \alpha_i y_i x_i \tag{2}$$

where α_i are Lagrange multipliers for the dual problem of (1). Furthermore,

$$\sum_i \alpha_i y_i = 0 \tag{3}$$
$$0 \leqslant \alpha_i \leqslant C, \quad i = 1,\ldots,n \tag{4}$$
$$\alpha_i(y_i(\langle x_i, w\rangle + b) - 1 + \xi_i) = 0 \quad i = 1,\ldots,n \tag{5}$$

A vector x_i is called a support vector (SV) when $\alpha_i \neq 0$. Term b is calculated a posteriori (Gonzalez-Abril et al., 2008). The classifier can be written as

$$f(x) = \sum_i \alpha_i y_i \langle x_i, x\rangle + b \tag{6}$$

Note that if α_i^+ and α_j^- are multipliers associated to the vectors of $\mathcal{Z}_{(+)} = \{(x_i, y_i) \in \mathcal{Z},\ y_i = +1\}$ and $\mathcal{Z}_{(-)} = \{(x_j, y_j) \in \mathcal{Z},\ y_j = -1\}$, respectively, then $\sum_i \alpha_i^+ = \sum_j \alpha_j^- > 0$, since otherwise $w = 0$.

3. Normalization of the SVM outputs

There are several alternatives, further to that which considers the signed output, which tackle the scaling problem when comparing the output of different SVMs. Let us consider the two most used approaches:

MS normalization: This normalization, which is called *MS* (for 'Mayoraz scheme'), was introduced in (Mayoraz and Alpaydin, 1999). It is obtained by dividing the classifier f from (6) by the norm of the solution vector w, that is, the scalar factor is $\lambda^M = \frac{1}{\|w\|}$. Hence, the new classifier is $f^M(x) = \frac{1}{\|w\|} f(x)$. This approach presents a geometrical interpretation since the norm of the normal vector of the classifiers f^M is one.

SRM normalization: An approach, called *SRM* (static reliability measure), was given in (Liu and Zhang, 2005). It takes into account the value of the objective function of (1). $\lambda^R = \exp\{-(\frac{1}{2}\|w\|^2 + C\sum \xi_i)/(Cn)\}$ is considered as the scalar factor in this approach. Hence, the normalized classifier is rewritten in the form $f^R(x) = \lambda^R f(x)$.

The main concept behind both the normalization procedures above, *MS* and *SRM*, is the same: since the classifier with a smaller $\|w\|$, which corresponds to a larger margin, is considered to be more accurate in generalization, then $\frac{1}{\|w\|}$ and λ^R, as reliability measures, should be larger. Nevertheless, there is not theoretical result which provides a good explanation for these scalar factors.

A new normalization procedure, denoted by *CH* (convex hull), is introduced here, and is based on the location of the nearest point between the reduced convex hulls of $\mathcal{Z}_{(+)}$ and $\mathcal{Z}_{(-)}$ (Bennett and Bredensteiner, 2000; Bredensteiner and Bennett, 1999). The geometrical interpretation for this point between convex hulls can be seen in (Bredensteiner and Bennett, 1999).

Let us develop this normalization. The solution for (1) can be written González et al., 2006 as $w = \lambda w^*,\ b = \lambda b^*$, for $\lambda = \sum_i \alpha_i^+$ and $w^* = \sum_i u_i^* x_i - \sum_j v_j^* x_j$, with $\{u_i^*\}$ and $\{v_j^*\}$ as the solution of the dual problem

$$\min_{\mathbf{u},\mathbf{v}} \frac{1}{2}\left\|\sum_i u_i x_i - \sum_j v_j x_j\right\|^2$$
$$\text{s.t.} \quad \sum_i u_i = \sum_j v_j = 1, \quad 0 \leqslant u_i, v_j \leqslant \frac{C}{\lambda} \tag{7}$$
$$z_i \in \mathcal{Z}_{(+)}, \quad z_j \in \mathcal{Z}_{(-)}$$

From (6), the classifier can be written as

$$f(x) = \lambda\left(\sum_i u_i^* \langle x_i, x\rangle - \sum_j v_j^* \langle x_j, x\rangle + b^*\right) \tag{8}$$

The λ value is such that the margin between the reduced convex hull of $\mathcal{Z}_{(+)}, R\mathcal{Z}_{(+)} = \{\sum_i u_i x_i,\ \sum_i u_i = 1,\ 0 \leqslant u_i \leqslant \frac{C}{\lambda}, z_i \in \mathcal{Z}_{(+)}\}$, and $\mathcal{Z}_{(-)}, R\mathcal{Z}_{(-)} = \{\sum_j v_j x_j, \sum_i v_j = 1, 0 \leqslant v_j \leqslant \frac{C}{\lambda}, z_j \in \mathcal{Z}_{(-)}\}$, is non-zero. In this form, the optimization problem (7) is prevented from collapsing into a trivial solution.

Note that a large value for λ involves small reduced convex hulls $R\mathcal{Z}_{(+)}$ and $R\mathcal{Z}_{(-)}$, and hence the number of support vectors in the solution w will be high. Since $\sum_i \alpha_i^+ = \sum_j \alpha_j^- = \lambda$, the λ value can be interpreted from (7) as the strength that support vectors must attain in order to obtain good accuracy in generalization.

A different expression for λ can be obtained from the constraints in the optimization problem (1). From (5), if $\alpha_i \neq 0$, then $y_i(\langle x_i, w\rangle + b) - 1 + \xi_i = 0$, and by considering (2) and (3),

$$\begin{aligned} 0 &= \sum_i \alpha_i (y_i(\langle x_i, w\rangle + b) - 1 + \xi_i) \\ &= \left\langle \sum_i \alpha_i y_i x_i, w \right\rangle + b\sum_i \alpha_i y_i - \sum_i \alpha_i + \sum_i \alpha_i \xi_i \\ &= \langle w, w\rangle + 0 - \sum_i \alpha_i^+ - \sum_j \alpha_j^- + \sum_i \alpha_i \xi_i = \|w\|^2 - 2\lambda + \sum_i \alpha_i \xi_i \end{aligned}$$

Therefore,

$$\lambda = \frac{1}{2}\left(\|w\|^2 + \sum_i \alpha_i \xi_i\right) \tag{9}$$

Note that if the problem (1) is separable then, from (9), $\lambda = \frac{1}{2}\|w\|^2$, since $\xi_i = 0$ for all i.

Let $N_{(+)}^{SV} = \#\{\alpha_i^+,\ \alpha_i^+ \neq 0\}$ (the number of SVs for the positive class), $N_{(-)}^{SV} = \#\{\alpha_i^-,\ \alpha_i^- \neq 0\}$ (the number of SVs for the negative class), and $N^{SV} = N_{(+)}^{SV} + N_{(-)}^{SV}$ (the total number of SVs). From (4), $\lambda = \sum_i \alpha_i^+ \leqslant C \cdot N_{(+)}^{SV}$, $\lambda = \sum_i \alpha_i^- \leqslant C \cdot N_{(-)}^{SV}$, and $\sum_i \alpha_i \xi_i \leqslant C \cdot \sum_i \xi_i$. Hence, lower and upper bounds for the value of λ can be given:

$$\frac{1}{2}\|w\|^2 \leqslant \lambda \leqslant \min\left\{C \cdot N_{(+)}^{SV}, C \cdot N_{(-)}^{SV}, \frac{1}{2}\left(\|w\|^2 + C \cdot \sum_i \xi_i\right)\right\}$$

Finally, by applying $\min\left\{N_{(+)}^{SV}, N_{(-)}^{SV}\right\} \leqslant \frac{1}{2}N^{SV}$, a different and more manageable upper bound can be obtained:

$$\frac{1}{2}\|w\|^2 \leqslant \lambda \leqslant \frac{C}{2}N^{SV}$$

Table 1
Characteristics of the selected data sets.

Data set	Patterns	Classes	Features	Data per classes
Iris	150	3	4	All 50
Tae	151	3	5	49, 50, 52
Wine	178	3	13	59, 71, 48
Glass	214	6	9	70, 76, 17, 13, 9, 29
Thyroid	215	3	5	150, 35, 30
Ecoli	330	6	8	141, 77, 52, 35,20,5
Dermat	358	6	34	111, 60, 71, 48, 48, 20
Vowel	990	11	11	All 90
Segment	2310	7	19	All 330

Table 2
Results of the experiment where the best mean accuracy rates, its standard deviation and C-parameter are presented.

Data set^{R,N}	*SS*	*MS*	*SRM*	*CH*
Iris^{50,10}	94.85 ± 0.537 (2^{10})	91.72 ± 0.722 (2^{2})	95.48 ± 0.524 (2^{10})	**96.96** ± 0.446 (2^{1})
Tae^{50,10}	48.05 ± 1.223 (2^{2})	46.75 ± 1.169 (2^{2})	48.05 ± 1.223 (2^{2})	**48.08** ± 1.214 (2^{2})
Wine^{50,10}	**97.94** ± 0.322 (2^{-1})	97.89 ± 0.330 (2^{0})	97.93 ± 0.324 (2^{-1})	97.71 ± 0.354 (2^{-1})
Glass^{30,10}	57.86 ± 1.102 (2^{3})	54.57 ± 1.140 (2^{2})	58.10 ± 1.080 (2^{3})	**58.18** ± 1.108 (2^{3})
Thyroid^{30,10}	96.11 ± 0.448 (2^{9})	95.21 ± 0.467 (2^{3})	95.92 ± 0.456 (2^{10})	**96.19** ± 0.402 (2^{5})
E. coli^{25,5}	85.59 ± 0.485 (2^{-1})	**85.75** ± 0.454 (2^{-1})	85.70 ± 0.480 (2^{-1})	85.53 ± 0.461 (2^{-1})
Dermat^{25,5}	96.40 ± 0.192 (2^{0})	**97.10** ± 0.210 (2^{3})	96.53 ± 0.192 (2^{0})	96.99 ± 0.210 (2^{3})
Vowel^{10,2}	**54.81** ± 0.270 (2^{0})	44.07 ± 0.300 (2^{0})	54.79 ± 0.268 (2^{0})	54.58 ± 0.284 (2^{0})
Segment*	89.64 ± 0.174 (2^{3})	90.21 ± 0.117 (2^{3})	89.73 ± 0.169 (2^{3})	**91.22** ± 0.115 (2^{3})

The value of C is given a priori in the problem (1), therefore, as aforementioned, a large value for λ results in a high number of support vectors N^{SV}. Similarly, looking into the lower bound, a small value for λ implies that the margin separating $\mathcal{Z}_{(+)}$ and $\mathcal{Z}_{(-)}$, $\frac{2}{\|w\|^2}$, is large. Hence, the solution will provide good generalization performance as well as being smooth (small VC-dimension), and therefore its reliability is better than for a sharp solution (λ value is high).

Hence, the proposed approach will consider $\lambda^{CH} = \frac{1}{\lambda}$ as the scalar factor. The classifier $f^{CH}(x) = \lambda^{CH} f(x)$ is considered, that is, from (8),

$$f^{CH}(x) = \sum_i u_i^* \langle x_i, x\rangle - \sum_j v_j^* \langle x_j, x\rangle + b^*$$

Unlike the previously considered normalization procedures, the f^{CH} function has a specific meaning from the solution of a dual problem (7) plus a bias. This normalization agrees with Bayesian arguments advocating that smoother solutions should be given higher-valued weights.

The schemes used for multiclassification represent a typical illustration of the problem where the comparison of the output of different SVMs is carried out. Thus, in the following section, a study on the use of the proposed normalization procedure when the 1-v-r SVM scheme is used for multiclassification, is presented since this is the most widely used implementation (Rifkin and Klautau, 2004).

4. Normalization in the 1-v-r SVM scheme

Let $\{\theta_1, \ldots, \theta_K\}$ be a set of possible labels with $K \geqslant 2$ and $\mathcal{Z}_k = \{(x_i, y_i) : y_i = \theta_k\}$. In the 1-v-r SVM scheme, in a first decomposition phase, K binary classifiers are trained to generate functions $f_j(x) = \langle w_j, x\rangle + b_j,\ 1 \leqslant j \leqslant K$, by separating training vectors $\mathcal{Z}_j$ with label θ_j from the rest of the training vectors $\mathcal{Z} \setminus \mathcal{Z}_j$ (j-v-r SVM). In the reconstruction phase, the label distribution generated by the trained machines is considered through a merging scheme.

A map $\Theta : \mathcal{X} \to \{\theta_1, \ldots, \theta_K\}$ is defined from the set of classifiers $F = \{f_j(x)\}_{j=1}^K$ such that, given a input vector x, it assigns a label as follows (non-normalization or standard scheme – *SS*):

$$\Theta^{SS}(x) = \arg\max_{j=1,\ldots,K} f_j(x) \tag{10}$$

It is demonstrated below that by means of experimental results and a statistical study, the use of a normalization process for this multiclassification problem is straightforward and sufficient. Firstly, it will be defined:

MS normalization: The classifiers are $f_j^M(x) = \frac{1}{\|w_j\|} f_j(x)$, and the decision function is $\Theta^M(x) = \arg\max_{j=1,\ldots,K} f_j^M(x)$.

SRM normalization: The classifiers, $f_j^R(x) = \lambda_j^R f_j(x)$, and the decision function, $\Theta^R(x) = \arg\max_{j=1,\ldots,K} f_j^R(x)$, are considered.

CH normalization: The classifiers are $f_j^{CH}(x) = \lambda_j^{CH} f_j(x)$, and the decision function is $\Theta^{CH}(x) = \arg\max_{j=1,\ldots,K} f_j^{CH}(x)$.

Table 3
p-Value and sample size m in hypothesis tests.

Data set	*SS* vs *MS*	*MS* vs *CH*	*SS* vs *CH*	m
Iris	**0.0000**	0.0000	**0.0000**	7500
Tae	0.1090	0.1028	0.9771	7500
Wine	0.8297	0.4029	0.2931	8500
Glass	0.0002	0.0000	0.7180	6300
Thyroid	0.0126	0.0065	0.8168	6300
Ecoli	0.7727	0.6895	0.9118	8250
Dermat	**0.0083**	0.6574	**0.0278**	8875
Vowel	0.0000	0.0000	0.0989	9900
Segment	**0.0064**	0.0000	**0.0000**	41200

It is worth noting that if $K = 2$, then (10) and the three normalization approaches provide the same solution since there are only two j-v-r SVM functions, $f_1(x)$ (1-v-2 SVM) and $f_2(x) = -f_1(x)$ (2-v-1 SVM).

Note that $f_j(x)$, $f_j^M(x)$, $f_j^R(x)$ and $f_j^{CH}(x)$ can be interpreted as the similarity score for the θ_j class (Crammer and Singer, 2001), and the class of x is therefore assigned to that of the classifiers with the highest score. Hence, the normalization process is very useful when SVMs are used in order to give a score for the different classes.

5. Experimental results and statistical study

The comparison between normalization and non-normalization is conducted on nine widely used data sets from UCI Repository.[2] The experiment was carried out by following a similar experimental framework to that used in (Angulo et al., 2006) as suggested in (Hsu and Lin, 2002). Hence, the training data are normalized in order to avoid problems with outliers. The test data is normalized accordingly. The selected data sets are: Iris Plants, Teaching assistant evaluation, Wine Recognition Data, Glass Identification Database, Thyroid Disease, Protein Localization Sites (*E. coli*), Dermatology, Vowel Recognition Data, and Image Segmentation. A summary of the characteristics of these data sets is shown in Table 1.

Performance for the 1-v-r SVM, in the form of accuracy rate, is evaluated on models using the linear kernel which is chosen as a baseline for the empirical evaluation, and C is explored on a one-dimensional grid with the following values: $C = \left[2^{-3}, 2^{-2}, \ldots, 2^{9}, 2^{10}\right]$.

The criteria employed to estimate the generalized accuracy is the N-fold cross-validation on the whole set of training data.[3] This procedure is repeated R times, according to the size of the data set, in order to ensure good statistical behavior.

The best cross-validation mean rate for the values of C, its standard deviation, and the value of the optimal parameter C are reported in Table 2 for the *SS*, *MS*, *SRM* and *CH* schemes.

[2] Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

[3] Except for the Segment data set, where the whole data is randomly partitioned 20 times by stratified sampling into a training set (with 250 instances) and a test set.

Table 4
Results of the experiment where the best mean accuracy rates, its standard deviation and *C*-parameter are presented for a polinomial kernel of degree 2.

Data set^{R,N}	*SS*	*MS*	*SRM*	*CH*
Iris^{50,10}	97.07 ± 0.398 (2^0)	96.97 ± 0.410 (2^1)	97.07 ± 0.398 (2^0)	$\mathbf{97.37} \pm 0.394$ (2^1)
Tae^{50,10}	58.71 ± 1.339 (2^6)	$\mathbf{59.11} \pm 1.385$ (2^6)	58.58 ± 1.337 (2^6)	58.84 ± 1.342 (2^6)
Wine^{50,10}	96.90 ± 0.364 (2^{-2})	96.82 ± 0.370 (2^{-2})	96.90 ± 0.364 (2^{-2})	$\mathbf{97.18} \pm 0.358$ (2^{-2})
Glass^{30,10}	68.08 ± 0.909 (2^3)	69.66 ± 0.869 (2^5)	68.85 ± 0.912 (2^3)	$\mathbf{70.05} \pm 0.942$ (2^4)
Thyroid^{30,10}	97.40 ± 0.290 (2^0)	97.34 ± 0.297 (2^{-1})	97.40 ± 0.290 (2^0)	$\mathbf{97.44} \pm 0.300$ (2^0)
Ecoli^{25,5}	84.83 ± 0.476 (2^{-2})	$\mathbf{84.93} \pm 0.455$ (2^0)	87.75 ± 0.456 (2^{-2})	84.77 ± 0.438 (2^{-1})
Dermat^{25,5}	$\mathbf{95.61} \pm 0.177$ (2^3)	95.47 ± 0.177 (2^3)	95.61 ± 0.177 (2^3)	95.44 ± 0.190 (2^3)
Vowel^{10,2}	86.72 ± 0.191 (2^1)	86.67 ± 0.218 (2^1)	86.81 ± 0.192 (2^1)	$\mathbf{88.89} \pm 0.205$ (2^1)
Segment	90.79 ± 0.101 (2^{-1})	91.62 ± 0.105 (2^0)	90.79 ± 0.101 (2^{-1})	$\mathbf{91.97} \pm 0.109$ (2^{-1})

Table 5
Results of the experiment where the best mean accuracy rates, its standard deviation and *C*-parameter are presented for a polynomial kernel of degree 3.

Data set^{R,N}	*SS*	*MS*	*SRM*	*CH*
Iris^{50,10}	97.07 ± 0.425 (2^{-1})	97.20 ± 0.417 (2^{-1})	97.07 ± 0.425 (2^{-1})	$\mathbf{97.22} \pm 0.417$ (2^{-1})
Tae^{50,10}	56.40 ± 1.216 (2^{-1})	56.80 ± 1.204 (2^{-1})	56.71 ± 1.219 (2^{-1})	$\mathbf{56.93} \pm 1.204$ (2^{-1})
Wine^{50,10}	$\mathbf{96.47} \pm 0.408$ (2^{-2})	96.08 ± 0.433 (2^{-2})	$\mathbf{96.47} \pm 0.408$ (2^{-2})	96.08 ± 0.433 (2^{-2})
Glass^{30,10}	68.56 ± 0.951 (2^2)	69.81 ± 0.904 (2^2)	68.75 ± 0.951 (2^2)	$\mathbf{70.15} \pm 0.946$ (2^2)
Thyroid^{30,10}	$\mathbf{96.25} \pm 0.405$ (2^0)	96.22 ± 0.397 (2^0)	96.24 ± 0.405 (2^0)	96.22 ± 0.421 (2^0)
Ecoli^{25,5}	83.13 ± 0.460 (2^{-2})	83.09 ± 0.442 (2^{-2})	83.19 ± 0.448 (2^{-2})	$\mathbf{83.27} \pm 0.409$ (2^{-2})
Dermat^{25,5}	$\mathbf{95.04} \pm 0.230$ (2^4)	94.10 ± 0.250 (2^4)	95.02 ± 0.230 (2^4)	94.31 ± 0.246 (2^4)
Vowel^{10,2}	89.27 ± 0.210 (2^0)	91.29 ± 0.170 (2^0)	89.27 ± 0.210 (2^0)	$\mathbf{91.31} \pm 0.179$ (2^0)
Segment	90.75 ± 0.121 (2^0)	91.76 ± 0.114 (2^0)	90.76 ± 0.120 (2^0)	$\mathbf{92.26} \pm 0.104$ (2^0)

Table 6
Results of the experiment where the best mean accuracy rates, *C*-parameter and σ-parameter are presented.

Data set^{R,N}	*SS*	*MS*	*SRM*	*CH*
Iris^{50,10}	96.27 ± 0.520 $(2^2, 2^1)$	96.13 ± 0.512 $(2^2, 2^1)$	96.26 ± 0.520 $(2^2, 2^1)$	$\mathbf{96.31} \pm 0.520$ $(2^2, 2^1)$
Tae^{50,10}	$60.40 \pm 1.221(2^6, 2^{-1})$	$\mathbf{60.71} \pm 1.223$ $(2^6, 2^{-1})$	60.30 ± 1.217 $(2^6, 2^{-1})$	60.63 ± 1.201 $(2^6, 2^{-1})$
Wine^{50,10}	$98.51 \pm 0.282(2^5, 2^1)$	98.54 ± 0.271 $(2^5, 2^1)$	98.51 ± 0.282 $(2^5, 2^1)$	$\mathbf{98.55} \pm 0.271(2^5, 2^1)$
Glass^{30,10}	$\mathbf{74.61} \pm 1.040$ $(2^5, 2^0)$	74.44 ± 1.028 $(2^4, 2^0)$	74.10 ± 1.037 $(2^4, 2^0)$	74.53 ± 1.074 $(2^5, 2^0)$
Thyroid^{30,10}	$97.27 \pm 0.349(2^5, 2^1)$	$\mathbf{97.43} \pm 0.328$ $(2^5, 2^1)$	97.27 ± 0.349 $(2^5, 2^1)$	97.37 ± 0.360 $(2^5, 2^1)$
Ecoli^{25,5}	$\mathbf{86.60} \pm 0.410$ $(2^3, 2^1)$	86.59 ± 0.350 $(2^3, 2^1)$	$86.56 \pm 0.395(2^3, 2^1)$	86.54 ± 0.358 $(2^3, 2^1)$
Dermat^{25,5}	96.65 ± 0.145 $(2^4, 2^1)$	$96.63 \pm 0.144(2^4, 2^0)$	$96.59 \pm 0.150(2^4, 2^1)$	$\mathbf{96.66} \pm 0.145$ $(2^4, 2^1)$
Vowel^{10,2}	95.24 ± 0.250 $(2^3, 2^0)$	94.6 ± 0.2845 $(2^7, 2^1)$	94.60 ± 0.216 $(2^5, 2^0)$	$\mathbf{95.30} \pm 0.220$ $(2^7, 2^2)$
Segment	90.06 ± 0.176 $(2^3, 2^2)$	89.85 ± 0.165 $(2^3, 2^2)$	90.07 ± 0.175 $(2^3, 2^2)$	$\mathbf{90.54} \pm 0.146(2^3, 2^2)$

Some conclusions can be drawn from the experimentation carried out:

- The accuracy rate attained for the *CH* normalization is the best in 5 of the 9 data sets (see Table 2). Furthermore, there are no statistically significant differences between the accuracy rates of the *CH* normalization and the best obtained accuracy rates (see Table 3), when it is not the winner.
- The accuracy rate attained for the standard approach is improved in 7 of the 9 data sets using some kind of normalization.
- The *MS* normalization is the best option in two cases. Nevertheless, its results have been very poor for another two cases (Glass and Vowel). This is due to the fact that only the value of $\|w_j\|$ is taken into account in the scalar factor and, in these cases, the value of $C\sum \xi_i$ is high due to the errors (low accuracy rate).
- The computational cost in each normalization is small since the norm of the normal vector (*MS*), the value of the decision function (*SRM*) and the sum of the Lagrange multipliers (*CH*) are usually all calculated in the optimization problem. However, if these values were not calculated directly then the complexity would be $O(n)$.

Hypothesis tests have been carried out in order to test whether there are statistically significant differences between the accuracy rates. Note that the *SRM* approach is omitted since its mean accuracy rates are not the best in any data set.

Let $X_i = 0$ be assigned when the i machine provides an error for the pattern x_i, and $X_i = 1$ otherwise. Hence, X_i follows a Bernoulli distribution, $X_i \sim B(p_i)$, where p_i is the probability that the i machine assigns a correct class for a new input. Given two machines, by testing $H_0 : p_i = p_j$ versus $H_1 : p_i \neq p_j$, the statistic is $Z = \frac{\sqrt{n}(\overline{X}_i - \overline{X}_j)}{\sqrt{\overline{X}_i(1-\overline{X}_i)+\overline{X}_j(1-\overline{X}_j)}} \sim N(0,1)$, where $\overline{X}_i$ and $\overline{X}_j$ are the mean sample of size n of the X_i and X_j variables, respectively. The p-value for each hypothesis test is given in Table 3.

From Table 3, it can be observed that the normalization procedure has significantly improved the accuracy rate in Iris, Dermatology and Segment data sets with respect to the standard scheme, since all p-values are less than 2.78%.

As an important result, the accuracy rate for the *CH* normalization is significantly better than the accuracy rate for the *SS* schemes in 3 out of the 9 data sets. Furthermore, there are no statistically significant differences between the accuracy rates between these two approaches in the rest of the data sets.

As aforementioned, a conclusion is that if the accuracy rate in a data set is small then the *MS* normalization must not be considered since its scalar factor fails to take the errors into account.

The same conclusions obtained from the experimentation with the linear kernel are derived when nonlinear kernels are employed. Thus, three experiments with polynomial kernels (degrees two and three) and the RBF kernel are carried out, similarly to the linear kernel, whose results are provided in Tables 4–6.

6. Conclusions

Output normalization is an absolutely necessary procedure for the comparison of outputs from different SVMs, since scaling varies for each machine considered.

Note that in the experimentation, the training and test data are normalized in order to prevent problems with outliers. Hence, the argument defended in this paper is that the output must also be normalized.

The *CH* normalization process is a good alternative since

- It has a rigorous theoretical foundation based on the reliability of the classifier.
- It follows Bayesian arguments advocating that smoother solutions should be given higher-valued weights.
- It has been empirically demonstrated that the accuracy rate of this normalization process is significantly better than the accuracy rate of the non-normalization in some data sets. No statistically significant differences between the accuracy rates of these two approaches exist in the rest of the data sets.

Acknowledgements

This study was partially supported by a grant from the Andalusian Regional Ministry of Economy, Innovation and Science (Simon-TIC-8052).

References

Angulo, C., Ruiz, F., González, L., Ortega, J.A., 2006. Multi-classification by using tri-class SVM. Neural Proc. Lett. 23 (1), 89–101.

Bennett, K.P., Bredensteiner, E.J., 2000. Duality and geometry in SVM classifiers. In: Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., pp. 57–64.

Bredensteiner, E.J., Bennett, K.P., 1999. Multicategory classification by support vector machines. Comput. Optim. Appl. 12 (1–3), 53–79.

Crammer, K., Singer, Y., 2001. On the algorithmic implementation of multiclass kernel-based vector machines. J. Machine Learn. Res. 2, 265–292.

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press.

Gonzalez-Abril, L., Angulo, C., Velasco, F., Ortega, J., 2008. A note on the bias in SVMs for multi-classification. IEEE Trans. Neural Networks 19 (4), 723–725.

Gonzalez-Abril, L., Velasco, F., Ortega, J., Franco, L., 2011. Support vector machines for classification of input vectors with different metrics. Comput. Math. Appl. 61 (9), 2874–2878.

González, L., Angulo, C., Velasco, F., Català, A., 2006. Dual unification of bi-class support vector machine formulations. Pattern Recognition 39 (7), 1325–1332.

Hsu, C., Lin, C., 2002. A comparison of methods for multiclass support vector machine. IEEE Trans. Neural Networks 13 (2), 415–425.

Liu, Y., Zhang, Y.F., 2005. One-against-all multi-class SVM classification using reliability measures. In: Proc. of the IJCNN '05.

Mayoraz, E., Alpaydin, E., 1999. Support vector machines for multi-class classification. In: IWANN, vol. 2, pp. 833–842.

Rifkin, R., Klautau, A., 2004. In defense of one-vs-all classification. J. Machine Learn. Res. 5, 101–141.

Vapnik, V., 1998. Statistical Learning Theory. John Wiley & Sons Inc..

Wang, L., Xue, P., Chan, K., 2008. Two criteria for model selection in multiclass support vector machines. IEEE Trans. Systems Man Cybernet. Part B 38 (6), 1432–1448.