Computers and Mathematics with Applications 61 (2011) 2874–2878

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

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

Support vector machines for classification of input vectors with different metrics

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

[a] Applied Economics I Department, Seville University, 41018 Seville, Spain

[b] Computer Languages and Systems Department, Seville University, 41012 Seville, Spain

ARTICLE INFO

Article history:
Received 8 January 2010
Received in revised form 15 March 2011
Accepted 21 March 2011

Keywords:
Pattern recognition
Learning machine
Maximal margin principle
ℓ_p norm

ABSTRACT

In this paper, a generalization of support vector machines is explored where it is considered that input vectors have different ℓ_p norms for each class. It is proved that the optimization problem for binary classification by using the maximal margin principle with ℓ_p and ℓ_q norms only depends on the ℓ_p norm if $1 \leq p \leq q$. Furthermore, the selection of a different bias in the classifier function is a consequence of the ℓ_q norm in this approach. Some commentaries on the most commonly used approaches of SVM are also given as particular cases.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Support vector machines (SVMs) are learning machines which implement the structural risk-minimization inductive principle to obtain good generalization on a limited number of learning patterns [1]. This theory was developed on the basis of a separable binary classification problem where the output scale is determined such that outputs for the support vectors are ± 1. The optimization criterion is the width of the margin between the positive and negative examples, since an SVM with a large margin separating two classes has a small VC dimension which provides a good generalization performance, as has been demonstrated in some applications [2].

Although the ℓ_p norm has been explored in SVMs [3,4], the norm usually used is ℓ_2 or ℓ_∞ norm and is always the same norm for both positive and negative examples. In this paper, it is considered that norm for the positive and negative examples are ℓ_p and ℓ_q norms, where[1] $p, q \geq 1$, respectively. It is worth noting that some real-world examples that justify using multiple metrics and theoretical results of this approach have recently been considered in other areas of research [5–9]. Nevertheless, to the best of our knowledge, there are not existing researches that directly address the problem of pattern recognition using this approach.

The remainder of this paper is arranged as follows: Section 2 presents a new SVM to solve linearly separable binary classification problems with a different metric for each class. Some particular cases are obtained and comments are made on this approach in Section 3, and its extensions to solve multiclassification and nonlinear separable problems are given in Section 4. Empirical testing is carried out using different biases in Section 5. Finally, some conclusions are drawn.

* Corresponding author. Tel.: +34 954554345; fax: +34 954551636.
E-mail addresses: luisgon@us.es (L. Gonzalez-Abril), velasco@us.es (F. Velasco), ortega@lsi.us.es (J.A. Ortega), lfranco@us.es (L. Franco).

[1] Note that ℓ_p is not a norm if $p < 1$.

0898-1221/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2011.03.071

2. ℓ_p-SVM-ℓ_q approach

Let $\mathcal{Z} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ be a training set, with[2] $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})' \in \mathbb{R}^d$ as the input space, $y_i \in \mathcal{Y} = \{\theta_1, \theta_2\} = \{+1, -1\}$ the output space, and $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ for $i = 1, \ldots, n$. Let $\mathcal{Z}_+$ and $\mathcal{Z}_-$ be the patterns belonging to the classes labelled as $+1$ and -1, respectively.

Let us first consider the linearly separable case, that is, a $\mathbf{w} \in \mathbb{R}^d$ (no null), $b \in \mathbb{R}$, $\mathbf{x}_{i_0} \in \mathcal{Z}_+$ and $\mathbf{x}_{j_0} \in \mathcal{Z}_-$ exist such that $\mathbf{w}'\mathbf{x} + b \geq \mathbf{w}'\mathbf{x}_{i_0} + b = 1$ if $\mathbf{x} \in \mathcal{Z}_+$, and $\mathbf{w}'\mathbf{x} + b \leq \mathbf{w}'\mathbf{x}_{j_0} + b = -1$ if $\mathbf{x} \in \mathcal{Z}_-$. Hence, both inequalities can be written as $y_i\left(\mathbf{w}'\mathbf{x}_i + b\right) \geq 1$ for any $i = 1, \ldots, n$. The best pair is sought among all pairs $(\mathbf{w}, b)$ by following the criterion of maximization of the margin [1].

Given b^* with $-1 \leq b^* \leq 1$, let us consider the hyperplane $\pi_{b^*} : \mathbf{w}'\mathbf{x} + b - b^* = 0$ and suppose, without any loss of generality, that norms in the region $\pi_{b^*} > 0$ and $\pi_{b^*} < 0$ are the ℓ_p and ℓ_q norms, respectively, with $1 \leq p \leq q$. Thus, given $\mathbf{x}_i \in \mathcal{Z}_+$ and $\mathbf{x}_j \in \mathcal{Z}_-$, the p-distance between $\mathbf{x}_i$ and π_{b^*}, that is $d_p(\mathbf{x}_i, \pi_{b^*}) = \min_{\mathbf{x}\in\pi_{b^*}} d_p(\mathbf{x}_i, \mathbf{x})$, is (see in [10,3,11]):

$$d_p(\mathbf{x}_i, \pi_{b^*}) = \frac{\left|\mathbf{w}'\mathbf{x}_i + b - b^*\right|}{\|\mathbf{w}\|_{p_1}} = \frac{\mathbf{w}'\mathbf{x}_i + b - b^*}{\|\mathbf{w}\|_{p_1}} \geq \frac{1 - b^*}{\|\mathbf{w}\|_{p_1}} = d_p(\mathbf{x}_{i0}, \pi_{b^*}) \tag{1}$$

and the q-distance between $\mathbf{x}_j$ and π_{b^*} is as follows:

$$d_q(\mathbf{x}_j, \pi_{b^*}) = \frac{\left|\mathbf{w}'\mathbf{x}_j + b - b^*\right|}{\|\mathbf{w}\|_{q_1}} = \frac{-\left(\mathbf{w}'\mathbf{x}_j + b - b^*\right)}{\|\mathbf{w}\|_{q_1}} \geq \frac{1 + b^*}{\|\mathbf{w}\|_{q_1}} = d_q(\mathbf{x}_{j0}, \pi_{b^*}) \tag{2}$$

on the condition that p_1 and q_1 are the conjugate exponents of p and q, respectively, that is,[3] $\frac{1}{p} + \frac{1}{p_1} = 1$, $\frac{1}{q} + \frac{1}{q_1} = 1$, and $\|\mathbf{w}\|_p = \left(\sum_{i=1}^M |w_i|^p\right)^{\frac{1}{p}}$, $\|\mathbf{w}\|_\infty = \max_i |w_i|$ are the p-norm and the ∞-norm of the vector $\mathbf{w}$, respectively.

By defining the p-distance between a set A and a hyperplane π_{b^*}, denoted by $d_p(A, \pi_{b^*})$, as $d_p(A, \pi_{b^*}) = \inf\left\{d_p(\mathbf{x}, \pi_{b^*}), \text{for all } \mathbf{x} \in A\right\}$, then by using (1) and (2) (the lowest bound is always attained in $\mathcal{Z}_+$ and $\mathcal{Z}_-$):

$$d_p(\mathcal{Z}_+, \pi_{b^*}) = \frac{1 - b^*}{\|\mathbf{w}\|_{p_1}} \quad \text{and} \quad d_q(\mathcal{Z}_-, \pi_{b^*}) = \frac{1 + b^*}{\|\mathbf{w}\|_{q_1}}.$$

Hence, the pq-margin between $\mathcal{Z}_+$ and $\mathcal{Z}_-$ with respect to the hyperplane π_{b^*} can be defined as $d_p(\mathcal{Z}_+, \pi_{b^*}) + d_q(\mathcal{Z}_-, \pi_{b^*})$. Therefore, the hyperplane π_{b^*} with the largest pq-margin in $\mathcal{Z}$ can be obtained by solving the problem $\max\left(d_p(\mathcal{Z}_+, \pi_{b^*}) + d_q(\mathcal{Z}_-, \pi_{b^*})\right)$, which can be written as follows:

$$\begin{aligned} &\max_{\mathbf{w},b,b^*} \quad \left(\frac{1 - b^*}{\|\mathbf{w}\|_{p_1}} + \frac{1 + b^*}{\|\mathbf{w}\|_{q_1}}\right) \\ &\text{s. t.} \quad y_i\left(\mathbf{w}'\mathbf{x}_i + b\right) \geq 1, \quad -1 \leq b^* \leq 1, \ \forall i = 1, \ldots, n. \end{aligned} \tag{3}$$

It is worth noting that the constraints do not depend on p and q. This optimization problem can be simplified since $1 \leq p \leq q \Rightarrow p_1 \geq q_1 \Rightarrow \|\mathbf{w}\|_{p_1} \leq \|\mathbf{w}\|_{q_1}$ for any $\mathbf{w} \in \mathbb{R}^d$. Hence,

$$\frac{1 - b^*}{\|\mathbf{w}\|_{p_1}} + \frac{1 + b^*}{\|\mathbf{w}\|_{q_1}} \leq \frac{1 - b^*}{\|\mathbf{w}\|_{p_1}} + \frac{1 + b^*}{\|\mathbf{w}\|_{p_1}} \leq \frac{2}{\|\mathbf{w}\|_{p_1}}$$

for any $-1 \leq b^* \leq 1$, and as this upper bound is attained by taking $b^* = -1$; therefore, the optimization problem (3) is equivalent to $\max_{\mathbf{w},b} \frac{2}{\|\mathbf{w}\|_{p_1}}$, which can be formulated as a pth-order programming problem [10]:

$$\begin{aligned} &\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|_{p_1}^{p_1} \\ &\text{s. t.} \quad y_i\left(\mathbf{w}'\mathbf{x}_i + b\right) \geq 1, \quad \forall i = 1, \ldots, n. \end{aligned} \tag{4}$$

The objective function $\frac{1}{2}\|\mathbf{w}\|_{p_1}^{p_1}$ is a convex function for any $p_1 \geq 1$ and the constraints are linear; therefore, the optimization problem (4) has a unique solution. It is important to note that the problem (4) does not depend on the ℓ_q norm ($q \geq p$) and regardless of whether the ℓ_p norm is considered in positive or negative examples.

Therefore, a binary linear classifier, $f_{(\mathbf{w},b)}(x) = \mathbf{w}'\mathbf{x} + b - 1$, is found with $\mathbf{w} = (w_1, \ldots, w_d)' \in \mathbb{R}^d$ and $b \in \mathbb{R}$ ($b^* = -1$), and where outputs are obtained as $h_{(\mathbf{w},b)}(\mathbf{x}) = \text{sign}(f_{(\mathbf{w},b)}(\mathbf{x}))$, that is, $+1$ (θ_1) if $f_{(\mathbf{w},b)}(\mathbf{x}) \geq 0$, and -1 (θ_2) otherwise. Nevertheless, a new bias, different to $b - 1$ (that is $b^* = 1$) and b (that is $b^* = 0$), must be considered since in

[2] Notation: vectors are denoted in bold, the transposed vector of $\mathbf{x}$ is denoted by $\mathbf{x}'$. Sets and real numbers are denoted in capitals and lower-case letters, respectively.

[3] If $p = 1$, then $p_1 = \infty$.

these cases $d_q(\mathcal{Z}_-, \pi_{-1}) = 0$ and $d_q(\mathcal{Z}_-, \pi_0) = \frac{1}{\|\mathbf{w}\|_{q_1}} \leq \frac{1}{\|\mathbf{w}\|_{p_1}} = d_p(\mathcal{Z}_+, \pi_0)$, that is, the vectors of $\mathcal{Z}_-$ are closer than the vectors of $\mathcal{Z}_+$ to the hyperplanes π_{-1} and π_0. Thus, a natural selection of the b^* is to choose it such that the equality $d_q(\mathcal{Z}_-, \pi_{b^*}) = d_p(\mathcal{Z}_+, \pi_{b^*})$ holds, and thus:

$$\frac{1 - b_0^*}{\|\mathbf{w}\|_{p_1}} = \frac{1 + b_0^*}{\|\mathbf{w}\|_{q_1}} \Longrightarrow b_0^* = \frac{\|\mathbf{w}\|_{q_1} - \|\mathbf{w}\|_{p_1}}{\|\mathbf{w}\|_{q_1} + \|\mathbf{w}\|_{p_1}} \geq 0. \tag{5}$$

Hence, the ℓ_q norm is useful in the search for an adequate bias. Note that if $p = q$, then $b_0^* = 0$ and the standard bias b is obtained. For this reason, the b^* is not taken into account in the developments of the standard SVM ($p = q = 2$) since it is not necessary. Henceforth, let us call this approach ℓ_p-SVM-ℓ_q.

With respect to the generalization error, it is known that this barely depends on p when the usual bias is chosen [3,4]. However, the bias in the ℓ_p-SVM-ℓ_q, $b - b_0^*$, is not the standard bias b, and the hyperplane $\pi_{b_0^*} : \mathbf{w}'\mathbf{x} + b - b_0^* = 0$ is geometrically nearer the positive class than the negative class, that is, the bias reduces skew towards the positive class. Hence, the performance on the negative class is increased but it is reduced on the positive class. For this reason, an empirical study is going to be carried out in Section 5.

3. Particular cases of ℓ_p-SVM-ℓ_q

In this section, some known approaches as particular cases of the ℓ_p-SVM-ℓ_q approach are obtained.

3.1. ℓ_1-SVM-ℓ_q

In this case, for any $q \geq 1$ the optimization problem is as follows:

$$\begin{aligned} &\min_{\mathbf{w},b} \quad \max_{i=1,\ldots,d} |w_i| \\ &\text{s. t.} \quad y_i\left(\mathbf{w}'\mathbf{x}_i + b\right) \geq 1, \quad \forall i = 1, \ldots, n. \end{aligned}$$

Furthermore,

$$\max_{1\leq p\leq q}\left(\max_{\mathbf{w},b,b^*}\left(\frac{1 - b^*}{\|\mathbf{w}\|_{p_1}} + \frac{1 + b^*}{\|\mathbf{w}\|_{q_1}}\right)\right) = \frac{2}{\|\mathbf{w}\|_\infty} = \frac{2}{\max\limits_{i=1,\ldots,d} |w_i|}.$$

That is, given $\mathcal{Z}_+$ and $\mathcal{Z}_-$ linearly separable sets, the maximal margin for any $p, q \geq 1$ is obtained for $p = 1$.

3.2. ℓ_∞-SVM-ℓ_q

If $p = \infty$, then the unique possibility of q is $q = +\infty$ and, therefore, the metric is the same in both regions. The optimization problem is as follows:

$$\begin{aligned} &\min_{\mathbf{w},b} \quad \sum_{i=1}^{d} |w_i| \\ &\text{s. t.} \quad y_i\left(\mathbf{w}'\mathbf{x}_i + b\right) \geq 1, \quad \forall i = 1, \ldots, n \end{aligned} \tag{6}$$

which is one of the most widely used approaches of SVM since the optimization problem is linear.

3.3. ℓ_2-SVM-ℓ_q

In this case, $q \geq 2$ and the optimization problem is as follows:

$$\begin{aligned} &\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 \\ &\text{s. t.} \quad y_i\left(\mathbf{w}'\mathbf{x}_i + b\right) \geq 1, \quad \forall i = 1, \ldots, n \end{aligned} \tag{7}$$

which is a quadratic optimization problem.

It is worth bearing in mind that the most common approach, called the standard primal SVM 2-norm formulation [12], is obtained for $q = 2$, and the binary linear classifier becomes $f_{(\mathbf{w},b)}(x) = \pi_0 = \mathbf{w}'\mathbf{x} + b$, that is $b^* = 0$. Hence, the selection of a different bias in [13,14] can be justified from the point of view of an ℓ_q norm with $q \geq 2$ which is explored in Section 5.

Usually, the most commonly used norms are ℓ_2- or ℓ_∞ norms and the same norm is considered in the positive and negative examples. This fact is due to the algorithms which efficiently solve these optimization problems. Hence, we think that the optimization problem (7) is more relevant than the optimization problem (6) because (i) it is only possible to consider $b^* = 0$ in the classification problem with ℓ_∞ norm, and (ii) it is possible to choose different values of ℓ_q norm ($q \geq 2$) in problem (7) which justifies the selection of the b^*.

Table 1
Results of the experiment where the best mean accuracy rates are presented.

$p+$	$q-$	Glass	Iris	Tiroide	Ecoli
2	∞	50.282	93.867	**95.869**	85.212
2	10	50.235	93.867	**95.869**	85.212
2	5	49.953	93.933	**95.869**	85.212
2	2.5	51.409	**94.533**	95.634	85.364
2	2.1	51.925	94.400	95.681	85.515
2	2	58.263	94.400	95.681	85.576
2.1	2	**58.545**	94.400	95.728	85.667
2.5	2	58.498	94.267	95.634	**85.818**
5	2	58.404	93.933	95.446	85.788
10	2	58.216	93.867	95.399	85.788
∞	2	58.216	93.867	95.399	85.788

4. The extension of ℓ_p-SVM-ℓ_q

In the same way as in the standard SVM, the ℓ_p-SVM-ℓ_q can be generalized to solve nonlinearly separable and multiclassification problems. Let us turn our attention to these topics.

4.1. Nonlinearly separable case

If some errors are allowed for the constraints [15], then the optimization problem of ℓ_p-SVM-ℓ_q can be written as follows:

$$\min_{\mathbf{w},b}\left(\frac{1}{2}\|\mathbf{w}\|_{p_1}^{p_1}+C\sum_i \xi_i\right)$$
$$\text{s. t.}\quad y_i\left(\mathbf{w}'\,\mathbf{x}_i+b\right)+\xi_i\geq 1,\quad \xi_i\geq 0,\ \forall i=1,\ldots,n$$

where C is the regularization term and ξ_i are slack variables.

Note that in the ℓ_p-SVM-ℓ_q approach, an upper bound of the number of errors in the classification problem is $2\sum\xi_i$. Nevertheless, if $p=q$, then in the same way as in the standard approach, $\sum\xi_i$ becomes an upper bound.

4.2. Multiclassification

A set of possible labels $\{\theta_1,\ldots,\theta_\ell\}$ (i.e. an unordered set of classes), with $\ell>2$, is considered. Let $\mathcal{Z}$ be a training set. Subsets $\mathcal{Z}_k\in\mathcal{Z}$, defined as $\mathcal{Z}_k=\{\mathbf{z}_i=(\mathbf{x}_i,y_i):y_i=\theta_k\}$, generate a partition in $\mathcal{Z}$. Let us suppose that the norm given in $\mathcal{Z}_k$ is ℓ_{p_k} norm with $p_k\geq 1$ for any k.

The 1-v-r SVM and the 1-v-1 SVM approaches [16] are two of the most commonly used alternatives in multiclassification problems and can be used jointly with the ℓ_p-SVM-ℓ_q. In both approaches, each machine solves bi-classification problems and the labels distribution generated by the trained machines in the parallel decomposition is into consideration, through a merging scheme which does not depend on the norm used.

5. Experiment

In this section, the comparison between different bias obtained from (5) is conducted on four widely used data sets from UCI Repository.[4] The experiment has been carried out by following a similar experimental framework to that used in [13]. The selected data sets are: **Glass** Identification Database, **Iris** Plants, **Thyroid** Disease and Protein Localization Sites (**Ecoli**).

Performance for the 1-v-r SVM, in the form of accuracy rate, has been evaluated on models using the linear kernel which has been chosen as a baseline for the empirical evaluation, and C is explored on a one-dimensional grid with the following values: $C=\left[2^{-2},2^{-1},\ldots,2^{9},2^{10}\right]$. The criteria employed to estimate the generalized accuracy is the threefold cross-validation on the whole set of training data. This procedure is repeated 20 times in order to ensure good statistical behaviour.

It is worth bearing in mind that the generalization error barely depends on p [3,4], the standard primal SVM 2-norm formulation is always used in this experimentation. Thus, the cross-validation mean rate for the values of C is reported in Table 1 for the ℓ_p-SVM-ℓ_q approach with different values of p and q.

Some analysis can be completed according to the empirical experimentation carried out: (i) The accuracy rate attained for the standard approach ($p=q=2$) can be improved by using a different bias. (ii) No single bias stands out as being the best. (iii) It has been observed that the difference in the performance between close parameter C is small. (iv) It can be seen that the generalization error barely depends on the bias.

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

6. Conclusions

The SVM 2-norm formulation in the same way as in other approaches based on SVM can be seen as a particular case of the approach development in this paper. The classification problem (4) in regions with different norms does not depend on the ℓ_q norm if $q \geq p$ and regardless of whether the ℓ_p norm is considered in positive or negative examples.

Furthermore, the selection of a different bias yields specific results related with the different metrics in each region. Clearly, the selection of p depends on the optimization problem and for simplicity this problem is usually solved by using $p = 2$ or $p = \infty$.

On the other hand, the kernel 'trick' can improve this approach in those cases where the solution can be written as a linear combination of the training vector, as happens for $p = 2$ and $p = \infty$.

References

[1] V. Vapnik, Statistical Learning Theory, John Wiley & Sons, Inc, 1998.
[2] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University press, 2000, 2000.
[3] J.P. Pedroso, N. Murata, Support vector machines with different norms: motivation, formulations and results, Pattern Recognition Letters 22 (2001) 1263–1272.
[4] K. Ikeda, N. Murata, Geometrical properties of Nu Support Vector Machines with different norms, Neural Computation 17 (11) (2005) 2508–2529.
[5] F. Love, J. Walker, An empirical comparison of block and round norms for modelling actual distances, Location Science 2 (1994) 21–43.
[6] M. Parlar, Single facility location problem with region-dpendent distance metrics, International Journal of Systems Sciences 25 (3) (1994) 513–525.
[7] J. Brimberg, Locating a single facility in the plane in the presence of a bounded region and different norms, Annals of Operations Research 122 (2003) 87–102.
[8] J. Brimberg, H. Taghizadeh, Kakhki, G. Wesolowsky, Locating a single facility in the plane in the presence of a bounded region and different norms, Journal of Operational Research Society of Japan 48 (2) (2005) 135–147.
[9] M. Zaferanieh, H.T. Kakhki, J. Brimberg, G. Wesolowsky, A bsss algorithm for the single facility location problem in two regions with different norms, European Journal of Operational Research 190 (1) (2008) 79–89.
[10] O.L. Mangasarian, Arbitrary-norm separating plane, Operations Research Letters 24 (1999) 15–23.
[11] A. Dax, The distance between two convex sets, Linear Algebra and its Applications 416 (2006) 184–213.
[12] L. González, C. Angulo, F. Velasco, A. Català, Unified dual for bi-class SVM approaches, Pattern Recognition 38 (10) (2005) 1772–1774.
[13] L. Gonzalez-Abril, C. Angulo, F. Velasco, J. Ortega, A note on the bias in SVMs for multi-classification, IEEE Transactions on Neural Networks 19 (4) (2008) 723–725.
[14] H. Nuñez, L. Gonzalez-Abril, C. Angulo, A post-processing strategy for SVM learning from unbalanced data, in: Proceedings of the 19th European Symposium on Artificial Neural Networks, ESANN, 2011.
[15] L. González, C. Angulo, F. Velasco, A. Català, Dual unification of bi-class support vector machine formulations, Pattern Recognition 39 (7) (2006) 1325–1332.
[16] C. Angulo, F. Ruiz, L. González, J.A. Ortega, Multi-classification by using tri-class SVM, Neural Proceeding Letters 23 (1) (2006) 89–101.