Neurocomputing 235 (2017) 112–121

Contents lists available at ScienceDirect

Neurocomputing

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

Group-penalized feature selection and robust twin SVM classification via second-order cone programming

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

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

ARTICLE INFO

Communicated by Feiping Nie

Keywords:
Support vector machines
Feature selection
Twin SVM
Second-order cone programming
Group penalty

ABSTRACT

Selecting the relevant factors in a particular domain is of utmost interest in the machine learning community. This paper concerns the feature selection process for twin support vector machine (TWSVM), a powerful classification method that constructs two nonparallel hyperplanes in order to define a classification rule. Besides the Euclidean norm, our proposal includes a second regularizer that aims at eliminating variables in both twin hyperplanes in a synchronized fashion. The baseline classifier is a twin SVM implementation based on second-order cone programming, which confers robustness to the approach and leads to potentially better predictive performance compared to the standard TWSVM formulation. The proposal is studied empirically and compared with well-known feature selection methods using microarray datasets, on which it succeeds at finding low-dimensional solutions with highest average performance among all the other methods studied in this work.

1. Introduction

Robustness has been a relevant topic in the SVM literature in recent years [11,30]. Second-order cone programming (SOCP) [2,23] is a popular convex optimization approach that has been used to develop robust maximum margin classifiers [5,29]. In the scheme presented by Nath and Bhattacharyya [29], the worst data distribution is assumed for a given mean and covariance matrix, while each training pattern is classified correctly for predefined false positive and false negative error rates. This strategy has demonstrated superior predictive performance thanks to its robust framework [5,25,29].

Twin support vector machine (TWSVM) [19] has gained popularity in recent years due to its performance and geometrical properties. For binary classification, this strategy constructs two nonparallel hyperplanes in such a way that each one is close to one of the two training patterns, and as far as possible from the other. The two classifiers can either be constructed independently by using two optimization problems [19], or simultaneously by using a single model. This latter approach was proposed in Shao et al. [34], and is known as Nonparallel Hyperplane SVM (NH-SVM).

Feature selection is an important task in the process of knowledge discovery [22]. The right identification of the relevant attributes and the removal of noisy variables lead to more effective classifiers in terms of predictive performance, faster training, and relevant insight for decision-making gained from the process that generates the data [15,16,27]. Feature selection is particularly useful in high-dimensional applications like text classification [18] and bioinformatics [37].

A plethora of feature selection methods has been proposed for SVM (see e.g. [16,27]). For twin SVM, however, feature selection is more challenging since the two hyperplanes are constructed independently, leading to different subsets of relevant variables for each classifier. Some feature selection methods have been extended to twin SVM, such as the well-known SVM-RFE [40] and l_1-SVM methods [4,15,41]. These methods lead to sparse solutions but do not perform synchronized feature selection in such a way that a subset of common relevant attributes is detected.

In this work, we propose a robust SOCP formulation for twin SVM that includes a group penalization term, shrinking the weights toward zero at the attribute level. The approach combines the ideas of Nonparallel Hyperplane SVM [34] as a SOCP formulation, as is presented in Carrasco et al. [9], and the use of the infinite norm as group penalty function for embedded feature selection [44].

This paper is structured as follows: previous work on twin SVM, also including its robust formulation based on second-order cones, is discussed in Section 2. In Section 3, an overview of feature selection methods for SVM classification is presented. The proposed strategy for simultaneous feature selection and robust twin SVM classification is described in Section 4. In Section 5, experimental results using high-dimensional microarray datasets are given. Finally, the main conclusions of this study are presented in Section 6.

* Corresponding author.
E-mail address: smaldonado@uandes.cl (S. Maldonado).

http://dx.doi.org/10.1016/j.neucom.2017.01.005
Received 30 June 2016; Received in revised form 2 November 2016; Accepted 1 January 2017
Available online 05 January 2017

2. Twin SVM classification

In this section, previous research on twin SVM is presented. First, the standard twin SVM is described. Subsequently, the Nonparallel Hyperplane SVM formulation is reviewed. Finally, the robust formulation that extends NH-SVM to SOCP (RNH-SVM), which represents the base classifier for our approach, is discussed at the end of this section.

2.1. Twin support vector machine

The original twin SVM formulation proposed by Jayadeva [19] constructs two quadratic programming (QP) problems that aim at classifying each class correctly. This approach generates two linear nonparallel hyperplanes of the form $\mathbf{w}_k^\top \mathbf{x} + b_k = 0$, k=1,2, in such a way that each one is closer to the data samples of one of the classes and as far as possible from those of the other class. Formally, let us denote by m_k the cardinality of class k=1,2, and by $A \in \Re^{m_1 \times n}$ ($B \in \Re^{m_2 \times n}$) the data matrix related to the positive (negative) class. The twin SVM model follows [35]:

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

and

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

where c_1, c_2, c_3, and c_4 are positive parameters, and $\mathbf{e}_1$ and $\mathbf{e}_2$ are vectors of ones of appropriate dimensions. Formulation (1) and (2) is the one proposed by Shao et al. [35] (Twin-Bounded SVM), which is equivalent to the original twin SVM model proposed by Jayadeva et al. [19] when setting $c_1 = c_2 = \epsilon$, with $\epsilon > 0$ a fixed small parameter. A new sample $\mathbf{x}$ is assigned to class k^* according to its proximity to the hyperplanes based on the following rule:

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

where d_k is the perpendicular distance of the data point $\mathbf{x}$ from hyperplane $\mathbf{w}_k^\top \mathbf{x} + b_k = 0$, k=1,2.

2.2. Nonparallel hyperplane SVM (NH-SVM)

The NH-SVM method [34] solves a single QP problem, constructing the two hyperplanes $\mathbf{w}_k^\top \mathbf{x} + b_k = 0$ simultaneously. The linear NH-SVM model follows:

$$\min_{\substack{\mathbf{w}_k, b_k, \boldsymbol{\xi}_k \\ k=1,2}} \frac{1}{2}(\| A\mathbf{w}_1 + \mathbf{e}_1 b_1 \|^2 + \| B\mathbf{w}_2 + \mathbf{e}_2 b_2 \|^2) + \frac{c_1}{2}(\| \mathbf{w}_1 \|^2 + b_1^2 + \| \mathbf{w}_2 \|^2 + b_2^2) + c_3(\mathbf{e}_1^\top \boldsymbol{\xi}_1 + \mathbf{e}_2^\top \boldsymbol{\xi}_2)$$
$$\text{s.t.} \quad A\mathbf{w}_1 + \mathbf{e}_1 b_1 - A\mathbf{w}_2 - \mathbf{e}_1 b_2 \geq \mathbf{e}_1 - \boldsymbol{\xi}_1,$$
$$B\mathbf{w}_2 + \mathbf{e}_2 b_2 - B\mathbf{w}_1 - \mathbf{e}_2 b_1 \geq \mathbf{e}_2 - \boldsymbol{\xi}_2,$$
$$\boldsymbol{\xi}_1 \geq 0, \boldsymbol{\xi}_2 \geq 0, \tag{4}$$

where c_1 and c_3 are positive parameters [34]. Equivalent to twin SVM, a new data sample $\mathbf{x}$ in $\Re^n$ is assigned to k^* by identifying the nearest of both hyperplanes according to Equation (3).

2.3. Robust nonparallel hyperplane SVM (RNH-SVM)

A robust nonparallel hyperplane SVM version based on second-order cones (SOCs) was presented by Carrasco et al. [9]. This method (RNH-SVM) extends the ideas of the NH-SVM approach by constructing two nonparallel classifiers simultaneously, in such a way that each hyperplane is close to one class and far away from the other class. The main difference between both methods lies in the strategy used to represent the two hyperplanes: while NH-SVM considers the reduced convex hulls, each training pattern is represented by ellipsoids in RNH-SVM.

Formally, let $\mathbf{X}_1$ and $\mathbf{X}_2$ be random vectors that generate the data objects of the positive and negative classes, respectively, with means and covariance matrices given by $(\boldsymbol{\mu}_k, \Sigma_k)$ for k=1,2, where $\Sigma_k \in \Re^{n \times n}$ are symmetric positive semidefinite matrices. The two linear nonparallel hyperplanes can be obtained by solving the following quadratic chance-constrained programming problem:

$$\min_{\substack{\mathbf{w}_k, b_k \\ k=1,2}} \frac{1}{2}(\| A\mathbf{w}_1 + \mathbf{e}_1 b_1 \|^2 + \| B\mathbf{w}_2 + \mathbf{e}_2 b_2 \|^2) + \frac{\theta}{2}(\| \mathbf{w}_1 \|^2 + b_1^2 + \| \mathbf{w}_2 \|^2 + b_2^2)$$
$$\text{s.t.} \quad \inf_{\mathbf{X}_1 \sim (\boldsymbol{\mu}_1, \Sigma_1)} \Pr\{\mathbf{X}_1 \in H^+(\mathbf{w}_1 - \mathbf{w}_2, b_1 - b_2)\} \geq \eta_1,$$
$$\inf_{\mathbf{X}_2 \sim (\boldsymbol{\mu}_2, \Sigma_2)} \Pr\{\mathbf{X}_2 \in H^-(\mathbf{w}_1 - \mathbf{w}_2, b_1 - b_2)\} \geq \eta_2, \tag{5}$$

where $\theta > 0$, $\eta_k \in (0, 1)$ (k=1,2), and

$$H^+(\mathbf{w}, b) := \{\mathbf{x}: \mathbf{x}^\top \mathbf{w} + b \geq 1\}, \quad H^-(\mathbf{w}, b) := \{\mathbf{x}: \mathbf{x}^\top \mathbf{w} + b \leq -1\}.$$

The constraints are used to assure that the two hyperplanes, H^+ and H^-, classify the instances correctly from both classes up to the rate η_k (k=1,2) under a probabilistic scheme.

The chance-constrained problem can be cast to a deterministic problem by using the multivariate Chebyshev inequality [21, Lemma 1]. The RNH-SVM formulation follows:

$$\min_{\substack{\mathbf{w}_k, b_k \\ k=1,2}} \frac{1}{2}(\| A\mathbf{w}_1 + \mathbf{e}_1 b_1 \|^2 + \| B\mathbf{w}_2 + \mathbf{e}_2 b_2 \|^2) + \frac{\theta}{2}(\| \mathbf{w}_1 \|^2 + b_1^2 + \| \mathbf{w}_2 \|^2 + b_2^2)$$
$$\text{s.t.} \quad (\mathbf{w}_1 - \mathbf{w}_2)^\top \boldsymbol{\mu}_1 + (b_1 - b_2) \geq 1 + \kappa_1 \| S_1^\top (\mathbf{w}_1 - \mathbf{w}_2) \|,$$
$$-((\mathbf{w}_1 - \mathbf{w}_2)^\top \boldsymbol{\mu}_2 + (b_1 - b_2)) \geq 1 + \kappa_2 \| S_2^\top (\mathbf{w}_1 - \mathbf{w}_2) \|, \tag{6}$$

where $\kappa_k = \sqrt{\frac{\eta_k}{1 - \eta_k}}$ and $\Sigma_k = S_k S_k^\top$, for k=1,2. This problem is an instance of quadratic SOCP with two SOC constraints [2]. Generally speaking, an SOC constraint on the variable $\mathbf{x} \in \Re^n$ is of the form $\| D\mathbf{x} + \mathbf{b} \| \leq \mathbf{c}^\top \mathbf{x} + d$, where $d \in \Re$, $\mathbf{c} \in \Re^n$, $\mathbf{b} \in \Re^m$, and $D \in \Re^{m \times n}$ are given.

3. Feature selection for SVM

In this section, we refer to the best-known strategies for feature selection for SVM classification, whose use as benchmark approaches will be shown in the empirical section; the concept of group penalty, which represents the cornerstone of our proposal; and previous research in feature selection for both SOCP classifiers for SVM classification and twin SVM, and their main differences compared with our proposal.

3.1. The Fisher score

This method is a statistical measure used to rank the variables according to their contribution before applying any classifier. The Fisher Score computes the absolute value of the difference between the mean of both classes for each variable, and divides it by a joint standard deviation [13]:

$$F(j) = \left| \frac{\mu_j^+ - \mu_j^-}{(\sigma_j^+)^2 + (\sigma_j^-)^2} \right|, \tag{7}$$

where μ_j^+ (resp. μ_j^-) is the mean for the j-th attribute in the positive (resp. negative) class and σ_j^+ (resp. σ_j^-) is the respective standard deviation. Support Vector Machines can be applied over the subset of relevant features (i.e. with highest Fisher Score).

3.2. Recursive feature elimination SVM

The SVM-RFE approach uses the SVM solution to rank the variables according to their contribution in the SVM margin [17]. A backward elimination scheme is developed in order to remove those variables whose elimination leads to the largest margin. The margin can be rewritten in terms of the dual variables $\boldsymbol{\alpha} \in \Re^m$ as follows:

$$W^2(\boldsymbol{\alpha}) = \sum_{i,s=1}^{m} \alpha_i \alpha_s y_i y_s \mathbf{x}_i \cdot \mathbf{x}_s. \tag{8}$$

Following the previous description, the value of $|W^2(\boldsymbol{\alpha}) - W^2_{(-p)}(\boldsymbol{\alpha})|$ represents the contribution of variable p in the margin, where $W^2_{(-p)}(\boldsymbol{\alpha})$ corresponds to the margin when variable p is removed from the data matrix [17].

3.3. Group penalty functions

The idea of the group penalty function is to penalize a group of related weights together in such a way that sparsity is encouraged at a variable level instead of removing weights independently [42]. Such strategies are well-known in binary classification with categorical attributes with multiple levels, which are usually transformed into sets of dummy variables. It is desirable to remove the full set of dummy variables to enhance interpretability [42]. Another application is multiclass classification in which several classifiers are constructed in order to shatter each class. Feature selection can be performed simultaneously at a variable level, jointly penalizing all the weights related to one attribute in each classification function [10].

The best-known group penalty is called *group-lasso*. This function has the following form:

$$\Gamma(\mathbf{w}) = \sum_{j=1}^{J} \sqrt{p_j} \parallel \mathbf{w}^{(j)} \parallel_2, \tag{9}$$

where $\parallel \mathbf{w}^{(j)} \parallel_2 = \sqrt{\sum_{l \in \mathcal{I}_j} w_l^2}$. The measure $\mathcal{I}_j$ represents disjoint sets of related features linked to a given attribute $j = 1, \ldots, J$, where $|\mathcal{I}_j| = p_j$ is the total number of levels for nominal variables, or the number of classifiers constructed for the case of multiclass classification, and $\sum_{j=1}^{J} p_j = n$ represents the total number of estimated weights.

Another strategy for group penalization is known as the $l_\infty - \text{norm}$ penalty [44], which has the following form:

$$\Gamma(\mathbf{w}) = \sum_{j=1}^{J} \parallel \mathbf{w}^{(j)} \parallel_\infty, \tag{10}$$

where $\parallel \mathbf{w}^{(j)} \parallel_\infty = \max_{l \in \mathcal{I}_j}\{|w_l|\}$. The $l_\infty - \text{norm}$ penalty was originally developed for dealing with categorical variables in binary SVM classification, under the name of $F_\infty - \text{norm}$ SVM. The formulation can be cast into a LP problem by introducing decision variables of the form $t_j = \parallel \mathbf{w}^{(j)} \parallel_\infty$, and adding a new set of constraints $|w_l| \leq t_j$ for each $l \in \mathcal{I}_j$ and $j = 1, \ldots, J$. This is an important property in our proposal since it reduces the complexity of the problem; which is why this alternative was chosen instead of the lasso penalty.

3.4. Feature selection for SOCP-based maximum margin classifiers

Some feature selection approaches have been proposed for SVM classifiers based on the concept of ellipsoids proposed by Nath and Bhattacharyya [29]. For example, Bhattacharyya [5] replaced the Euclidean norm as a regularization approach by the l_1-norm for binary classification, extending the ideas of Bradley and Mangasarian [6] for standard SVM to SOCP. The method provides a good compromise between regularization and sparsity [26], leading to very good predictive performance. The l_1-SOCP formulation follows:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \parallel \mathbf{w} \parallel_1 \\ \text{s.t.} \quad & \mathbf{w}^\top \boldsymbol{\mu}_1 + b \geq 1 + \kappa_1 \parallel S_1^\top \mathbf{w} \parallel, \\ & -(\mathbf{w}^\top \boldsymbol{\mu}_2 + b) \geq 1 + \kappa_2 \parallel S_2^\top \mathbf{w} \parallel, \end{aligned} \tag{11}$$

where $\parallel \mathbf{w} \parallel_1 = \sum_{i=1}^{n} |w_i|$ denotes the l_1-norm of $\mathbf{w}$.

The SVM-RFE has also been extended to the robust framework already discussed for binary [26] and multiclass classification [24]. These approaches follow the same ideas: for binary classification, the variables with less impact in the margin are removed in a backward elimination process; while the squared sum of all weights related to a given variable is used as the contribution metric for multiclass learning. None of the previous approaches have been applied in either twin classification or in tasks in which coordinated feature selection at a variable level is required.

3.5. Feature selection for twin SVM

Like for SOCP-SVM approaches, some feature selection methods developed for twin SVM have been reported in the literature. The SVM-RFE approach was extended for twin SVM in Yang et al. [40], in which $W^2(\boldsymbol{\alpha})$ (see Eq. (8)) is redefined as the sum of the weights linked with attribute j in both twin classifiers. Relevant variables are the ones with a value higher for this sum in absolute value: $|w_{1j}^*| + |w_{2j}^*|$, in which all weights are normalized as follows: $w_{lj}^* = \frac{|w_{lj}|}{\parallel w_{lj} \parallel_2}$, for l=1,2 and $j = 1, \ldots, n$.

The popular l_1-SVM method [6] has been extended to standard twin SVM [4] and other variations, such as least squares twin SVM [15,41]. The main difference between standard twin SVM and least squares twin SVM is that the latter uses a quadratic loss function instead of the traditional hinge loss to penalize the slack variables $\boldsymbol{\xi}$ that control model fit.

For the sake of completeness, we propose the l_1-NHSVM method, which adapts the NH-SVM model presented in Formulation (4) to perform embedded feature selection by replacing the l_2 regularization on $\mathbf{w}_1$ and $\mathbf{w}_2$ by the LASSO penalty. The main difference between this proposal and l_1 twin SVM [4] is that the latter solves both twin problems independently. The l_1-NHSVM formulation follows:

$$\begin{aligned} \min_{\substack{\mathbf{w}_k, b_k, \boldsymbol{\xi}_k \\ k=1,2}} \quad & \frac{1}{2}(\parallel A\mathbf{w}_1 + \mathbf{e}_1 b_1 \parallel^2 + \parallel B\mathbf{w}_2 + \mathbf{e}_2 b_2 \parallel^2) + c_1(\parallel \mathbf{w}_1 \parallel_1 + \parallel \mathbf{w}_2 \parallel_1) \\ & + c_3(\mathbf{e}_1^\top \boldsymbol{\xi}_1 + \mathbf{e}_2^\top \boldsymbol{\xi}_2) \\ & \text{s.t.} \quad A\mathbf{w}_1 + \mathbf{e}_1 b_1 - A\mathbf{w}_2 - \mathbf{e}_1 b_2 \geq \mathbf{e}_1 - \boldsymbol{\xi}_1, \\ & B\mathbf{w}_2 + \mathbf{e}_2 b_2 - B\mathbf{w}_1 - \mathbf{e}_2 b_1 \geq \mathbf{e}_2 - \boldsymbol{\xi}_2, \\ & \boldsymbol{\xi}_1 \geq 0, \boldsymbol{\xi}_2 \geq 0, \end{aligned} \tag{12}$$

where $c_1, c_3 > 0$.

The main issue with the current state of the art is that none of these approaches perform a coordinated feature selection at a variable level; each twin problem is solved independently, leading to different subsets of relevant features in each twin classifier. To the best of our knowledge, no approach that includes group penalty has been developed for twin SVM. Our approach also differs from other methods based on double regularization for embedded feature selection, such as elastic net [43], and other combinations of the l_2-norm and the l_1-norm [7,31], for the same reason: such methods do not perform a coordinated feature selection at a variable level.

4. A novel SOCP method for simultaneous twin feature selection

In this section, we propose a novel method for embedded feature selection and robust twin SVM classification. The main idea is to include a group penalty, namely $l_\infty - \text{norm}$ penalization, in the RNH-SVM method in order to achieve a coordinated elimination of variables in both hyperplanes, conferring sparsity to the robust twin SVM method.

The RNH-SVM model has an important advantage over the standard twin formulation for the following reason: instead of splitting the problem into two different QPPs, it uses a single optimization problem to construct the two twin hyperplanes, allowing the use of a group penalty function. Regarding the group penalty function, we choose $l_\infty-$norm penalization over the group LASSO because the resulting optimization problem is less complex: the latter strategy requires additional conic constraints in order to cast the non-smooth function presented in Eq. (9) into a smooth convex optimization problem, while the $l_\infty-$norm penalization requires only linear constraints for this purpose, as detailed in Section 3.3.

Formally, let us consider the following formulation:

$$\begin{aligned}\min_{\substack{\mathbf{w}_k,b_k\\k=1,2}}\ & \frac{1}{2}(\| A\mathbf{w}_1+\mathbf{e}_1b_1\|^2+\|B\mathbf{w}_2+\mathbf{e}_2b_2\|^2)+\frac{\theta}{2}(\|\mathbf{w}_1\|^2+b_1^2+\|\mathbf{w}_2\|^2+b_2^2)\\ &+\lambda\sum_{j=1}^{n}\|\mathbf{w}_{(j)}\|_\infty\\ \text{s.t.}\quad & (\mathbf{w}_1-\mathbf{w}_2)^\top\boldsymbol{\mu}_1+(b_1-b_2)\geq 1+\kappa_1\|S_1^\top(\mathbf{w}_1-\mathbf{w}_2)\|,\\ & -((\mathbf{w}_1-\mathbf{w}_2)^\top\boldsymbol{\mu}_2+(b_1-b_2))\geq 1+\kappa_2\|S_2^\top(\mathbf{w}_1-\mathbf{w}_2)\|,\end{aligned}\tag{13}$$

where $\kappa_k=\sqrt{\frac{\eta_k}{1-\eta_k}}$ for k=1,2, $\mathbf{w}_{(j)}=(w_{1j},w_{2j})\in\Re^2$, and $\|\mathbf{w}_{(j)}\|_\infty=\max_{k=1,2}|w_{kj}|$, for $j=1,\ldots,n$. It can be noticed that the above formulation corresponds to the RNH-SVM method (cf. formulation (6)) with the inclusion of the $l_\infty-$norm regularization term (cf. Eq. (10)) in the objective function. Parameters θ, and λ control the trade-off between l_2 regularization (margin maximization), model fit, and sparsity. Both constraints are used to guarantee that each twin hyperplane is close to one of the ellipsoids representing the training patterns, and as far as possible from the other.

In order to avoid the use of a non-smooth function in the previous problem, we cast Formulation (13) into a quadratic SOCP problem by introducing an additional variable $\mathbf{z}\in\Re^n$, leading to the following formulation:

$$\begin{aligned}\min_{\substack{\mathbf{w}_k,b_k,\mathbf{z}\\k=1,2}}\ & \frac{1}{2}(\|A\mathbf{w}_1+\mathbf{e}_1b_1\|^2+\|B\mathbf{w}_2+\mathbf{e}_2b_2\|^2)\\ &+\frac{\theta}{2}(\|\mathbf{w}_1\|^2+b_1^2+\|\mathbf{w}_2\|^2+b_2^2)\quad+\lambda\mathbf{e}^\top\mathbf{z}\quad\text{s.t.}\quad(\mathbf{w}_1-\mathbf{w}_2)^\top\boldsymbol{\mu}_1\\ &+(b_1-b_2)\geq 1+\kappa_1\|S_1^\top(\mathbf{w}_1-\mathbf{w}_2)\|,\\ &-((\mathbf{w}_1-\mathbf{w}_2)^\top\boldsymbol{\mu}_2+(b_1-b_2))\geq 1+\kappa_2\|S_2^\top(\mathbf{w}_1-\mathbf{w}_2)\|,\\ &\mathbf{z}-\mathbf{w}_k\geq 0,\ \mathbf{z}+\mathbf{w}_k\geq 0,\quad k=1,2,\end{aligned}\tag{14}$$

where $\mathbf{e}$ denotes a vector of ones of dimension n. We refer to this formulation as l_2l_∞–RNH–SVM.

4.1. Dual formulation of l_2l_∞–RNH–SVM and geometric interpretation

In this section, the dual formulation of the l_2l_∞–RNH–SVM is presented, which is derived from Formulation (14) in the Appendix, and provides geometrical insights into the method. The dual formulation for l_2l_∞–RNH–SVM is given by:

$$\begin{aligned}\max_{\substack{t_k,\mathbf{z}_k,\mathbf{u}_k,\boldsymbol{\alpha}_k,\boldsymbol{\beta}_k\\k=1,2}}\ & \tilde{\mathbf{e}}^\top\mathbf{t}-\frac{1}{2}(Z\mathbf{t}-(\hat{\boldsymbol{\alpha}}_1-\hat{\boldsymbol{\beta}}_1))^\top(H^\top H+\theta I)^{-1}(Z\mathbf{t}-(\hat{\boldsymbol{\alpha}}_1-\hat{\boldsymbol{\beta}}_1))\\ &-\frac{1}{2}(Z\mathbf{t}+(\hat{\boldsymbol{\alpha}}_2-\hat{\boldsymbol{\beta}}_2))^\top(G^\top G+\theta I)^{-1}(Z\mathbf{t}+(\hat{\boldsymbol{\alpha}}_2-\hat{\boldsymbol{\beta}}_2))\\ \text{s.t.}\quad & \mathbf{z}_1=\boldsymbol{\mu}_1-\kappa_1S_1\mathbf{u}_1,\ \|\mathbf{u}_1\|\leq 1,\\ & \mathbf{z}_2=\boldsymbol{\mu}_2+\kappa_2S_2\mathbf{u}_2,\ \|\mathbf{u}_2\|\leq 1,\\ & t_k,\boldsymbol{\alpha}_k,\boldsymbol{\beta}_k\geq 0,\quad k=1,2,\\ & \boldsymbol{\alpha}_1+\boldsymbol{\alpha}_2+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2=\lambda\mathbf{e},\end{aligned}\tag{15}$$

where

$$\tilde{\mathbf{e}}=\begin{pmatrix}1\\1\end{pmatrix},\quad \mathbf{t}=\begin{pmatrix}t_1\\t_2\end{pmatrix},\quad \hat{\boldsymbol{\alpha}}_k=\begin{pmatrix}\boldsymbol{\alpha}_k\\0\end{pmatrix},\quad\text{and}\quad \hat{\boldsymbol{\beta}}_k=\begin{pmatrix}\boldsymbol{\beta}_k\\0\end{pmatrix}.$$

Remark 1. The optimal value for $\mathbf{t}$ can be obtained by fixing variables $\mathbf{z}_k$, $\mathbf{u}_k$, $\boldsymbol{\alpha}_k$, and $\boldsymbol{\beta}_k$ (for k=1,2); and solving the following linear system:

$$Z^\top[(H^\top H+\theta I)^{-1}+(G^\top G+\theta I)^{-1}]Z\mathbf{t}=\mathbf{c},\tag{16}$$

where

$$\mathbf{c}=\tilde{\mathbf{e}}+Z^\top(H^\top H+\theta I)^{-1}(\hat{\boldsymbol{\alpha}}_1-\hat{\boldsymbol{\beta}}_1)-Z^\top(G^\top G+\theta I)^{-1}(\hat{\boldsymbol{\alpha}}_2-\hat{\boldsymbol{\beta}}_2).$$

The solution of the linear system (16) allows us to rewrite the dual problem (15) as

$$\begin{aligned}\max_{\substack{\mathbf{z}_k,\mathbf{u}_k,\boldsymbol{\alpha}_k,\boldsymbol{\beta}_k\\k=1,2}}\ & \frac{1}{2}\mathbf{c}^\top(Z^\top[(H^\top H+\theta I)^{-1}+(G^\top G+\theta I)^{-1}]Z)^{-1}\mathbf{c}-\\ &\frac{1}{2}(\|(H^\top H+\theta I)^{-1/2}(\hat{\boldsymbol{\alpha}}_1-\hat{\boldsymbol{\beta}}_1)\|^2+\|(G^\top G+\theta I)^{-1/2}(\hat{\boldsymbol{\alpha}}_2-\hat{\boldsymbol{\beta}}_2)\|^2)\\ \text{s.t.}\quad & \mathbf{z}_1\in\mathbf{B}(\boldsymbol{\mu}_1,S_1,-\kappa_1),\\ & \mathbf{z}_2\in\mathbf{B}(\boldsymbol{\mu}_2,S_2,\kappa_2),\\ & \boldsymbol{\alpha}_k,\boldsymbol{\beta}_k\geq 0,\quad k=1,2,\\ & \boldsymbol{\alpha}_1+\boldsymbol{\alpha}_2+\boldsymbol{\beta}_1+\boldsymbol{\beta}_2=\lambda\mathbf{e},\end{aligned}\tag{17}$$

where

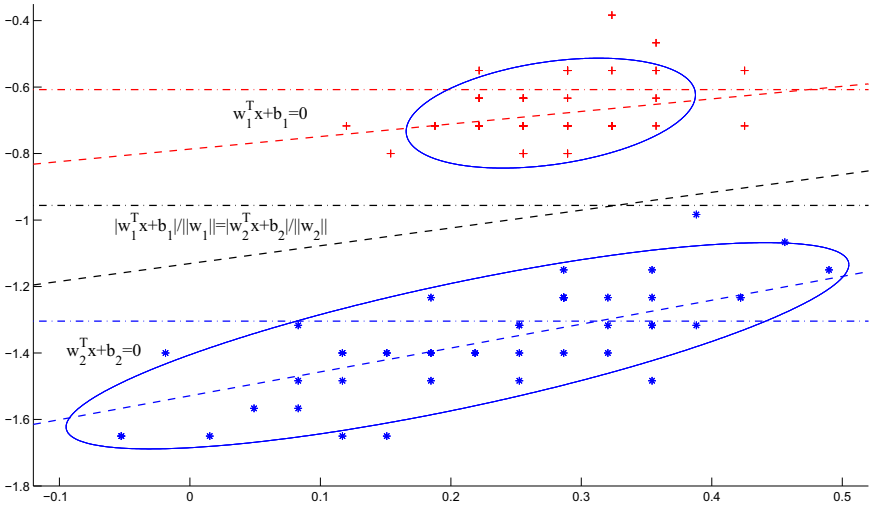

Fig. 1. Geometric interpretation for RNH-SVM and l_2l_∞–RNH–SVM.

Table 1
Number of features, number of examples, and number of examples per class for all five datasets.

Dataset	#features	#examples	#class (min., maj.)
ALON	2000	62	(22;40)
GRAVIER	2905	168	(57;111)
ALIZADEH	4026	96	(35;61)
POMEROY	7128	60	(21;39)
WEST	7129	49	(24;25)

$$\mathbf{B}(\boldsymbol{\mu}, S, \kappa) = \{\mathbf{z}: \mathbf{z} = \boldsymbol{\mu} + \kappa S\mathbf{u}, \| \mathbf{u} \| \leq 1\}. \quad (18)$$

The set $\mathbf{B}(\boldsymbol{\mu}, S, \kappa)$ denotes an ellipsoid centered at $\boldsymbol{\mu}$ whose shape is determined by S, and size by κ.

Remark 2. Formulation (15) (similarly (17)) tells us that the dual problem of (14) can be seen as the maximization of a function subject to a set of constraints that defines two ellipsoids, including inequality and affine equality constraints.

Fig. 1 presents the geometrical interpretation of RNH-SVM and l_2l_∞–RNH–SVM in a two-dimensional toy data set. The attribute presented on the X-axis is an irrelevant variable generated at random, having completely overlapping class conditional densities. The attribute on the Y-axis was generated to be relevant, resulting in two disjoint clumps. The dashed lines represent the three hyperplanes constructed with RNH-SVM: the two nonparallel classifiers over the training patterns, and the one that defines the decision rule between both twin hyperplanes. Similarly, the dot-dash lines correspond to the hyperplanes defined by l_2l_∞–RNH–SVM.

Two important points can be noticed in this example. First, the ellipsoids that represent each training pattern, the non-parallel hyperplanes, and the final classifier resulting from the margin maximization can be clearly distinguished by both methods. Additionally, the main difference between both approaches is easy to recognize: the RNH-SVM method uses both attributes in the construction of the non-parallel hyperplanes, while our proposal l_2l_∞–RNH–SVM ignores the irrelevant variable, leading to completely flat hyperplanes. This is the result of the synchronized feature selection process.

5. Experimental results

The proposed l_2l_∞–RNH–SVM methodology was applied to five microarray datasets for binary classification. This section is organized as follows: a description of the experimental setting and the datasets used in this work is provided in Section 5.1. Subsequently, a performance summary is presented in Section 5.2, in which the best results among different subsets of variables are analyzed. In Section 5.3, the predicted performance for each subset of variables is studied for each method. The usefulness of the group penalty in terms of providing a synchronized feature selection is discussed in Section 5.4. Finally, a sensitivity analysis is performed for the relevant parameters of the proposed l_2l_∞–RNH–SVM method in order to illustrate the influence of these parameters and the stability of the method. These experiments are discussed in Section 5.5.

5.1. Experimental setting and datasets

The following experimental procedure was performed: leave-one-out cross-validation (LOO) was used in each dataset for model selection and validation purposes, using the Area Under the Curve (AUC) as performance metric. Feature selection was performed on the training set. For each method, the performance was monitored for various subsets of features of the following cardinality: $n = \{20, 50, 100, 250, 500, 1000\}$. The RFE strategy for twin SVM classification described in Section 3.5 was applied to the twin SVM, NH-SVM, l_1-NHSVM, and l_2l_∞–RNH–SVM methods in order to obtain subsets of n variables.

Regarding the validation procedure, the following values were explored for the parameters C (soft-margin SVM); c_1, c_2, c_3, and c_4 (twin SVM, where $c_1 = c_2$ and $c_3 = c_4$); c_1 and c_3 (NH-SVM and l_1-NHSVM); and θ and λ (l_2l_∞–RNH–SVM): $\{2^{-7}, 2^{-6}, \ldots, 2^6, 2^7\}$. This validation procedure was performed for each baseline classifier before applying feature selection.

Results from the following alternative feature selection methods are reported, along with the proposed l_2l_∞–RNH–SVM: Fisher Score as a filter strategy for SVM (Fisher+SVM), the RFE-SVM method, the TWSVM-RFE strategy [40] using twin SVM and NH-SVM as baseline classifiers, the l_1-NHSVM method, and Bhattacharyya's l_1-SOCP-SVM method for robust classification and feature selection [5].

The following datasets were used in the analysis: Alon's colon cancer data [3], Gravier's breast cancer data [14], Alizadeh's lymphoma data [1], West's breast cancer data [38], and Pomeroy's central nervous system embryonal tumor data [33]. The relevant descriptive information is presented for each dataset in Table 1, including the total number of variables and examples, and the number of observations per class.

Table 2
Average LOO AUC over all subsets of selected attributes, in percentages, for all five datasets.

Method	ALON	GRAVIER	ALIZADEH	POMEROY	WEST
Fisher+SVM	86.1	73.6	93.7	62.7	72.8
RFE-SVM	87.0	70.7	93.5	60.5	70.1
RFE-TWSVM	88.4	75.4	69.3	67.5	68.7
RFE-NHSVM	88.7	73.7	70.0	68.4	68.1
l_1-NHSVM	91.0	**79.5**	95.6	73.3	85.1
l_1-SOC-SVM	**91.7**	78.3	95.4	64.1	87.5
l_2l_∞–RNH–SVM	**91.7**	78.2	**98.3**	**71.3**	**89.5**

5.2. Performance summary

The average and maximum AUC is reported in Tables 2, 3, respectively, for all predefined subsets of attributes and for all datasets. The average AUC can be considered as a measure for model stability: it is desirable not only as a model that finds a single best solution while exploring various subsets of features, but also as model that consis-

Table 3
Maximum LOO AUC over all subsets of selected attributes, in percentages, for all five datasets.

Method	ALON	GRAVIER	ALIZADEH	POMEROY	WEST
Fisher+SVM	88.2	78.8	95.6	72.0	89.8
RFE-SVM	89.4	74.9	95.6	67.4	89.8
RFE-TWSVM	93.0	78.4	94.6	76.9	81.7
RFE-NHSVM	**95.0**	77.5	95.3	75.8	85.7
l_1-NHSVM	94.0	80.6	97.1	78.0	**100.0**
l_1-SOC-SVM	91.7	78.3	97.1	66.1	87.8
l_2l_∞–RNH–SVM	94.0	**80.1**	**98.5**	**81.7**	91.8

Table 4
Holm's post-hoc test for pairwise comparisons.

Method	Mean rank	Mean AUC	p value	$\alpha/(k-i)$	Action
l_2l_∞–RNH–SVM	1.7	89.22	–	–	Not reject
l_1-NHSVM	1.8	89.94	0.9417	0.0500	Not reject
Fisher+SVM	4.6	84.88	0.0338	0.0250	Not reject
RFE-NHSVM	4.6	85.86	0.0338	0.0167	Not reject
l_1-SOC-SVM	4.9	84.20	0.0192	0.0125	Not reject
RFE-TWSVM	5.0	84.92	0.0157	0.0100	Not reject
RFE-SVM	5.4	83.42	0.0068	0.0083	Reject

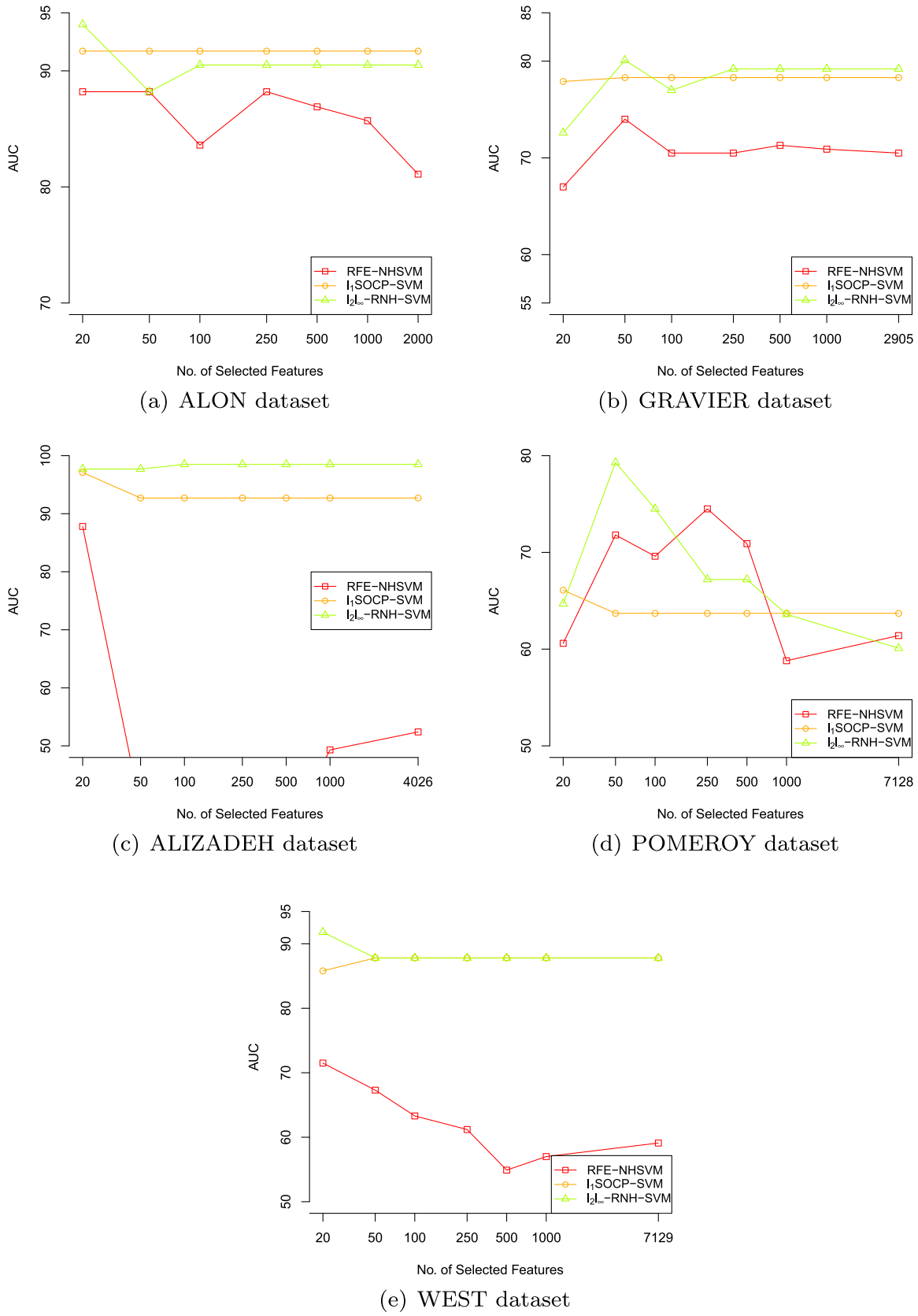

Fig. 2. Performance (AUC) for an increasing number of features. All datasets.

tently achieves good performance for all trials. The highest AUC is highlighted in bold type for both datasets.

In Tables 2, 3, we observe that the best overall performance is achieved with the proposed method considering both the average and maximum AUC: l_2l_∞−RNH−SVM has the highest average predictive performance in four of the five datasets, and the highest maximum

predictive performance in three of the five datasets. The proposed approach is slightly below l_1-NHSVM in terms of average AUC for the GRAVIER dataset, and slightly below RFE-NHSVM and l_1-NHSVM in terms of maximum AUC for the ALON and WEST datasets, respectively.

In order to demonstrate that our approach has the best overall performance, we compute the average rank for each method based on the maximum AUC (Table 3). The Holm's test is used to study statistical significance, as suggested by Demšar [12]. This test performs pairwise comparisons between each technique and the one with the best performance. This analysis is reported in Table 4.

From the experiments presented in Table 4, we conclude that the proposed l_2l_∞–RNH–SVM achieves the best overall performance. Our approach, however, is not able to outperform all the others statistically; only RFE-SVM is statistically worse than l_2l_∞–RNH–SVM. Notice also that the two methods proposed in this work (l_2l_∞–RNH–SVM and l_1-NHSVM) are very close in terms of average ranking (1.7 and 1.8 respectively), being very far from the remaining methods.

5.3. Feature selection performance

Next, the feature selection performance is detailed by plotting the AUC for an increasing number of selected attributes n for all datasets. For visualization purposes, only the most relevant alternative methods were selected for comparison: RFE-NHSVM and l_1-SOC-SVM, which are also the ones that have either the highest average or maximum performance in any of the datasets. The RFE-NHSVM method is a relevant benchmark since it shares the same classification principle with our proposal (twin SVM solved in a single optimization problem), while l_1-SOC-SVM is relevant because it is the best-known embedded feature selection method for SOCP-based SVM classification. These graphs are presented in Fig. 2.

In Fig. 2, it can be seen that the proposed l_2l_∞–RNH–SVM method has better overall performance compared to the two alternatives, although no method outperforms others in any of the datasets. It can be noticed, however, that l_2l_∞–RNH–SVM has a higher or similar AUC compared to RFE-NHSVM and l_1-SOC-SVM, for all subsets of variables in the ALIZADEH and WEST datasets. We conclude that l_2l_∞–RNH–SVM is a very stable and powerful predictive method for embedded feature selection and SVM classification thanks to its robust framework, leading to best average performance in high-dimensional datasets.

Table 5
Average Pearson's correlation over all subsets of selected attributes for all five datasets.

Method	ALON	GRAVIER	ALIZADEH	POMEROY	WEST
TWSVM	0.59	0.23	0.20	0.58	0.69
NHSVM	0.49	0.58	0.20	0.79	0.76
l_1-NHSVM	0.51	0.58	0.52	0.92	1.00
l_2l_∞–RNH–SVM	0.91	0.97	1.00	0.95	0.96

Table 6
Maximum Pearson's correlation over all subsets of selected attributes for all five datasets.

Method	ALON	GRAVIER	ALIZADEH	POMEROY	WEST
TWSVM	0.65	0.30	0.23	0.63	0.75
NH-SVM	0.55	0.62	0.21	0.82	0.79
l_1-NHSVM	0.70	0.69	0.62	1.00	1.00
l_2l_∞–RNH–SVM	1.00	1.00	1.00	0.97	1.00

5.4. Is the feature selection process actually synchronized?

In this work, it is hypothesized that the l_2l_∞–RNH–SVM method removes attributes in a synchronized fashion, in the sense that each twin classifier should have similar relevant variables in its functions. Twin SVM, by contrast, assigns weights to each variable independently while constructing the hyperplanes, and, therefore, high agreement in terms of variable relevancy cannot be expected. In this section this hypothesis is explored by computing the level of agreement for each method that constructs twin classifiers.

The following methodology was applied to assess the synchronization at feature weighting: each twin method was first trained, and the weights related to both hyperplanes were then sorted (using the absolute values) in descending order. Two binary vectors, one for each function, were created for each method, and for each subset of variables $n = \{20, 50, 100, 250, 500, 1000\}$, representing with a 1 a variable that has a weight that belongs to the n largest ones in magnitude, and with a 0 otherwise. Finally, the Pearson's correlation [32] between both binary variables was computed as a metric of coordinated feature selection. A value close to the unit for this measure indicates high synchronization since the two twin classifiers are identifying the same variables as relevant, while values close to zero indicate that the process is performed independently.

Similar to Section 5.2, the average and maximum Pearson's correlations are presented in Tables 5, 6, respectively, summarizing the synchronization for each subset of variables n in two values for all datasets and all twin approaches.

Strong agreement in the feature selection process for the proposed method can be observed in Tables 5, 6, with both average and

Table 7
AUC for different values of λ and θ, ALON dataset.

$\lambda\backslash\theta$	2^{-7}	2^{-6}	2^{-5}	2^{-4}	2^{-3}	2^{-2}	2^{-1}	2^{0}	2^{1}	2^{2}	2^{3}	2^{4}	2^{5}	2^{6}	2^{7}	MAX
2^{-7}	83.6	94.0	86.9	85.9	92.7	86.9	88.2	88.2	89.2	88.2	90.5	90.5	87.2	89.4	89.4	94.0
2^{-6}	87.2	88.2	94.0	90.5	88.2	88.2	86.9	87.2	89.4	89.2	91.7	90.5	85.9	87.2	89.4	94.0
2^{-5}	89.4	89.4	92.7	94.0	90.5	89.4	89.2	87.2	85.9	88.0	88.2	90.5	90.5	91.7	89.4	94.0
2^{-4}	89.4	91.7	87.2	90.5	92.7	85.9	86.9	88.2	86.9	88.2	90.5	88.2	89.4	88.2	88.2	92.7
2^{-3}	88.2	94.0	89.4	89.4	91.7	90.5	89.4	85.9	86.9	90.5	88.0	88.2	88.2	88.2	88.2	94.0
2^{-2}	88.2	88.2	88.2	89.4	85.9	90.5	89.4	89.4	88.2	88.2	91.7	88.2	88.2	89.4	90.5	91.7
2^{-1}	90.5	90.5	90.5	90.5	89.4	91.7	94.0	89.4	91.7	91.7	90.5	88.2	88.2	87.2	87.2	94.0
2^{0}	90.5	91.7	90.5	90.5	90.5	89.4	91.7	89.4	89.2	89.4	85.9	89.4	89.4	89.4	89.4	91.7
2^{1}	91.7	91.7	90.5	91.7	90.5	91.7	90.5	89.4	90.5	88.2	84.7	89.4	88.2	89.4	87.2	91.7
2^{2}	90.5	90.5	90.5	90.5	91.7	90.5	90.5	91.7	88.2	88.2	88.2	84.9	88.2	88.2	87.2	91.7
2^{3}	90.5	90.5	90.5	90.5	90.5	91.7	91.7	91.7	91.7	87.2	88.2	90.5	85.9	88.2	88.2	91.7
2^{4}	90.5	90.5	90.5	90.5	90.5	90.5	90.5	90.5	90.5	90.5	89.4	87.2	88.2	88.2	85.9	90.5
2^{5}	90.5	90.5	90.5	90.5	90.5	90.5	90.5	90.5	90.5	90.5	90.5	91.7	87.2	84.7	88.2	91.7
2^{6}	90.5	90.5	90.5	90.5	90.5	90.5	90.5	90.5	90.5	90.5	90.5	89.4	91.7	85.9	85.9	91.7
2^{7}	90.5	90.5	90.5	90.5	90.5	90.5	90.5	90.5	90.5	90.5	90.5	90.5	89.4	88.2	85.9	90.5
MAX	94.0	94.0	94.0	92.7	94.0	91.7	94.0	91.7	91.7	91.7	91.7	90.5	91.7	91.7	90.5	94.0

maximum Pearson's correlations close to the unit for all datasets. Significantly lower agreement is achieved by NH-SVM and then by TWSVM. For the case of l_1-NHSVM, strong agreement is only observed in two of the five datasets, demonstrating that the LASSO penalty does not guarantee synchronized feature elimination.

With this set of experiments the usefulness of a group penalty function to achieve a synchronized feature selection is confirmed, as is the advantage of using NH-SVM over TWSVM as the baseline classifier for this purpose.

5.5. Influence of the parameters

The proposed l_2l_∞–RNH–SVM method includes two new parameters for controlling the trade-off between complexity, model fit, and synchronized feature selection. These parameters, λ and θ, are set via grid search during the model selection procedure. In this section, we analyze how the performance of the model varies as a function of these parameters. For illustration purposes, we report the AUC for $\theta, \lambda \in \{2^{-7}, 2^{-6}, \ldots, 2^{6}, 2^{7}\}$ for the ALON data set. Similar analyses were conducted for the other data sets in order to assess whether or not the results are stable along different values of these parameters. These results are presented in Table 7.

In Table 7, we observe relatively stable results for the various values of θ and λ, although the highest difference between the lowest and the highest performance is significant (10.34% AUC). We conclude that an adequate grid search is highly recommended in order to guarantee adequate predictive results.

6. Conclusions

A novel method for simultaneous feature selection and twin SVM classification is presented in this work. The robust framework presented in Nath and Bhattacharyya [29] was adapted in order to provide a robust framework for the construction of twin classifiers. In this framework, a pessimistic approach is assumed, in which each training pattern needs to be classified correctly for predefined false positive and false negative error rates, even for the worst data distribution for a given mean and covariance matrix, leading to a single SOCP model. A group penalty function is included in it to perform a synchronized feature elimination in each twin classifier in such a way that only the variables that have a large weight in both hyperplanes are considered relevant. The l_∞ – norm is added to the objective function, leading to an SOCP model with three objectives: model fit, structural risk minimization, and coordinated feature selection.

Experiments were performed in high-dimensional genomic datasets for cancer prediction. Predictive performance was analyzed, and the proposed method achieved superior performance in general compared to well-known feature selection and SVM classification strategies. The proposed l_2l_∞–RNH–SVM has the highest AUC in four of the five datasets, also having stable results: the method has the best performance when both the AUC curves for an increasing number of selected variables and the average AUC among all studied feature subsets are analyzed. The effect of the group penalty function is assessed by Pearson's correlation between two binary vectors that represents the variable selection in each twin classifier. The proposed l_2l_∞–RNH–SVM performed a synchronized feature elimination effectively, constructing hyperplanes that has almost the same relevant variables in them (Pearson's correlation close to the unit), in contrast to the twin SVM and NH-SVM classifiers. The main issue with such approaches when performing RFE is that a backward elimination process forces the removal of potentially relevant attributes if they have a low average of both weights.

Regarding future developments, several research opportunities were identified from this work. First, the robust SOCP framework can be used in multiclass classification (see e.g. [24]), in which several hyperplanes are constructed to shatter each class. The most common strategies for multiclass classification are One-versus-All, One-versus-One, and all-together approaches. For the latter methods, all necessary hyperplanes are constructed via a single optimization problem [39], and feature selection can be performed in a coordinated fashion using a group penalty function [10]. Additionally, other penalty functions can be explored to perform embedded feature selection; for instance, a concave approximation of the l_0-norm [6] can be used, and, although it may lead to non-convex optimization problems, there are some strategies to deal with this issue in the conic programming literature [8,20]. Finally, the use of a generic solver like SeDuMi [36] for the SOCP implementation may bottleneck the use of the SOCP methods in large-scale problems, and therefore the development of more efficient implementations tailored for the structure of the SVM formulations is an interesting line of future research.

Acknowledgements

The first author was supported by FONDECYT project 1160894, while the second one was funded by FONDECYT projects 1140831 and 1160738. This research was partially funded by the Complex Engineering Systems Institute, ISCI (ICM-FIC: P05-004-F, CONICYT: FB0816). The authors are grateful to the anonymous reviewers who contributed to improving the quality of the original paper.

Appendix A. Derivation for the dual formulation of l_2l_∞–RNH–SVM

In this appendix, the dual formulation of the l_2l_∞–RNH–SVM problem (cf. Formulation (14)) is derived. Let L be the Lagrangian function for Problem (14), which is given by

$$L(\mathbf{w}_k, b_k, \mathbf{z}, t_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k) = \frac{1}{2}(\| A\mathbf{w}_1 + \mathbf{e}_1 b_1 \|^2 + \| B\mathbf{w}_2 + \mathbf{e}_2 b_2 \|^2) + \frac{\theta}{2}(\| \mathbf{w}_1 \|^2 + b_1^2 + \| \mathbf{w}_2 \|^2 + b_2^2) + \lambda \mathbf{e}^\top \mathbf{z} - t_1(\mathbf{w}_1 - \mathbf{w}_2)^\top \boldsymbol{\mu}_1 + t_2(\mathbf{w}_1 - \mathbf{w}_2)^\top \boldsymbol{\mu}_2$$
$$+ t_1(-(b_1 - b_2) + 1 + \kappa_1 \| S_1^\top(\mathbf{w}_1 - \mathbf{w}_2)\|) + t_2((b_1 - b_2) + 1 + \kappa_2 \| S_2^\top(\mathbf{w}_1 - \mathbf{w}_2)\|) + \sum_{k=1}^{2} (\boldsymbol{\alpha}_k^\top(\mathbf{w}_k - \mathbf{z}) - \boldsymbol{\beta}_k^\top(\mathbf{w}_k + \mathbf{z})), \tag{A.1}$$

where $t_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k \geq 0$, for k=1,2. Since $\| \mathbf{v} \| = \max_{\|\mathbf{u}\| \leq 1} \mathbf{u}^\top \mathbf{v}$ holds for any $\mathbf{v} \in \Re^n$, the Lagrangian can be rewritten as follows:

$$L(\mathbf{w}_k, b_k, \mathbf{z}, t_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k) = \max_{\mathbf{u}_1, \mathbf{u}_2} \{\widehat{L}(\mathbf{w}_k, b_k, \mathbf{z}, t_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k, \mathbf{u}_k) : \| \mathbf{u}_k \| \leq 1\},$$

with $\widehat{L}$ given by

$$\widehat{L}(\mathbf{w}_k, b_k, \mathbf{z}, t_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k, \mathbf{u}_k) = \frac{1}{2}(\| A\mathbf{w}_1 + \mathbf{e}_1 b_1 \|^2 + \| B\mathbf{w}_2 + \mathbf{e}_2 b_2 \|^2) + \frac{\theta}{2}(\| \mathbf{w}_1 \|^2 + b_1^2 + \| \mathbf{w}_2 \|^2 + b_2^2) + \lambda \mathbf{e}^\top \mathbf{z} - t_1(\mathbf{w}_1 - \mathbf{w}_2)^\top \boldsymbol{\mu}_1 + t_2(\mathbf{w}_1 - \mathbf{w}_2)^\top \boldsymbol{\mu}_2$$
$$+ t_1(-(b_1 - b_2) + 1 + \kappa_1(\mathbf{w}_1 - \mathbf{w}_2)^\top S_1 \mathbf{u}_1) + t_2((b_1 - b_2) + 1 + \kappa_2(\mathbf{w}_1 - \mathbf{w}_2)^\top S_2 \mathbf{u}_2) + \sum_{k=1}^{2} (\boldsymbol{\alpha}_k^\top(\mathbf{w}_k - \mathbf{z}) - \boldsymbol{\beta}_k^\top(\mathbf{w}_k + \mathbf{z})). \tag{A.2}$$

Thus, Problem (14) can be written equivalently as

$$\min_{\mathbf{w}_k, b_k, \mathbf{z}\mathbf{u}_k, t_k, \boldsymbol{\alpha}_k \boldsymbol{\beta}_k} \max \{\widehat{L}(\mathbf{w}_k, b_k, \mathbf{z}, t_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k, \mathbf{u}_k) : \| \mathbf{u}_k \| \le 1, t_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k \ge 0\},$$

and therefore the Wolfe-dual of Formulation (14) (see [28]) corresponds to

$$\max_{\mathbf{u}_k, t_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k} \{\widehat{L} : \nabla_{\mathbf{w}_k} \widehat{L} = 0, \nabla_{b_k} \widehat{L} = 0, \nabla_{\mathbf{z}} \widehat{L} = 0, \| \mathbf{u}_k \| \le 1, t_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k \ge 0\}. \tag{A.3}$$

The following linear system is obtained when computing the gradient of $\widehat{L}$ with respect to $\mathbf{w}_k$, b_k (k=1,2), and $\mathbf{z}$:

$$(A^\top A + \theta I)\mathbf{w}_1 + b_1 A^\top \mathbf{e}_1 + \boldsymbol{\alpha}_1 - \boldsymbol{\beta}_1 = t_1 \mathbf{z}_1 - t_2 \mathbf{z}_2, \tag{A.4}$$

$$(B^\top B + \theta I)\mathbf{w}_2 + b_2 B^\top \mathbf{e}_2 + \boldsymbol{\alpha}_2 - \boldsymbol{\beta}_2 = - t_1 \mathbf{z}_1 + t_2 \mathbf{z}_2, \tag{A.5}$$

$$\mathbf{e}_1^\top A \mathbf{w}_1 + b_1(\theta + \mathbf{e}_1^\top \mathbf{e}_1) = t_1 - t_2, \tag{A.6}$$

$$\mathbf{e}_2^\top B \mathbf{w}_2 + b_2(\theta + \mathbf{e}_2^\top \mathbf{e}_2) = - t_1 + t_2, \tag{A.7}$$

$$\lambda \mathbf{e} - (\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2) - (\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2) = 0, \tag{A.8}$$

where $\mathbf{z}_1 = \boldsymbol{\mu}_1 - \kappa_1 S_1 \mathbf{u}_1$ and $\mathbf{z}_2 = \boldsymbol{\mu}_2 + \kappa_2 S_2 \mathbf{u}_2$. Note that relations (A.4), (A.6) and (A.5), (A.7) can be written compactly as

$$(H^\top H + \theta I)\mathbf{v}_1 + \begin{pmatrix} \boldsymbol{\alpha}_1 - \boldsymbol{\beta}_1 \\ 0 \end{pmatrix} = Z\mathbf{t} \tag{A.9}$$

and

$$(G^\top G + \theta I)\mathbf{v}_2 + \begin{pmatrix} \boldsymbol{\alpha}_2 - \boldsymbol{\beta}_2 \\ 0 \end{pmatrix} = - Z\mathbf{t}, \tag{A.10}$$

respectively, where $H = [A, \mathbf{e}_1] \in \Re^{m_1 \times (n+1)}$, $G = [B, \mathbf{e}_2] \in \Re^{m_2 \times (n+1)}$, $\mathbf{v}_k = [\mathbf{w}_k^\top, b_k]^\top \in \Re^{n+1}$ for k=1,2, $\mathbf{t} = [t_1; t_2] \in \Re^2$, and $Z = [\mathbf{z}_1, - \mathbf{z}_2; 1, - 1] \in \Re^{n+1 \times 2}$. Here, the operator ',' in $[a, b]$ concatenates matrices a and b horizontally, while the operator ';' in $[a; b]$ concatenates both matrices vertically.

Since the symmetric matrices $(H^\top H + \theta I)$ and $(G^\top G + \theta I)$ are positive definite, for any $\theta > 0$, the following expressions are obtained for $\mathbf{v}_1$ and $\mathbf{v}_2$:

$$\mathbf{v}_1 = (H^\top H + \theta I)^{-1}(Z\mathbf{t} - (\widehat{\boldsymbol{\alpha}}_1 - \widehat{\boldsymbol{\beta}}_1)), \tag{A.11}$$

and

$$\mathbf{v}_2 = - (G^\top G + \theta I)^{-1}(Z\mathbf{t} + (\widehat{\boldsymbol{\alpha}}_2 - \widehat{\boldsymbol{\beta}}_2)), \tag{A.12}$$

where

$$\widehat{\boldsymbol{\alpha}}_k = \begin{pmatrix} \boldsymbol{\alpha}_k \\ 0 \end{pmatrix}, \quad \widehat{\boldsymbol{\beta}}_k = \begin{pmatrix} \boldsymbol{\beta}_k \\ 0 \end{pmatrix} \in \Re^{n+1}, \quad \text{for } k = 1, 2.$$

Using these expressions, the terms in the objective function can be rewritten in terms of $\mathbf{v}_1$ and $\mathbf{v}_2$ as follows:

$$\frac{1}{2} \|A\mathbf{w}_1 + \mathbf{e}_1 b_1\|^2 + \frac{\theta}{2}(\| \mathbf{w}_1 \|^2 + b_1^2) = \frac{1}{2}\mathbf{v}_1^\top (H^\top H + \theta I)\mathbf{v}_1, \tag{A.13}$$

and

$$\frac{1}{2} \|B\mathbf{w}_2 + \mathbf{e}_2 b_2\|^2 + \frac{\theta}{2}(\| \mathbf{w}_2 \|^2 + b_2^2) = \frac{1}{2}\mathbf{v}_2^\top (G^\top G + \theta I)\mathbf{v}_2. \tag{A.14}$$

Then, by using (A.13)–(A.14) and (A.8), the function $\widehat{L}$ (see Eq. (A.2)) can be rewritten as

$$\widehat{L} = \frac{1}{2}\mathbf{v}_1^\top (H^\top H + \theta I)\mathbf{v}_1 + \frac{1}{2}\mathbf{v}_2^\top (G^\top G + \theta I)\mathbf{v}_2 + \tilde{\mathbf{e}}^\top \mathbf{t} + \mathbf{w}_1^\top (\boldsymbol{\alpha}_1 - \boldsymbol{\beta}_1) + (\mathbf{v}_2 - \mathbf{v}_1)^\top Z\mathbf{t} + \mathbf{w}_2^\top (\boldsymbol{\alpha}_2 - \boldsymbol{\beta}_2), \tag{A.15}$$

where $\tilde{\mathbf{e}} = [1; 1] \in \Re^2$. Hence, from (A.9) and (A.10), the expression (A.15) reduces to

$$\widehat{L} = \tilde{\mathbf{e}}^\top \mathbf{t} - \frac{1}{2}\mathbf{v}_1^\top (H^\top H + \theta I)\mathbf{v}_1 - \frac{1}{2}\mathbf{v}_2^\top (G^\top G + \theta I)\mathbf{v}_2. \tag{A.16}$$

In consequence, the dual formulation for $l_2 l_\infty$–RNH–SVM can be derived by using Eqs. (A.11) and (A.12) in (A.16), as follows:

$$\max_{\substack{t_k, \mathbf{z}_k, \mathbf{u}_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k \\ k=1,2}} \quad \tilde{\mathbf{e}}^\top \mathbf{t} - \frac{1}{2}(Z\mathbf{t} - (\widehat{\boldsymbol{\alpha}}_1 - \widehat{\boldsymbol{\beta}}_1))^\top (H^\top H + \theta I)^{-1}(Z\mathbf{t} - (\widehat{\boldsymbol{\alpha}}_1 - \widehat{\boldsymbol{\beta}}_1)) \quad - \frac{1}{2}(Z\mathbf{t} + (\widehat{\boldsymbol{\alpha}}_2 - \widehat{\boldsymbol{\beta}}_2))^\top (G^\top G + \theta I)^{-1}(Z\mathbf{t} + (\widehat{\boldsymbol{\alpha}}_2 - \widehat{\boldsymbol{\beta}}_2)) \text{s.t.} \quad \mathbf{z}_1 = \boldsymbol{\mu}_1 - \kappa_1 S_1 \mathbf{u}_1,$$
$$\| \mathbf{u}_1 \| \le 1, \quad \mathbf{z}_2 = \boldsymbol{\mu}_2 + \kappa_2 S_2 \mathbf{u}_2, \ \| \mathbf{u}_2 \| \le 1, \quad t_k, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_k \ge 0, \quad k = 1, 2, \quad \boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 + \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 = \lambda \mathbf{e}. \tag{A.17}$$

References

[1] 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.
[2] F. Alizadeh, D. Goldfarb, Second-order cone programming, Math. Program. 95 (2003) 3–51.
[3] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. 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.
[4] 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.
[5] C. Bhattacharyya, Second order cone programming formulations for feature selection, J. Mach. Learn. Res. 5 (2004) 1417–1433.
[6] P. Bradley, O. Mangasarian, Feature selection via concave minimization and support vector machines, in: Machine Learning Proceedings of the Fifteenth International Conference (ICML'98), San Francisco, California, Morgan Kaufmann 1998, pp. 82–90.
[7] X. Cai, F. Nie, H. Huang, C. Ding, Multi-class $\ell_{2,1}$– norm support vector machine, in: 2011 IEEE Proceedings of the 11th International Conference on Data Mining, 2011, pp. 91–100.
[8] A. Canelas, M. Carrasco, J. López, A feasible direction algorithm for nonlinear second-order cone optimization problems. Submitted. Preprint avaliable at ⟨http://www.optimization-online.org/DB_HTML/2014/08/4511.html⟩.
[9] 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.
[10] O. Chapelle, S. Keerthi, Multi-class feature selection with support vector machines, 2008.
[11] M. Claesen, F. De Smet, J.A.K. Suykens, B. De Moora, A robust ensemble approach to learn from positive and unlabeled data using svm base models, Neurocomputing 160 (2015) 73–84.
[12] J. Demšar, Statistical comparisons of classifiers over multiple data set, J. Mach. Learn. Res. 7 (2006) 1–30.
[13] R.O. Duda, P.E. Hard, D.G. Stork, Pattern Classification, Wiley-Interscience Publication, 2001.
[14] 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. Delattre, A prognostic dna signature for t1t2 node-negative breast cancer patients, Genes, Chromosomes Cancer 49 (12) (2010) (1125–1125).
[15] J. Guo, P. Yi, R. Wang, Q. Ye, C. Zhao, Feature selection for least squares projection twin support vector machine, Neurocomputing 144 (2014) 174–183.
[16] I. Guyon, S. Gunn, M. Nikravesh, L.A. Zadeh, Feature Extraction, Foundations and Applications, Springer, Berlin, 2006.
[17] 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.
[18] K. Javed, S. Maruf, H.A. Babri, A two-stage markov blanket based feature selection algorithm for text classification, Neurocomputing 157 (2015) 91–104.
[19] Jayadeva, R. Khemchandani, S. Chandra, Twin support vector machines for pattern classification, IEEE Trans. Pattern Anal. Mach. Intell. 29 (5) (2007) 905–910.
[20] H. Kato, M. Fukushima, An sqp-type algorithm for nonlinear second-order cone programs, Optim. Lett. 1 (2) (2007) 129–144.
[21] G. Lanckriet, L.E. Ghaoui, C. Bhattacharyya, M. Jordan, A robust minimax approach to classification, J. Mach. Learn. Res. 3 (2003) 555–582.
[22] Y. Liu, F. Nie, J. Wu, L. Chen, Efficient semi-supervised feature selection with noise insensitive trace ratio criterion, Neurocomputing 105 (2013) 12–18.
[23] M. Lobo, L. Vandenberghe, S. Boyd, H. Lebret, Applications of second-order cone programming, Linear Algebra Appl. 284 (1998) 193–228.
[24] J. López, S. Maldonado, Feature selection for multiclass support vector machines using second-order cone programming, Intell. Data Anal. 19 (S1) (2015) 117–133 (Special Issue in Business Analytics).
[25] S. Maldonado, J. López, Alternative second-order cone programming formulations for support vector classification, Inf. Sci. 268 (2014) 328–341.
[26] S. Maldonado, J. López, An embedded feature selection approach for support vector classification via second-order cone programming, Intell. Data Anal. 19 (6) (2015) 1259–1273.
[27] S. Maldonado, J. Pérez, M. Labbé, R. Weber, Feature selection for support vector machines via mixed integer linear programming, Inf. Sci. 279 (2014) 163–175.
[28] O.L. Mangasarian, Nonlinear Programming, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, 1994.
[29] S. 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.
[30] F. Nie, H. Huang, X. Cai, C.H. Ding, Efficient and robust feature selection via joint $\ell_{2,1}$– norms minimization, in: J. Lafferty, C. Williams, J. Shawe-taylor, R.s. Zemel, A. Culotta (Eds.), , Advances in Neural Information Processing Systems 23 (2010), 2010, pp. 1813–1821.
[31] F. Nie, S. Xiang, Y. Jia, C. Zhang, S. Yan, Trace ratio criterion for feature selection, in: Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2 AAAI'08, AAAI Press, 2008, pp. 671–676.
[32] 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, vol. 58, 1895, pp. 240–242.
[33] 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.C. Allen, D. Zagzag, J.M. Olson, T. Curran, C. Wetmore, J.A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D.N. Louis, J.P. Mesirov, E.S. Lander, T.R. Golub, Prediction of central nervous system embryonal tumour outcome based on gene expression, Nature 415 (6870) (2002) 436–442.
[34] Y.H. Shao, W.J. Chen, N.Y. Deng, Nonparallel hyperplane support vector machine for binary classification problems, Inf. Sci. 263 (0) (2014) 22–35.
[35] Y.H. Shao, C.H. Zhang, X.B. Wang, N.Y. Deng, Improvements on twin support vector machines, IEEE Trans. Neural Netw. 22 (6) (2011) 962–968.
[36] J.F. Sturm, Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones, Optim. Methods Softw. 11 (12) (1999) 625–653 (Special issue on Interior Point Methods (CD supplement with software).
[37] G. Valentini, M. Muselli, F. Ruffino, Cancer recognition with bagged ensembles of support vector machines, Neurocomputing 56 (1) (2004) 461–466.
[38] M.M. West, C.C. Blanchette, H.H. Dressman, E.E. Huang, S.S. Ishida, R.R. Spang, H.H. Zuzan, J.A. Olson, J.R. Marks, J.R. Nevins, Predicting the clinical status of human breast cancer by using gene expression profiles, Proc. Natl. Acad. Sci. USA 98 (20) (2001) 11462–11467.
[39] J. Weston, C. Watkins, Multi-class support vector machines, in: Proceedings of the Seventh European Symposium on Artificial Neural Networks, 1999.
[40] Z.-M. Yang, J.-Y. He, Y.-H. Shao, Feature selection based on linear twin support vector machines, in: Procedia Computer Science, vol. 17, 2013, pp. 1039–1046.
[41] Q. Ye, C. Zhao, N. Ye, H. Zheng, X. Chen, A feature selection method for nonparallel plane support vector machine classification, Optim. Methods Softw. 27 (3) (2012) 431–443.
[42] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J. R. Stat. Soc., Ser. B 68 (2006) 49–67.
[43] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc., Ser. B 67 (2) (2005) 301–320.
[44] H. Zou, M. Yuan, The f-infinite norm support vector machine, Stat. Sin. 18 (2008) 379–398.

Julio López received his B.S. degree in Mathematics in 2000 from the University of Trujillo, Perú. He also received the M.S. degree in Sciences in 2003 from the University of Trujillo, Perú and the Ph.D. degree in Engineering Sciences, minor Mathematical Modelling in 2009 from the University of Chile. Currently, he is an assistant Professor of Institute of Basic Sciences at the University Diego Portales, Santiago, Chile. His research interests include conic programming, convex analysis, algorithms and machine learning.

Sebastián Maldonado received his B.S. and M.S. degree from the University of Chile, in 2007, and his Ph.D. degree from the University of Chile, in 2011. He is currently Associate Professor at the School of Engineering and Applied Sciences, Universidad de los Andes, Santiago, Chile. His research interests include statistical learning, data mining and business analytics.