Knowledge-Based Systems 132 (2017) 119–128

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Synchronized feature selection for Support Vector Machines with twin hyperplanes

Sebastián Maldonado[a,*], Julio López[b]

[a] Facultad de Ingeniería y Ciencias Aplicadas, Universidad de los Andes, Monseñor Álvaro del Portillo 12455, Las Condes, Santiago, Chile
[b] Facultad de Ingeniería y Ciencias, Universidad Diego Portales, Ejército 441, Santiago, Chile

ARTICLE INFO

Article history:
Received 20 January 2017
Revised 11 May 2017
Accepted 16 June 2017
Available online 17 June 2017

Keywords:
Support vector machine
Embedded methods
Feature selection
L-infinity norm

ABSTRACT

In this work, a novel feature selection method for twin Support Vector Machine (SVM) is presented. The main idea is to combine two regularizers, namely the Euclidean and infinite norm to perform twin classification and variable selection simultaneously. This latter task is performed in a coordinated fashion, enabling that the same attributes are selected in each twin classifiers. A single optimization problem is used to solve both subproblems, leading to a sparse final classification rule. Experiments on low- and high-dimensional datasets indicate that our approaches present the best average performance compared to well-known feature selection strategies, also achieving a synchronized feature elimination in the two twin classifiers. Our approaches are also able to improve the performance of the twin classifier, demonstrating the importance of feature selection in high-dimensional tasks.

1. Introduction

Feature selection is an important task in knowledge-based decision systems [1]. The goal of feature selection is to construct simple models based on only the attributes that are relevant for a particular application, which are preferable to more complex ones according to the principle of Occam's razor [2]. A low dimensional representation of the data leads to several advantages, such as better predictive performance thanks to the lower risk of overfitting, better understanding of the outcome of the modelling process for decision-making, and reduced storage and acquisition costs [3,4].

In this work, we explore feature selection methods for Support Vector Machines (SVMs) [5]. This supervised learning approach has proved to be very successful for several reasons, such as its ability to generalize the training patterns better thanks to the structural risk minimization principle, and the representation being based on only a few data points. Unfortunately, SVM is not designed to identify the relevant variables automatically [6]. To overcome this issue, feature selection can be embedded into the training process as a part of the optimization problem (see e.g.[7–9]).

In this work, we focus on twin SVM [10], an extension of the traditional SVM method in which two nonparallel hyperplanes are constructed instead of a single one. The construction of these classifiers is traditionally done by splitting the optimization problem (that appears in the classical SVM) into two subproblems [10,11]. Alternatively, the twin classifiers can also be obtained by solving a single optimization problem; a model known as nonparallel hyperplane SVM (NH-SVM) [12].

Feature selection is more challenging in twin SVM than in standard SVM because two hyperplanes are constructed, and each one of them can consider a different subset of variables as relevant. In this work, we propose a novel strategy for twin SVM classification and synchronized feature selection, in which a group penalty function [13,14] is introduced as a second regularizer. In contrast with well-known regularizers for feature selection, such as the l_1- and l_0-norms, a group penalty function aims at jointly penalizing the weights related to a given variable in both hyperplanes. The l_∞-norm [15] is used as a group penalty function, and the NH-SVM model is modified in order to solve a single optimization problem to obtain both twin classifiers.

The contents of the remainder of this work as follows: in Section 2 we describe the methodological background that is relevant for our proposal, which includes twin SVM formulations, and feature selection approaches. The proposed method based on double regularization for NH-SVM is presented in Section 3. Experimental results using benchmark data sets are given in Section 4. Finally, Section 5 provides the main conclusion and addresses future developments.

* Corresponding author.
E-mail addresses: smaldonado@uandes.cl (S. Maldonado), julio.lopez@udp.cl (J. López).

http://dx.doi.org/10.1016/j.knosys.2017.06.025
0950-7051/

2. Literature overview

In this section, we briefly describe the methods that are relevant for our proposal: the twin SVM, the nonparallel hyperplane SVM, feature selection strategies for SVM, and the concept of group penalty functions.

2.1. Twin support vector machine

The twin SVM method [10] is designed to construct two nonparallel hyperplanes, in contrast with traditional SVM, in which a single hyperplane defines the classification rule. These two "twin" classifiers are constructed independently via two different quadratic programming (QP) problems. Given data matrices $A \in \Re^{m_1 \times n}$ and $B \in \Re^{m_2 \times n}$ for the positive and negative training patterns, respectively, twin SVM constructs two classifiers of the form $\mathbf{w}_k^\top \mathbf{x} + b_k = 0$ $(k = 1, 2)$ in such a way that each function is closer to instances of one of the two classes, and as far as possible from those of the other class at the same time. The twin SVM formulation has the following form:

$$\begin{aligned} \min_{\mathbf{w}_1, b_1, \boldsymbol{\xi}_2} \quad & \frac{1}{2}\|A\mathbf{w}_1 + \mathbf{e}_1 b_1\|^2 + \frac{c_1}{2}(\|\mathbf{w}_1\|^2 + b_1^2) + c_3 \mathbf{e}_2^\top \boldsymbol{\xi}_2 \\ \text{s.t.} \quad & -(B\mathbf{w}_1 + \mathbf{e}_2 b_1) \geq \mathbf{e}_2 - \boldsymbol{\xi}_2, \\ & \boldsymbol{\xi}_2 \geq 0, \end{aligned} \tag{1}$$

and

$$\begin{aligned} \min_{\mathbf{w}_2, b_2, \boldsymbol{\xi}_1} \quad & \frac{1}{2}\|B\mathbf{w}_2 + \mathbf{e}_2 b_2\|^2 + \frac{c_2}{2}(\|\mathbf{w}_2\|^2 + b_2^2) + c_4 \mathbf{e}_1^\top \boldsymbol{\xi}_1 \\ \text{s.t.} \quad & (A\mathbf{w}_2 + \mathbf{e}_1 b_2) \geq \mathbf{e}_1 - \boldsymbol{\xi}_1, \\ & \boldsymbol{\xi}_1 \geq 0, \end{aligned} \tag{2}$$

where $c_i > 0$ $(i = 1, 2, 3, 4)$ are trade-off parameters designed to balance the compromise between complexity reduction (minimization of the Euclidean norm of both weight vectors) and model fit. The elements $\mathbf{e}_1$ and $\mathbf{e}_2$ are vectors of ones of appropriate dimensions. As decision rule, a new observation $\mathbf{x}$ is assigned to class k^* corresponding to closest hyperplane:

$$k^* = \underset{k=1,2}{\operatorname{argmin}} \left\{ d_k(\mathbf{x}) := \frac{|\mathbf{w}_k^\top \mathbf{x} + b_k|}{\|\mathbf{w}_k\|} \right\}, \tag{3}$$

where d_k is the distance of $\mathbf{x}$ from classifier $\mathbf{w}_k^\top \mathbf{x} + b_k = 0$, $k = 1, 2$. Formulation (1) and (2) is known as twin-bounded SVM (TB-SVM) [11], which is similar compared to the twin SVM (TW-SVM) formulation proposed by Jayadeva et al. [10] by setting if $c_1 = c_2 = \epsilon$. This formulation can also be extended to construct nonlinear classifiers thanks to the "kernel trick" (see [10,11] for details).

2.2. Nonparallel hyperplane SVM (NH-SVM)

The NH-SVM method follows the same ideas as twin SVM, but it constructs the two hyperplanes in a single optimization problem. Formally, the following QP problem is solved:

$$\begin{aligned} \min_{\substack{\mathbf{w}_k, b_k, \boldsymbol{\xi}_k \\ k=1,2}} \quad & \frac{1}{2}\left(\|A\mathbf{w}_1 + \mathbf{e}_1 b_1\|^2 + \|B\mathbf{w}_2 + \mathbf{e}_2 b_2\|^2\right) \\ & + \frac{c_1}{2}\left(\|\mathbf{w}_1\|^2 + b_1^2 + \|\mathbf{w}_2\|^2 + b_2^2\right) + c_2\left(\mathbf{e}_1^\top \boldsymbol{\xi}_1 + \mathbf{e}_2^\top \boldsymbol{\xi}_2\right) \\ \text{s.t.} \quad & A\mathbf{w}_1 + \mathbf{e}_1 b_1 - A\mathbf{w}_2 - \mathbf{e}_1 b_2 \geq \mathbf{e}_1 - \boldsymbol{\xi}_1, \\ & B\mathbf{w}_2 + \mathbf{e}_2 b_2 - B\mathbf{w}_1 - \mathbf{e}_2 b_1 \geq \mathbf{e}_2 - \boldsymbol{\xi}_2, \\ & \boldsymbol{\xi}_1 \geq 0, \boldsymbol{\xi}_2 \geq 0, \end{aligned} \tag{4}$$

where $c_k > 0$ $(k = 1, 2)$ are trade-off parameters [12]. The decision rule is equivalent to twin SVM.

Besides NH-SVM, some extensions for twin SVM have been proposed in the literature. In particular, the reasoning behind twin SVM has been used in regression [16], multi-class classification [17], and robust classification via second-order cone programming [18]. For the latter approach, the SVM principle of maximum-margin classification is used to construct twin hyperplanes that correctly classify the training patterns for specified error rates [19]. The proposed robust setting has the ability of generalizing better by assuming a pessimistic data distribution of the class-conditional densities with given mean and covariance matrices [18]. Like twin SVM, the NH-SVM method is also suitable for kernel functions once the kernel trick is applied to Formulation (4).

2.3. Feature selection for SVM

Several approaches have been proposed for feature selection in binary SVM classification. The *Fisher Score* [20], for example, measures the correlation between predictors and the target variable by calculating the difference between the mean of both classes for each variable, as follows:

$$F(j) = \left| \frac{\mu_j^+ - \mu_j^-}{(\sigma_j^+)^2 + (\sigma_j^-)^2} \right|, \tag{5}$$

where μ_j^+ (μ_j^-) is the mean for the jth attribute in the positive (negative) class and σ_j^+ (σ_j^-) is the respective standard deviation. Attributes can be ranked according to this measure, and SVM can be trained subsequently using the subset of r variables with the highest Fisher Score. Since this method ranks attributes before applying any classification tasks, it can be used jointly with twin SVM or NH-SVM, also allowing the use of kernel functions.

The *Recursive Feature Elimination SVM* (SVM-RFE) method is another well-known strategy for feature selection for SVM [21]. Instead of computing a measure that is independent of the model, SVM-RFE removes those variables whose elimination leads to the largest margin of class separation in a backward fashion. Since the margin is inversely proportional to the Euclidean norm of the weight vector, this value can be rewritten as follows:

$$W^2(\boldsymbol{\alpha}) = \sum_{i,s=1}^{m} \alpha_i \alpha_s y_i y_s \mathbf{x}_i \cdot \mathbf{x}_s, \tag{6}$$

where $\boldsymbol{\alpha}$ are the dual variables of the standard SVM formulation. Thus, SVM-RFE ranks the variables in terms of the measure $|W^2(\boldsymbol{\alpha}) - W^2_{(-p)}(\boldsymbol{\alpha})|$, where $W^2_{(-p)}(\boldsymbol{\alpha})$ is equivalent to $W^2(\boldsymbol{\alpha})$, with the only difference being that variable p is eliminated from each training point [21]. Since this method can be constructed using the dual formulation of SVM, it is suitable for the use of kernel functions.

The SVM-RFE method has been extended for twin SVM [22]. For each attribute j, the TWSVM-RFE method computes the sum of the absolute values of both weights, w_{1j} and w_{2j}, associated with each twin hyperplane. All weights needs to be normalized previously, i.e., the attributes are ranked according to $W(j) = w^*_{1j} + w^*_{2j}$, where $w^*_{lj} = \frac{|w_{lj}|}{||\mathbf{w}_l||_2}$, for $l = 1, 2$.

The main issue with the TWSVM-RFE method is that feature selection is not a coordinated strategy between the twin hyperplanes, and relevance is given simply by the average magnitude of their weights. Our method, in contrast, performs an embedded feature selection process that encourages sparsity in both twin problems simultaneously, aiming at finding a small set of common variables that work well in both classification functions.

Regarding embedded methods, an important approach is the use of the LASSO penalty or l_1-norm instead of the Euclidean norm for SVM regularization. The LASSO penalty finds a good compromise between predictive performance and sparsity [23]. Feature selection strategy is to use an approximation of the l_0-norm or the

cardinality of the non-zero elements of a vector. This strategy can be applied by replacing the Euclidean norm [23], or in combination with it [6,24].

The l_1- and l_0-norms can be applied to perform feature selection on twin SVM. For example, these norms can be used instead of the Euclidean norm in each twin subproblem, resulting in two low-dimensional classifiers. This approach was suggested by Bai et al. [1] for the l_1-norm (L_1-TWSVM), and has the following formulation:

$$\begin{aligned} \min_{\mathbf{w}_1,b_1,\boldsymbol{\xi}_2} \quad & \|\mathbf{w}_1\|_1 + c_1\|A\mathbf{w}_1 + \mathbf{e}_1 b_1\|_1 + c_3\mathbf{e}_2^\top\boldsymbol{\xi}_2 \\ \text{s.t.} \quad & -(B\mathbf{w}_1 + \mathbf{e}_2 b_1) \geq \mathbf{e}_2 - \boldsymbol{\xi}_2, \\ & \boldsymbol{\xi}_2 \geq 0, \end{aligned} \tag{7}$$

and

$$\begin{aligned} \min_{\mathbf{w}_2,b_2,\boldsymbol{\xi}_1} \quad & \|\mathbf{w}_2\|_1 + c_2\|B\mathbf{w}_2 + \mathbf{e}_2 b_2\|_1 + c_4\mathbf{e}_1^\top\boldsymbol{\xi}_1 \\ \text{s.t.} \quad & (A\mathbf{w}_2 + \mathbf{e}_1 b_2) \geq \mathbf{e}_1 - \boldsymbol{\xi}_1, \\ & \boldsymbol{\xi}_1 \geq 0. \end{aligned} \tag{8}$$

The main issue with this approach is that the variables selected in each subproblem may differ, and the combined set of selected attributes could be large. Although this strategy may lead to good predictive performance, a reduced set of selected variables also enhances interpretability and reduces storage and variable acquisition costs. This strategy has also been applied in a multi-class context [25], but it has the same issue: feature selection is performed independently in each classification function, and not in a synchronized fashion.

In order to overcome this issue, Bai et al. [1] replaced all Euclidean norms in the twin SVM formulation with the l_1-norm, and introduced a feature selection matrix set E consisting of a set of binary variables in the diagonal of this matrix that indicates whether or not a variable is selected. This method, called FTSVM, has the advantage of performing a synchronized feature selection, which is also suitable for the use of kernel functions. Unfortunately, the inclusion of binary variables leads to a multi-objective mixed-integer programming problem, which has higher complexity compared with L_1-TWSVM, and becomes intractable in high-dimensional settings.

Alternatively, we propose using a group penalty function to overcome this issue. This type of regularization functions have been proposed for binary classification problems with grouped variables [26], and subsequently extended to multi-class classification [27]. Grouped variables are, for example, nominal attributes with multiple categories expressed through a set of dummy variables [26]. Since we may want to eliminate the original attribute instead of the dummy variables individually, we require a regularizer that penalizes the use of the full set of dummy variables.

The group-lasso penalty [26] works as follows: Suppose that each of the n attributes are put in disjoint sets $\mathcal{I}_j$ of dummy variables, where $|\mathcal{I}_j| = p_j$, for $j = 1,\ldots,J$, and $\sum_{j=1}^{J} p_j = n$. The group-lasso function has the following form:

$$\Gamma(\mathbf{w}) = \sum_{j=1}^{J} \sqrt{p_j}||\mathbf{w}^{(j)}||_2 \tag{9}$$

where $||\mathbf{w}^{(j)}||_2 = \sqrt{\sum_{l\in\mathcal{I}_j} w_l^2}$. Another strategy for grouped variables is the l_∞-norm penalty, proposed for the F_∞-norm SVM method [15]. The penalty function has the following form:

$$\Gamma(\mathbf{w}) = \sum_{j=1}^{J} ||\mathbf{w}^{(j)}||_\infty \tag{10}$$

where $||\mathbf{w}^{(j)}||_\infty = \max_{l\in\mathcal{I}_j}\{|w_l|\}$. The F_∞-norm SVM method can be cast into a linear programming problem by introducing a set of slack variables $t_j = ||\mathbf{w}^{(j)}||_\infty$, and adding new constraints $|w_l| \leq t_j$ for each $l \in \mathcal{I}_j$ and $j = 1,\ldots,J$.

3. A novel SVM method for simultaneous twin feature selection

In this section, we propose a novel method for embedded feature selection and twin SVM classification. The main idea is to add the l_∞-norm penalization in the NH-SVM method in order to achieve a coordinated elimination of variables in both hyperplanes, conferring sparsity to the twin SVM model. The choice of NH-SVM as a baseline classification approach, instead of the traditional twin SVM model, is not arbitrary: since the NH-SVM method solves a single optimization problem to construct the twin classifiers, we can include the l_∞-norm penalization and act in both hyperplanes simultaneously. This is not possible to do in the twin SVM model by Jayadeva [10] because the problem is constructed in two independent QP problems.

The formulation that we propose is the following:

$$\begin{aligned} \min_{\substack{\mathbf{w}_k,b_k,\boldsymbol{\xi}_k \\ k=1,2}} \quad & \frac{1}{2}\left(\|A\mathbf{w}_1 + \mathbf{e}_1 b_1\|^2 + \|B\mathbf{w}_2 + \mathbf{e}_2 b_2\|^2\right) + \lambda\sum_{j=1}^{n}\|\mathbf{w}^{(j)}\|_\infty \\ & + \frac{c_1}{2}\left(\|\mathbf{w}_1\|^2 + b_1^2 + \|\mathbf{w}_2\|^2 + b_2^2\right) + c_2\left(\mathbf{e}_1^\top\boldsymbol{\xi}_1 + \mathbf{e}_2^\top\boldsymbol{\xi}_2\right) \\ \text{s.t.} \quad & A\mathbf{w}_1 + \mathbf{e}_1 b_1 - A\mathbf{w}_2 - \mathbf{e}_1 b_2 \geq \mathbf{e}_1 - \boldsymbol{\xi}_1, \\ & B\mathbf{w}_2 + \mathbf{e}_2 b_2 - B\mathbf{w}_1 - \mathbf{e}_2 b_1 \geq \mathbf{e}_2 - \boldsymbol{\xi}_2, \\ & \boldsymbol{\xi}_1 \geq 0, \boldsymbol{\xi}_2 \geq 0, \end{aligned} \tag{11}$$

where $\mathbf{w}^{(j)} = (w_{1j}, w_{2j}) \in \Re^2$, and $\|\mathbf{w}^{(j)}\|_\infty = \max_{k=1,2}\{|w_{kj}|\}$, for $j = 1,\ldots,n$. It can be seen that the above formulation corresponds to the NH-SVM method (Formulation (4)) with the inclusion of the l_∞-norm regularization term (Eq. (10)) in the objective function. Parameters c_1, c_2, and λ control the trade-offs among l_2 regularization (margin maximization), model fit, and sparsity; respectively.

The formulation (11) can be cast into a QP problem by introducing an additional variable $\mathbf{z} \in \Re^n$. The idea is to avoid using the maximum in the objective function. Specifically, one has the following QP problem:

$$\begin{aligned} \min_{\substack{\mathbf{w}_k,b_k,\boldsymbol{\xi}_k,\mathbf{z} \\ k=1,2}} \quad & \frac{1}{2}\left(\|A\mathbf{w}_1 + \mathbf{e}_1 b_1\|^2 + \|B\mathbf{w}_2 + \mathbf{e}_2 b_2\|^2\right) + \lambda\mathbf{e}^\top\mathbf{z} \\ & + \frac{c_1}{2}\left(\|\mathbf{w}_1\|^2 + b_1^2 + \|\mathbf{w}_2\|^2 + b_2^2\right) + c_2\left(\mathbf{e}_1^\top\boldsymbol{\xi}_1 + \mathbf{e}_2^\top\boldsymbol{\xi}_2\right) \\ \text{s.t.} \quad & A\mathbf{w}_1 + \mathbf{e}_1 b_1 - A\mathbf{w}_2 - \mathbf{e}_1 b_2 \geq \mathbf{e}_1 - \boldsymbol{\xi}_1, \\ & B\mathbf{w}_2 + \mathbf{e}_2 b_2 - B\mathbf{w}_1 - \mathbf{e}_2 b_1 \geq \mathbf{e}_2 - \boldsymbol{\xi}_2, \\ & \boldsymbol{\xi}_1 \geq 0, \boldsymbol{\xi}_2 \geq 0, \\ & |\mathbf{w}_k| \leq \mathbf{z}, \quad k = 1,2. \end{aligned} \tag{12}$$

Note that the last constraint of Formulation (12) can be replaced by $\mathbf{w}_k \leq \mathbf{z}$ and $-\mathbf{w}_k \leq \mathbf{z}$ for $k = 1,2$, leading to our final QP problem. We refer to this formulation as the **twin $l_2 l_\infty$-SVM**.

Next, we propose a simpler variation by setting $c_1 = 0$. This proposal is aligned with the F_∞-norm SVM method, in which the l_2 regularization is replaced by the l_∞-norm, instead of using a linear combination of both regularizers. Specifically, we consider the following problem:

$$\begin{aligned} \min_{\substack{\mathbf{w}_k,b_k,\boldsymbol{\xi}_k \\ k=1,2}} \quad & \frac{1}{2}\left(\|A\mathbf{w}_1 + \mathbf{e}_1 b_1\|^2 + \|B\mathbf{w}_2 + \mathbf{e}_2 b_2\|^2\right) + \lambda\sum_{j=1}^{n}\|\mathbf{w}^{(j)}\|_\infty \\ & + c_2\left(\mathbf{e}_1^\top\boldsymbol{\xi}_1 + \mathbf{e}_2^\top\boldsymbol{\xi}_2\right) \\ \text{s.t.} \quad & A\mathbf{w}_1 + \mathbf{e}_1 b_1 - A\mathbf{w}_2 - \mathbf{e}_1 b_2 \geq \mathbf{e}_1 - \boldsymbol{\xi}_1, \\ & B\mathbf{w}_2 + \mathbf{e}_2 b_2 - B\mathbf{w}_1 - \mathbf{e}_2 b_1 \geq \mathbf{e}_2 - \boldsymbol{\xi}_2, \\ & \boldsymbol{\xi}_1 \geq 0, \boldsymbol{\xi}_2 \geq 0. \end{aligned} \tag{13}$$

We refer to Formulation (13) as the **twin l_∞-SVM**.

Dual formulation of the twin l_2l_∞-SVM method

In this section, the dual formulation of the twin l_2l_∞-SVM method is derived. Duality is very useful for Support Vector Machine mainly for three reasons: more efficient training can be performed (arguably the most popular optimization strategy for SVM is SMO [28] which is constructed from the dual SVM formulation), the use of the kernel trick [29], and the geometrical interpretation that leads the dual form of SVM [30].

The Lagrangian function associated with Formulation (12) is given by

$$\begin{aligned} L(\mathbf{w}_k, b_k, \boldsymbol{\xi}_k, \mathbf{z}, \mathbf{r}_k, \mathbf{s}_k, \mathbf{t}_k, \hat{\mathbf{t}}_k) &= \frac{1}{2}\left(\|A\mathbf{w}_1 + \mathbf{e}_1 b_1\|^2 + \|B\mathbf{w}_2 + \mathbf{e}_2 b_2\|^2\right) \\ &+ \frac{c_1}{2}\left(\|\mathbf{w}_1\|^2 + b_1^2 + \|\mathbf{w}_2\|^2 + b_2^2\right) + c_2\mathbf{e}_1^\top\boldsymbol{\xi}_1 \\ &+ c_2\mathbf{e}_2^\top\boldsymbol{\xi}_2 - \mathbf{r}_1^\top\boldsymbol{\xi}_1 - \mathbf{r}_2^\top\boldsymbol{\xi}_2 + \lambda\mathbf{e}^\top\mathbf{z} \\ &- \mathbf{s}_1^\top(A(\mathbf{w}_1 - \mathbf{w}_2) + \mathbf{e}_1(b_1 - b_2) - \mathbf{e}_1 + \boldsymbol{\xi}_1) \\ &- \mathbf{s}_2^\top(B(\mathbf{w}_2 - \mathbf{w}_1) + \mathbf{e}_1(b_2 - b_1) - \mathbf{e}_2 + \boldsymbol{\xi}_2) \\ &- \sum_{k=1}^{2}[\mathbf{t}_k^\top(\mathbf{z} + \mathbf{w}_k) + \hat{\mathbf{t}}_k^\top(\mathbf{z} - \mathbf{w}_k)], \end{aligned} \tag{14}$$

where $\mathbf{r}_k, \mathbf{s}_k \in \Re_+^{m_k}$, $\mathbf{t}_k, \hat{\mathbf{t}}_k \in \Re_+^n$, for $k = 1, 2$. Thus, Problem (12) can be written equivalently as

$$\min_{\mathbf{w}_k, b_k, \boldsymbol{\xi}_k, \mathbf{z}} \max_{\mathbf{r}_k, \mathbf{s}_k, \mathbf{t}_k, \hat{\mathbf{t}}_k} \{L(\mathbf{w}_k, b_k, \boldsymbol{\xi}_k, \mathbf{z}, \mathbf{r}_k, \mathbf{s}_k, \mathbf{t}_k, \hat{\mathbf{t}}_k) : \mathbf{r}_k, \mathbf{s}_k, \mathbf{t}_k, \hat{\mathbf{t}}_k \geq 0, k = 1, 2\}, \tag{15}$$

and hence the Wolfe-dual of Problem (12) (see [31]) corresponds to

$$\max_{\mathbf{r}_k, \mathbf{s}_k, \mathbf{t}_k, \hat{\mathbf{t}}_k} \min_{\mathbf{w}_k, b_k, \boldsymbol{\xi}_k, \mathbf{z}} \{L : \nabla_{\mathbf{w}_k} L = \nabla_{\mathbf{z}} L = 0, \nabla_{b_k} L = 0, \nabla_{\boldsymbol{\xi}_k} L = 0, \mathbf{r}_k, \mathbf{s}_k, \mathbf{t}_k, \hat{\mathbf{t}}_k \geq 0\}. \tag{16}$$

Computing the gradient of L with respect to $\mathbf{w}_k$, b_k, $\boldsymbol{\xi}_k$ $(k = 1, 2)$, and $\mathbf{z}$ leads to the following linear system:

$$(A^\top A + c_1 I)\mathbf{w}_1 + b_1 A^\top\mathbf{e}_1 - A^\top\mathbf{s}_1 + B^\top\mathbf{s}_2 - \mathbf{t}_1 + \hat{\mathbf{t}}_1 = 0, \tag{17}$$

$$(B^\top B + c_1 I)\mathbf{w}_2 + b_2 B^\top\mathbf{e}_2 + A^\top\mathbf{s}_1 - B^\top\mathbf{s}_2 - \mathbf{t}_2 + \hat{\mathbf{t}}_2, = 0, \tag{18}$$

$$\lambda\mathbf{e} - (\mathbf{t}_1 + \mathbf{t}_2) - (\hat{\mathbf{t}}_1 + \hat{\mathbf{t}}_2) = 0, \tag{19}$$

$$\mathbf{e}_1^\top A\mathbf{w}_1 + b_1(c_1 + \mathbf{e}_1^\top\mathbf{e}_1) - \mathbf{s}_1^\top\mathbf{e}_1 + \mathbf{s}_2^\top\mathbf{e}_2 = 0, \tag{20}$$

$$\mathbf{e}_2^\top B\mathbf{w}_2 + b_2(c_1 + \mathbf{e}_2^\top\mathbf{e}_2) + \mathbf{s}_1^\top\mathbf{e}_1 - \mathbf{s}_2^\top\mathbf{e}_2 = 0, \tag{21}$$

$$c_2\mathbf{e}_1 - \mathbf{s}_1 - \mathbf{r}_1 = 0, \tag{22}$$

$$c_2\mathbf{e}_2 - \mathbf{s}_2 - \mathbf{r}_2 = 0. \tag{23}$$

Since $\mathbf{r}_k, \mathbf{s}_k \geq 0$ for $k = 1, 2$, the following relation can be obtained from (22) and (23):

$$0 \leq \mathbf{s}_k \leq c_2\mathbf{e}_k, \quad k = 1, 2. \tag{24}$$

In addition, the Lagrangian (14) can be rewritten as

$$\begin{aligned} L = &\frac{1}{2}\mathbf{v}_1^\top(H^\top H + c_1 I)\mathbf{v}_1 + \frac{1}{2}\mathbf{v}_2^\top(G^\top G + c_1 I)\mathbf{v}_2 + \boldsymbol{\xi}_1^\top(c_2\mathbf{e}_1 - \mathbf{r}_1 - \mathbf{s}_1) \\ &+ \boldsymbol{\xi}_2^\top(c_2\mathbf{e}_2^\top - \mathbf{r}_2 - \mathbf{s}_2) + \mathbf{z}^\top(\lambda\mathbf{e}^\top - \mathbf{t}_1 - \mathbf{t}_2 - \hat{\mathbf{t}}_1 - \hat{\mathbf{t}}_2) + \mathbf{s}_2^\top\mathbf{e}_2 \\ &+ \mathbf{s}_1^\top\mathbf{e}_1 - \mathbf{s}_2^\top(B(\mathbf{w}_2 - \mathbf{w}_1) + \mathbf{e}_1(b_2 - b_1)) - \mathbf{s}_1^\top(A(\mathbf{w}_1 - \mathbf{w}_2) \\ &+ \mathbf{e}_1(b_1 - b_2)) - \sum_{k=1}^{2}(\mathbf{t}_k - \hat{\mathbf{t}}_k)^\top\mathbf{w}_k, \end{aligned} \tag{25}$$

with $\mathbf{v}_k = [\mathbf{w}_k^\top, b_k]^\top \in \Re^{n+1}$ for $k = 1, 2$, $H = [A, \mathbf{e}_1] \in \Re^{m_1 \times (n+1)}$, and $G = [B, \mathbf{e}_2] \in \Re^{m_2 \times (n+1)}$. Then, the use of relations (19), (22), and (23) in (25) leads to the following form for the Lagrangian:

$$\begin{aligned} L = &\frac{1}{2}\mathbf{v}_1^\top(H^\top H + c_1 I)\mathbf{v}_1 + \frac{1}{2}\mathbf{v}_2^\top(G^\top G + c_1 I)\mathbf{v}_2 + \mathbf{s}_1^\top\mathbf{e}_1 + \mathbf{s}_2^\top\mathbf{e}_2 \\ &- \mathbf{w}_2^\top(\mathbf{t}_2 - \hat{\mathbf{t}}_2) + \mathbf{v}_1^\top[-H^\top \; G^\top]\begin{pmatrix}\mathbf{s}_1\\ \mathbf{s}_2\end{pmatrix} + \mathbf{v}_2^\top[H^\top \; -G^\top]\begin{pmatrix}\mathbf{s}_1\\ \mathbf{s}_2\end{pmatrix} \\ &- \mathbf{w}_1^\top(\mathbf{t}_1 - \hat{\mathbf{t}}_1). \end{aligned} \tag{26}$$

Note also that relations (17) and (20), and relations (18) and (21) can be written compactly as

$$(H^\top H + c_1 I)\mathbf{v}_1 + [-H^\top \; G^\top]\begin{pmatrix}\mathbf{s}_1\\ \mathbf{s}_2\end{pmatrix} + \begin{pmatrix}-\mathbf{t}_1 + \hat{\mathbf{t}}_1\\ 0\end{pmatrix} = 0 \tag{27}$$

and

$$(G^\top G + c_1 I)\mathbf{v}_2 + [H^\top \; -G^\top]\begin{pmatrix}\mathbf{s}_1\\ \mathbf{s}_2\end{pmatrix} + \begin{pmatrix}-\mathbf{t}_2 + \hat{\mathbf{t}}_2\\ 0\end{pmatrix} = 0, \tag{28}$$

respectively. From (27) and (28), we have that relation (26) reduces to

$$L = \mathbf{s}_1^\top\mathbf{e}_1 + \mathbf{s}_2^\top\mathbf{e}_2 - \frac{1}{2}\mathbf{v}_1^\top(H^\top H + c_1 I)\mathbf{v}_1 - \frac{1}{2}\mathbf{v}_2^\top(G^\top G + c_1 I)\mathbf{v}_2. \tag{29}$$

Finally, since the symmetric matrices $H^\top H + c_1 I$ and $G^\top G + c_1 I$ are definite positive, for any $c_1 > 0$, the dual formulation for the twin l_2l_∞-SVM method can be derived by using Eqs. (27) and (28) in Eq. (29):

$$\begin{aligned} \min_{\substack{\mathbf{s}_k, \mathbf{t}_k, \hat{\mathbf{t}}_k \\ k=1,2}} \; &\frac{1}{2}\boldsymbol{\alpha}^\top\left(\bar{A}_1^\top(H^\top H + c_1 I)^{-1}\bar{A}_1 + \bar{A}_2^\top(G^\top G + c_1 I)^{-1}\bar{A}_2\right)\boldsymbol{\alpha} \\ &- \mathbf{s}_1^\top\mathbf{e}_1 - \mathbf{s}_2^\top\mathbf{e}_2 \\ \text{s.t. } &0 \leq \mathbf{s}_k \leq c_2\mathbf{e}_k, \;\; \mathbf{t}_k, \hat{\mathbf{t}}_k \geq 0, \;\; k = 1, 2, \\ &\mathbf{t}_1 + \mathbf{t}_2 + \hat{\mathbf{t}}_1 + \hat{\mathbf{t}}_2 = \lambda\mathbf{e}, \end{aligned} \tag{30}$$

where $\bar{A}_1 = [-H^\top \; G^\top \; -\hat{I} \;\; 0 \;\; \hat{I} \;\; 0]$, $\bar{A}_2 = [H^\top \; -G^\top \; 0 \;\; -\hat{I} \;\; 0 \;\; \hat{I}] \in \Re^{n+1 \times 4n+m}$, with $\hat{I} = \begin{pmatrix} I \\ 0 \end{pmatrix} \in \Re^{n+1 \times n}$, $m = m_1 + m_2$; and $\boldsymbol{\alpha} = [\mathbf{s}_1^\top, \mathbf{s}_2^\top, \mathbf{t}_1^\top, \mathbf{t}_2^\top, \hat{\mathbf{t}}_1^\top, \hat{\mathbf{t}}_2^\top]^\top \in \Re^{4n+m}$.

Remark 1. The solution $\mathbf{v}_k = [\mathbf{w}_k^\top, b_k]^\top$ $(k = 1, 2)$ to twin l_2l_∞-SVM (Formulation (12)) can be derived by first solving Problem (30) in order to obtain $\mathbf{s}_k$, $\mathbf{t}_k$, and $\hat{\mathbf{t}}_k$ $(k = 1, 2)$, and then by evaluating the expressions

$$\mathbf{v}_1 = (H^\top H + c_1 I)^{-1}\begin{pmatrix}A^\top\mathbf{s}_1 - B^\top\mathbf{s}_2 + \mathbf{t}_1 - \hat{\mathbf{t}}_1\\ \mathbf{e}_1^\top\mathbf{s}_1 - \mathbf{e}_2^\top\mathbf{s}_2\end{pmatrix} \tag{31}$$

and

$$\mathbf{v}_2 = (G^\top G + c_1 I)^{-1}\begin{pmatrix}-A^\top\mathbf{s}_1 + B^\top\mathbf{s}_2 + \mathbf{t}_2 - \hat{\mathbf{t}}_2\\ -\mathbf{e}_1^\top\mathbf{s}_1 + \mathbf{e}_2^\top\mathbf{s}_2\end{pmatrix} \tag{32}$$

that result from (27) and (28).

Table 1
Number of features, number of examples, and number of examples per class (minority; majority) for all seven datasets.

Dataset	#features	#examples	#class(min.,maj.)
SONAR	60	208	(97;111)
ALON	2,000	62	(22;40)
GRAVIER	2,905	168	(57;111)
ALIZADEH	4,026	96	(35;61)
POMEROY	7,128	60	(21;39)
WEST	7,129	49	(24;25)
SHIPP	7,129	77	(19;58)

Remark 2. Note that if $c_1 = 0$, i.e., there is no regularization term in the objective function of the twin $l_2 l_\infty$-SVM formulation, then it is possible that the matrices $H^\top H$ and $G^\top G$ may not be well conditioned. Hence, we should explicitly introduce a regularization term δI, with $\delta > 0$ a fixed small scalar, in order to avoid the possible ill-conditioning of the matrices $H^\top H$ and $G^\top G$. In our experiments we use $\delta = 10^{-7}$. In case the symmetric matrices $H^\top H$ and $G^\top G$ are positive definite, the dual of Problem (13) is given by

$$\begin{aligned} \min_{\substack{\mathbf{s}_k, \mathbf{t}_k, \hat{\mathbf{t}}_k \\ k=1,2}} \; & \frac{1}{2}\boldsymbol{\alpha}^\top \left(\bar{A}_1^\top (H^\top H)^{-1} \bar{A}_1 + \bar{A}_2^\top (G^\top G)^{-1} \bar{A}_2\right)\boldsymbol{\alpha} - \mathbf{s}_1^\top \mathbf{e}_1 - \mathbf{s}_2^\top \mathbf{e}_2 \\ \text{s.t. } & 0 \leq \mathbf{s}_k \leq c_2 \mathbf{e}_k, \;\; \mathbf{t}_k, \hat{\mathbf{t}}_k \geq 0, \;\; k = 1, 2, \\ & \mathbf{t}_1 + \mathbf{t}_2 + \hat{\mathbf{t}}_1 + \hat{\mathbf{t}}_2 = \lambda \mathbf{e}. \end{aligned} \tag{33}$$

The dual of the twin $l_2 l_\infty$-SVM method has two important properties. First, the RFE algorithm described in Section 2.3 is usually written in terms of the dual variables [29], and therefore the RFE strategy can be implemented based on Formulation (30). Additionally, kernel-based formulations are usually derived from the dual via the kernel trick, and Problem (11) can be useful for this task.

4. Experimental results

The proposed twin $l_2 l_\infty$-SVM methodology and its simplified version, twin l_∞-SVM, were applied to one dataset from the UCI Repository [32], the Sonar dataset, which was studied in the context of feature selection in [2], and six microarray datasets: Alon's colon cancer data [33], Gravier's breast cancer data [34], Alizadeh's lymphoma data [35], Pomeroy's central nervous system embryonal tumor data [36], West's breast cancer data [37], and Shipp's lymphoma data [38]. The relevant meta-data (the number of variables, the total sample size, and the number of observations per class) is presented in Table 1.

In Table 1, we observe that all studied datasets are high-dimensional, ranging from 60 to 7129 attributes. These datasets also have few examples, which make them more challenging in terms of the modelling process, making feature selection of utmost importance in such cases. All the datasets are relatively balanced in terms of the class distribution. For studies that combine feature selection and the class-imbalance problem, we refer the reader to [39].

4.1. Data preparation and model calibration

Together with our proposal, we show the Fisher Score method using standard SVM in its linear and kernel-based versions as the baseline classifier, the SVM-RFE method using linear and kernel-based SVM [21], the TWSVM-RFE strategy [22] using twin SVM and NH-SVM as baseline classifiers (we refer to the latter strategy as NHSVM-RFE), the l_1-TWSVM method (Formulation (7) and (8)), as well as the proposed twin $l_2 l_\infty$-SVM and twin l_∞-SVM methods.

The experimental setting follows: Leave-one-out cross-validation (LOO) was used in each dataset for model selection and validation purposes. The following values for parameters C (standard SVM), c_i with $i = 1, \ldots, 4$ (Twin SVM, NH-SVM, l_1-TWSVM, and the proposed method), and λ were explored before performing feature selection: $C, c_1, c_2, c_3, c_4, \lambda \in \{2^{-7}, 2^{-6}, \ldots, 2^6, 2^7\}$. The AUC (Area Under the Curve) was used as the performance metric. For kernel-based methods, we use the radial basis function (RBF) kernel with $\gamma = 1/2\sigma^2 = 1/r$, with r the number of selected variables. We set $c_1 = c_2$ for the methods Twin SVM, l_1-TWSVM, and twin $l_2 l_\infty$-SVM; and $c_3 = c_4$ for Twin SVM, and l_1-TWSVM to limit the grid search to a maximum of two parameters.

All feature selection methods are trained using all available attributes, and then a feature ranking is constructed. Feature selection was performed in a backward fashion on the training set, and the AUC was monitored for various subsets of selected variables of cardinality $n = \{20, 50, 100, 250, 500, 1000\}$, with the exception of the Sonar dataset ($n = \{5, 10, 20, 30, 40, 50\}$).

Table 2
Maximum LOO AUC over all subsets of selected attributes, in percentage, for Sonar, Alon, Gravier, and Alizadeh datasets.

Method	SONAR		ALON		GRAVIER		ALIZADEH	
	AUC	n^*	AUC	n^*	AUC	n^*	AUC	n^*
Fisher+SVM(l)	80.1	50	88.2	20	78.8	100	95.6	1000
SVM-RFE(l)	80.4	50	89.4	20	67.9	500	95.6	20
Fisher+SVM(k)	81.7	50	88.1	20	78.3	100	96.3	100
SVM-RFE(k)	82.3	30	89.4	100	75.7	500	96.3	100
L1-SVM	80.1	50	89.4	20	76.6	50	93.3	20
L1-TWSVM	80.7	50	92.7	20	76.2	50	96.3	20
TWSVM-RFE	80.5	50	89.4	50	77.5	50	94.6	20
NHSVM-RFE	80	20	88.2	20	74	50	87.8	20
Twin l_∞-SVM	80	20	**94.0**	**20**	**79.3**	**20**	95.6	100
Twin $l_2 l_\infty$-SVM	**91.0**	20	**94.0**	**20**	78.4	50	**98.5**	**50**

Table 3
Maximum LOO AUC over all subsets of selected attributes, in percentage, for Pomeroy, Westm and Shipp datasets.

Method	POMEROY		WEST		SHIPP	
	AUC	n^*	AUC	n^*	AUC	n^*
Fisher+SVM(l)	72.0	50	**89.8**	**20**	96.5	1000
SVM-RFE(l)	67.4	20	**89.8**	**20**	96.5	1000
Fisher+SVM(k)	72.0	50	**89.8**	**20**	96.5	1000
SVM-RFE(k)	73.4	100	79.4	20	97.4	20
L1-SVM	75.8	20	81.5	20	97.4	50
L1-TWSVM	70.9	50	81.6	50	100	50
TWSVM-RFE	72	50	81.7	20	**100**	**20**
NHSVM-RFE	74.5	250	71.5	20	**100**	**20**
Twin l_∞-SVM	75.8	20	85.8	500	**100**	**20**
Twin $l_2 l_\infty$-SVM	**77.1**	**50**	**89.8**	**20**	**100**	**20**

4.2. Results summary

A summary of the results obtained from our experiments is presented in Tables 2–4, in which the maximum and average leave-one-out AUC values among all subsets of n attributes and for all five microarray datasets are presented, respectively. On the one hand, the maximum performance provides the best single solution, and allows us to estimate its predictive power if implemented. The average performance, on the other hand, allows us to assess the stability of the feature selection process [39]. The best performance is highlighted in bold type. The value of $n^* \in \{20, 50, 100, 250, 500, 1000\}$ ($n^* \in \{5, 10, 20, 30, 40, 50\}$ for the Sonar dataset), the cardinality of the subset of selected variables that leads to higher AUC for each method, is also reported in Tables 2 and 3.

In Tables 2 and 3, we observe that our proposed methods, either twin l_∞-SVM or twin $l_2 l_\infty$-SVM, always achieves the best maximum performance on all the subsets of selected features. This is an important conclusion since these models are the ones that

Table 4
Average LOO AUC over all subsets of selected attributes, in percentage, for all datasets.

Method	SONAR	ALON	GRAVIER	ALIZADEH	POMEROY	WEST	SHIPP
Fisher+SVM(l)	71.1	86.1	73.6	93.7	62.7	72.8	92.4
SVM-RFE(l)	70.5	87	70.7	93.5	60.5	70.1	91.5
Fisher+SVM(k)	70.4	86.6	73.0	93.8	62.6	74.5	93
SVM-RFE(k)	**79.8**	87.4	70.6	93	63.3	66.3	96.1
L1-SVM	70.4	89.4	76.5	92.7	71.7	81.5	97.1
L1-TWSVM	77.9	90.8	74.1	95.6	66.0	81.3	99.9
TWSVM-RFE	77.7	88.4	75.4	69.3	67.5	68.7	96.1
NHSVM-RFE	78.0	88.7	73.7	70.0	68.4	68.1	97.1
Twin l_∞-SVM	77.5	**91.9**	76.9	95.6	**73.8**	83.5	**100**
Twin $l_2 l_\infty$-SVM	77.9	91	**78**	**98.5**	69.2	**83.6**	**100**

Table 5
Best combination of parameters for all feature selection methods.

Method	Parameter	SONAR	ALON	GRAVIER	ALIZADEH	POMEROY	SHIPP	WEST
Fisher+SVM(l)	C	2^0	2^{-3}	2^2	2^{-6}	2^0	2^1	2^1
SVM-RFE(l)	C	2^4	2^{-4}	2^{-3}	2^{-6}	2^{-1}	2^{-1}	2^0
Fisher+SVM(k)	C	2^7	2^1	2^7	2^1	2^5	2^7	2^5
SVM-RFE(k)	C	2^7	2^1	2^7	2^1	2^3	2^7	2^0
L1-SVM	C	2^4	2^{-2}	2^{-1}	2^{-2}	2^0	2^1	2^{-2}
L1-TWSVM	$c_1 = c_2$	2^0	2^1	2^{-1}	2^{-3}	2^2	2^1	2^{-1}
	$c_3 = c_4$	2^3	2^1	2^{-1}	2^{-4}	2^0	2^1	2^{-1}
TWSVM-RFE	$c_1 = c_2$	2^{-2}	2^{-4}	2^{-6}	2^{-2}	2^{-5}	2^{-1}	2^{-6}
	$c_3 = c_4$	2^0	2^0	2^2	2^1	2^2	2^{-1}	2^{-1}
NHSVM-RFE	c_1	2^2	2^{-4}	2^{-1}	2^2	2^{-4}	2^0	2^{-3}
	c_2	2^0	2^{-2}	2^{-7}	2^{-4}	2^2	2^{-1}	2^{-7}
Twin l_∞-SVM	λ	2^-6	2^{-4}	2^{-4}	2^0	2^7	2^{-6}	2^2
	c_2	2^{-4}	2^4	2^{-7}	2^0	2^7	2^0	2^{-7}
Twin $l_2 l_\infty$-SVM	λ	2^7	2^4	2^0	2^1	2^7	2^{-6}	2^7
	$c_1 = c_2$	2^0	2^{-5}	2^{-1}	2^3	2^{-3}	2^2	2^{-5}

are eventually implemented for decision-making. A comparison between the two proposed strategies suggests that relatively similar performance can be obtained, and therefore the l_2-regularization can be omitted from the model.

In Table 4, we observe that the best average performance is also achieved by our methods in all datasets, with the only exception being the Sonar dataset, demonstrating the robustness of our strategy in terms of consistently identifying the relevant variables while achieving good predictive performance in terms of AUC. A comparison between the proposed twin l_∞-SVM and twin $l_2 l_\infty$-SVM also confirms our previous result which suggested that relatively similar performance can be achieved without using the Euclidean norm as regularizer.

The final parameters for all feature selection methods are reported in Table 5.

4.3. Model complexity and running times

The proposed approaches have a similar complexity compared with NH-SVM since the inclusion of the l_∞-regularization can be cast into a linear expression in the objective function without affecting the convexity of the problem. Our model, however, includes n decision variables (vector $\mathbf{z}$), and $2n$ constraints in order to construct a smooth, quadratic problem.

A comparison in terms of running times, in seconds, is provided in Table 6 for all the methods and datasets. For each fold of the LOO-crossvalidation, the running time is computed for obtaining all the solutions of different subsets of size n^*, using the best configuration of parameters for each method. Subsequently, a mean running time is obtained by simply averaging all running times for each fold. All experiments were performed on an HP Envy dv6 with 16 GB RAM, a i7-2620M processor with 2.70 GHz, 750GB SSD, and using Microsoft Windows 10.1 Operating System (64-bits). In terms of implementations, the methods l_1-SVM, l_1-TWSVM, and the proposed twin $l_2 l_\infty$-SVM and twin l_∞-SVM methods were developed using a generic solver (CVX, see [40]); LIBSVM [41] was used for standard SVM-based methods; and the codes by Yuan-Hai Shao et al., authors of Twin-Bounded SVM [11], were used for twin SVM and NH-SVM classifiers. These codes are publicly available at http://www.optimal-group.org/. The twin SVM and NH-SVM methods were also implemented using the CVX solver for comparison purposes.

It can be observed first in Table 6 that all running times are tractable, with about one minute being the longest training time. Furthermore, there is a gap between highly optimized codes, such as the LIBSVM toolbox and the twin implementations by Shao et al., and the use of a generic solver such as CVX. For example, l_1-SVM is the most efficient method in theory, but it has significantly longer running times compared with l_2-SVM and twin l_2-SVM approaches. In the same direction, our proposals are slower than NHSVM-RFE and TWSVM-RFE when the implementations by Shao et al. are used (see the first set of results in Table 6), but they are roughly similar when CVX is used (see the second set of results in Table 6). As future work, the development of an efficient optimization strategy for our proposal to reduce running times it is suggested, following the work by Shao et al. [11].

4.4. Feature selection performance and synchronization

One of the hypotheses of this work is that our proposals perform a coordinated feature selection, enabling each twin classifier to have similar relevant variables in its functions. In contrast, approaches like twin SVM that construct the classification models independently will necessarily identify different relevant variables in each twin function. In order to study this hypothesis, we trained each method and sorted the absolute values of the weights of both twin classifiers in a descending order for TWSVM-RFE, NHSVM-RFE, L1-TWSVM, twin $l_2 l_\infty$-SVM, and twin l_∞-SVM methods. Then, we identified the $n = \{20, 50, 100, 250, 500, 1000\}$ ($n \in \{5, 10, 20, 30, 40, 50\}$ for the Sonar dataset) most relevant variables

Table 6
Average running times, in seconds, for all methods and datasets.

Method	SONAR	ALON	GRAVIER	ALIZADEH	POMEROY	WEST	SHIPP
Fisher+SVM(l)	0″.06	0″.53	2″.61	0″.69	1″.80	1″.41	1″.05
SVM-RFE(l)	0″.06	0″.11	0″.47	0″.13	0″.17	0″.11	0″.23
Fisher+SVM(k)	0″.05	0″.63	2″.39	0″.98	1″.06	3″.41	1″.02
SVM-RFE(k)	0″.16	0″.08	0″.50	0″.28	0″.22	0″.14	0″.22
L1-SVM	6″.30	6″.81	14″.38	11″.23	8″.98	8″.91	11″.09
L1-TWSVM	13″.45	17″.64	32″.33	24″.25	21″.88	20″.14	25″.91
TWSVM-RFE[a]	0″.45	1″.94	3″.22	5″.95	18″.13	18″.20	20″.03
NHSVM-RFE[a]	0″.91	1″.73	4″.30	6″.42	20″.14	20″.38	19″.86
TWSVM-RFE[b]	16″.53	19″.80	22″.88	21″.05	20″.14	24″.25	23″.13
NHSVM-RFE[b]	2″.84	12″.13	22″.38	15″.11	13″.83	14″.92	17″.78
Twin l_∞-SVM	2″.89	13″.58	72″.44	38″.42	30″.33	26″.16	37″.38
Twin $l_2 l_\infty$-SVM	2″.98	12″.20	66″.19	32″.03	21″.78	20″.36	48″.63

[a] Method implemented using the codes by Shao et al.
[b] Method implemented using the CVX solver.

Table 7
Maximum Pearson's correlation over all subset of selected attributes for all seven datasets.

Method	SONAR	ALON	GRAVIER	ALIZADEH	POMEROY	WEST	SHIPP
TWSVM-RFE	0.85	0.65	0.30	0.23	0.63	0.75	0.49
NHSVM-RFE	0.56	0.55	0.62	0.21	0.82	0.79	0.56
L1-TWSVM	1.00	0.96	0.96	0.94	1.00	0.99	0.95
Twin l_∞-SVM	0.76	1.00	1.00	1.00	1.00	1.00	1.00
Twin $l_2 l_\infty$-SVM	1.00	1.00	0.98	1.00	1.00	1.00	1.00

Table 8
Average Pearson's correlation over all subset of selected attributes for all seven datasets.

Method	SONAR	ALON	GRAVIER	ALIZADEH	POMEROY	WEST	SHIPP
TWSVM-RFE	0.58	0.59	0.23	0.20	0.58	0.69	0.42
NHSVM-RFE	0.48	0.49	0.58	0.20	0.79	0.76	0.42
L1-TWSVM	0.40	0.76	0.73	0.57	0.79	0.86	0.48
Twin l_∞-SVM	0.53	0.80	0.93	0.93	1.00	1.00	0.90
Twin $l_2 l_\infty$-SVM	1.00	0.82	0.83	1.00	0.97	1.00	0.88

in each function (the n highest absolute values), and created two binary variables that reflect this selection (1 if variable j is relevant for twin classifier $k = 1, 2$; 0 otherwise). Finally, we computed the Pearson's correlation [42] between both indicator vectors as a measure of synchronized feature selection: a higher correlation means the two twin classifiers are identifying the same variables as relevant, while a low (or negative) correlation means that both classifiers are inconsistent in terms of the variables that are useful in the construction of the models.

As in the previous experiments, Tables 7 and 8 present a summary of the results obtained from our experiments, in which the maximum and average Pearson's correlations for the different subsets of attributes of cardinality n and for all datasets are presented, respectively.

In Tables 7 and 8, we can observe clearly that our proposals achieved a higher average and maximum Pearson's correlation compared to the alternative methods, being close to one in both metrics in all datasets. In contrast, both TWSVM-RFE and NHSVM-RFE have much lower values for this measure, with NHSVM-RFE slightly better than TWSVM-RFE at identifying the same feature in the twin functions. This result is to be expected since NHSVM solves a single optimization problem while twin SVM constructs both hyperplanes independently. We confirmed our hypothesis, concluding that the l_∞ regularization is very useful for performing a coordinated feature selection, identifying the same features as relevant in each twin classifier.

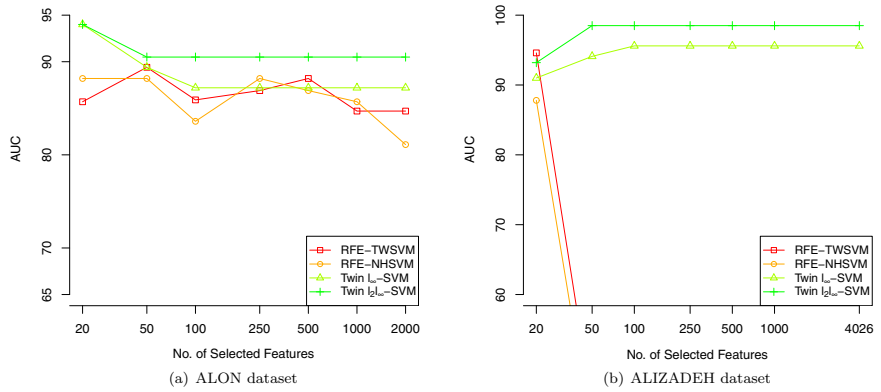
Next, we report the spinodal for all the subsets studied of n variables for the Alon and Alizadeh datasets. For the sake of space and visibility, we focus only in these two datasets, which are also the best-known ones in the feature selection literature (see e.g. [6,24,43]), and in the twin SVM approaches. For these two datasets, Fig. 1 illustrates the performance in terms of AUC for an increasing number of selected features and for all feature selection approaches studied.

In Fig. 1, we observe that twin $l_2 l_\infty$-SVM is consistently better than the other methods in terms of AUC for the Alon and Alizadeh datasets, not being the best only when 20 attributes on the latter dataset are used. Interestingly, predictive performance actually improves when using 20 attributes for the Alon dataset, in contrast to what happens with the Alizadeh dataset. We conclude that twin $l_2 l_\infty$-SVM is a very effective and robust classification approach, which leads to best predictive performance on average compared to the alternative methods studied here, while also achieving very stable results for an increasing number of selected attributes.

Similar to those in our previous analysis, Fig. 2(a) and (b) show the Pearson's correlation for an increasing number of selected features and for all feature selection approaches studied based on twin SVM, for the Alon and Alizadeh datasets, respectively.

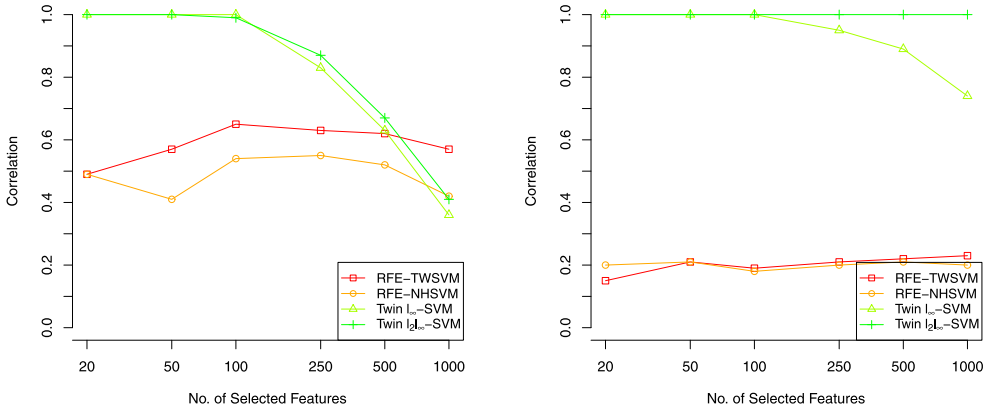
In Fig. 2, we observe that the correlation is almost 1 in both datasets for a number of selected variables of 100 or less. The synchronization improves when selecting few attributes for the Alon dataset (Fig. 2(a)), while it is consistently high and close to the unit for the Alizadeh dataset (Fig. 2(b)). Thus, we conclude that this coordinated feature elimination leads to positive predictive results when using a limited number of selected variables.

Finally, the influence of the parameters is discussed for the proposed Twin l_∞-SVM and $l_2 l_\infty$-SVM methods. For the Alon and Alizadeh datasets, the AUC is reported when parameters λ and C are varied, when the remaining parameter is fixed at its best value.



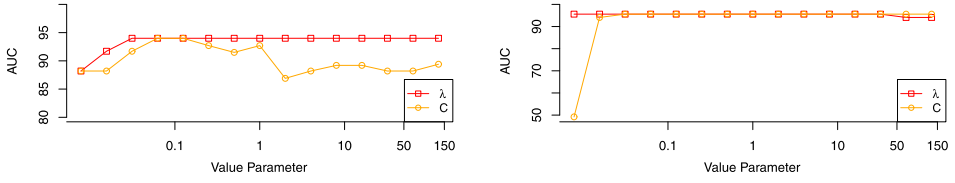
(a) ALON dataset (b) ALIZADEH dataset

Fig. 1. Performance (AUC) versus n for various feature selection approaches.



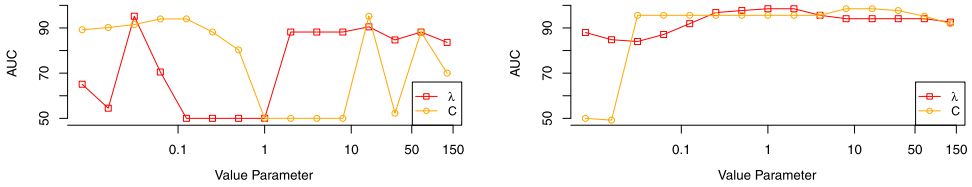
(a) ALON dataset (b) ALIZADEH dataset

Fig. 2. Pearson's correlation versus n for various feature selection approaches.



(a) ALON dataset (b) ALIZADEH dataset

Fig. 3. Sensitivity analysis for Twin l_∞-SVM. Parameters λ and $c_2 = C$.



(a) ALON dataset (b) ALIZADEH dataset

Fig. 4. Sensitivity analysis for Twin $l_2 l_\infty$-SVM. Parameters λ and $c_2 = c_1 = C$.

Figs. 3 and 4 present these results for the proposed Twin l_∞-SVM and $l_2 l_\infty$-SVM methods, respectively.

In Figs. 3 and 4, we observe that performance remains relatively stable when varying the parameters related to the Twin l_∞-SVM method, while Twin $l_2 l_\infty$-SVM presents more unstable behavior, especially for the Alon dataset. It can be concluded that only a proper grid search can guarantee positive predictive results.

5. Conclusions

In this work, we presented a novel embedded feature selection method for twin SVM, where the l_∞ regularization is used to perform a coordinated feature elimination in each twin classifier. Two strategies were proposed to pursue this goal: the use of the l_∞-norm as a sole regularizer, and in combination with the l_2-norm. Although both proposed strategies achieved best performance on

average compared to well-known feature selection strategies like the Fisher Score or TWSVM-RFE, the twin $l_2 l_\infty$-SVM method was slightly better than twin l_∞-SVM in general. This result demonstrates the importance of using both regularizers in the objective function of the twin SVM problem.

From the experimental section we conclude that, besides good predictive performance, our proposal is designed to foster the selection of similar relevant variables in each twin hyperplane in order to make the elimination step straightforward. This is an important advantage compared to the RFE strategy, which measures the contribution of an attribute as the average of its weights (in magnitude), and therefore a variable can be relevant in one of the functions, but it can be removed due to its aggregated importance. Our experiments demonstrate that this synchronization is important in order to achieve best predictive results with consistently fewer attributes.

Our proposal can be applied to any binary classification task, although it is designed to deal with high-dimensional datasets in which the interpretation of the results plays an important role. One of the disadvantages of twin SVM classification is that the construction of two hyperplanes makes the interpretation of the effect of the relevant variables trickier compared with that of linear methods. One of the main virtues of our work is that it enables us to gain insight into the process that generates the data, enhancing interpretability. Business analytics is a very important area in which choosing the relevant variables of a given task leads to important managerial insights. Suitable analytics applications are, for example, credit scoring [44] or churn prediction [45]. Our approach, however, is limited to linear classifiers.

There are interesting opportunities for future research in the following directions: First, the proposal can be extended to deal with class-imbalance problems. In this context, twin SVM has interesting properties when facing a skewed class distribution since both training patterns are treated independently. This issue is present in several analytics applications, such as credit scoring, fraud detection, and churn prediction [46]. In these cases, feature selection can be a major contribution in order to gain insight into the process that generated the data [39]. Secondly, our approach can also be extended to multi-class classification, where high-dimensional applications such as gene selection of multiple types of cancer [47] can commonly be found. Twin SVM has already been extended to multi-class classification [48], and the l_∞ regularization can be used to coordinate the feature selection process of all hyperplanes that needs to be constructed. Finally, there is a pressing need for more efficient implementation of classification methods, especially in high-dimensional domains. Our approach can be adapted to make it more efficient computationally, using, for example, linear programming formulations [49], or efficient optimization strategies, such as the Frank–Wolfe algorithm [50].

Acknowledgements

The first author was supported by CONICYT, FONDECYT project 1160738, while the second one was funded by CONICYT, FONDECYT project 1160894. This research was partially funded by the Complex Engineering Systems Institute, ISCI (ICM-FIC: P05-004-F, CONICYT: FB0816).

References

[1] L. Bai, Z. Wang, Y.-H. Shao, N.-Y. Deng, A novel feature selection method for twin support vector machine, Knowl. Based Syst. 59 (0) (2014) 1–8.
[2] S. García, J. Luengo, F. Herrera, Tutorial on practical tips of the most influential data preprocessing algorithms in data mining, Knowl. Based Syst. 98 (2016) 1–29.
[3] E. Carrizosa, D. Romero-Morales, Supervised classification and mathematical optimization, Comput. Oper. Res. 40 (1) (2013) 150–165.
[4] S. Maldonado, M. Pérez J.and Labbé, R. Weber, Feature selection for support vector machines via mixed integer linear programming, Inf. Sci. 279 (2014) 163–175.
[5] V. Vapnik, Statistical Learning Theory, John Wiley and Sons, 1998.
[6] S. Maldonado, R. Weber, J. Basak, Kernel-penalized SVM for feature selection, Inf. Sci. 181 (1) (2011) 115–128.
[7] D. Bertsimas, A. King, R. Mazumder, Best subset selection via a modern optimization lensbertsimas2016, Ann. Stat. 44 (2) (2016) 813–852.
[8] H.L. Thi, M.L. Hoai, T.P. Dinh, Feature selection in machine learning: an exact penalty approach using a difference of convex function algorithm, Mach. Learn. 101 (1) (2015) 163–186.
[9] Q. Ye, C. Zhao, N. Ye, H. Zheng, X. Chen, A feature selection method for non-parallel plane support vector machine classification, Optim. Methods Softw. 27 (3) (2012) 431–443.
[10] Jayadeva, R. Khemchandani, S. Chandra, Twin support vector machines for pattern classification, IEEE Trans. Pattern Anal. Mach. Intell. 29 (5) (2007) 905–910.
[11] Y. Shao, C. Zhang, X. Wang, N. Deng, Improvements on twin support vector machines, Neural Netw. IEEE Trans. 22 (6) (2011) 962–968.
[12] Y. Shao, W. Chen, N. Deng, Nonparallel hyperplane support vector machine for binary classification problems, Inf. Sci. 263 (0) (2014) 22–35.
[13] Y. Yang, H. Zou, A fast unified algorithm for solving group-lasso penalize learning problems, Stat. Comput. 25 (6) (2015) 1129–1141.
[14] M. García-Torres, F. Gómez-Vela, B. Melián-Batista, J.M. Moreno-Vega, High--dimensional feature selection via feature grouping: a variable neighborhood search approach, Inf. Sci. 326 (1) (2016) 102–118.
[15] H. Zou, M. Yuan, The f-infinite norm support vector machine, Stat. Sin. 18 (2008) 379–398.
[16] Y. Xu, L. Wang, A weighted twin support vector regression, Knowl. Based Syst. 33 (2012) 92–101.
[17] D. Tomar, S. Agarwal, A comparison on multi-class classification methods based on least squares twin support vector machine, Knowl. Based Syst. 81 (2015) 131–147.
[18] M. Carrasco, J. López, S. Maldonado, A second-order cone programming formulation for nonparallel hyperplane support vector machine, Expert Syst. Appl. 54 (2016) 95–104.
[19] J. Saketha Nath, C. Bhattacharyya, Maximum margin classifiers with specified false positive and false negative error rates, in: Proceedings of the SIAM International Conference on Data mining, 2007.
[20] R. Duda, P. Hard, D. Stork, Pattern Classification, Wiley-Interscience Publication, 2001.
[21] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines., Mach. Learn. 46 (1–3) (2002) 389–422.
[22] Z. Yang, J. He, Y. Shao, Feature selection based on linear twin support vector machines, Procedia Comput. Sci. 17 (2013) 1039–1046.
[23] P. Bradley, O. Mangasarian, Feature selection via concave minimization and support vector machines, in: Machine Learning proceedings of the fifteenth International Conference (ICML'98) 82–90, San Francisco, California, Morgan Kaufmann, 1998.
[24] J. Neumann, C. Schnörr, G. Steidl, Combined svm-based feature selection and classification, Mach. Learn. 61 (1–3) (2005) 129–150.
[25] H.A.L. Thi, M. Nguyen, Advanced Computational Methods for Knowledge Engineering, in: Studies in Computational Intelligence, 479, Springer International Publishing, 2013, pp. 41–52.
[26] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J. R. Stat. Soc. Ser. B 68 (2006) 49–67.
[27] O. Chapelle, S. Keerthi, Multi-class feature selection with support vector machines, 2008.
[28] J. Platt, Advances in Kernel Methods-Support Vector Learning, MIT Press, Cambridge, MA, 1999, pp. 185–208.
[29] B. Schölkopf, A.J. Smola, Learning with Kernels, MIT Press, 2002.
[30] K. Bennett, E. Bredensteiner, Duality and geometry in svm classifiers, in: In Proc. 17th International Conf. on Machine Learning, Morgan Kaufmann, 2000, pp. 57–64.
[31] O.L. Mangasarian, Nonlinear Programming, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, 1994.
[32] K. Bache, M. Lichman, UCI machine learning repository, 2013, Url: http://archive.ics.uci.edu/ml.
[33] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, A. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc. Natl. Acad. Sci. 96 (12) (1999) 6745–6750.
[34] E. Gravier, G. Pierron, A. Vincent-Salomon, N. Gruel, V. Raynal, A. Savignoni, Y. De Rycke, J.-Y. Pierga, C. Lucchesi, F. Reyal, A. Fourquet, S. Roman-Roman, F. Radvanyi, X. Sastre-Garau, O. Asselain B.and Delattre, A prognostic dna signature for t1t2 node-negative breast cancer patients, Genes, Chromosomes Cancer 49 (12) (2010). 1125–1125.
[35] A. Alizadeh, M. Eisen, R. Davis, et al., Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling, Nature 403 (2000) 503–511.
[36] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, L.M. Sturla, M. Angelo, M.E. McLaughlin, J.Y.H. Kim, L.C. Goumnerova, P.M. Black, C. Lau, J. Allen, D. Zagzag, J. Olson, T. Curran, C. Wetmore, J.A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D.N. Louis, J. Mesirov, E. Lander, T. Golub, Prediction of central nervous system embryonal tumour outcome based on gene expression, Nature 415 (6870) (2002) 436–442.

[37] M.M. West, C.C. Blanchette, H.H. Dressman, E.E. Huang, S.S. Ishida, R.R. Spang, H. Zuzan, J. Olson, J. Marks, J. Nevins, Predicting the clinical status of human breast cancer by using gene expression profiles, Proc. Natl. Acad. Sci. U.S.A. 98 (20) (2001) 11462–11467.
[38] M.A. Shipp, K.N. Ross, P. Tamayo, A.P. Weng, J.L. Kutok, R.C.T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G.S. Pinkus, T.S. Ray, M.A. Koval, K.W. Last, A. Norton, T.A. Lister, J. Mesirov, D.S. Neuberg, E.S. Lander, J.C. Aster, T.R. Golub, Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning, Nat. Med. 8 (1) (2002) 68–74.
[39] S. Maldonado, F. Famili, R. Weber, Feature selection for high-dimensional class-imbalanced data sets using support vector machines, Inf. Sci. 286 (2014) 228–246.
[40] M. Grant, S. Boyd, CVX: Matlab software for disciplined convex programming, version 2.1, 2014, (http://cvxr.com/cvx).
[41] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[42] K. Pearson, Notes on regression and inheritance in the case of two parentsnotes on regression and inheritance in the case of two parents, in: Proceedings of the Royal Society of London, 58, 1895, pp. 240–242.
[43] A. Rakotomamonjy, Variable selection using SVM-based criteria, J. Mach. Learn. Res. 3 (2003) 1357–1370.
[44] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, B. Baesens, New insights into churn prediction in the telecommunication sector: a profit driven data mining approach, Eur. J. Oper. Res. 218 (1) (2012) 211–229.
[45] Z.-Y. Chen, Z.-P. Fan, Distributed customer behavior prediction using multiplex data: a collaborative mk-svm approach, Knowl. Based Syst. 35 (2012) 111–119.
[46] B. Baesens, Analytics in a Big Data World, John Wiley and Sons, 2014.
[47] T. Lin, R. Liu, C. Chen, Y. Chao, S. Chen, Pattern classification in DNA microarray data of multiple tumor types, Pattern Recognit. 39 (12) (2006) 2426–2438.
[48] Y. Xu, R. Guo, L. Wang, A twin multi-class classification support vector machines, Cognit. Comput. 5 (2013) 580–588.
[49] N. Djuric, L. Lan, S. Vucetic, Z. Wang, Budgetedsvm: a toolbox for scalable svm approximations, J. Mach. Learn. Res. 14 (2013) 3813–3817.
[50] R. Ñanculef, E. Frandi, C. Sartori, H. Allende, A novel frank-wolfe algorithm. analysis and applications to large-scale svm training, Inf. Sci. 285 (2014) 66–99.