Information Sciences 429 (2018) 377–389

Contents lists available at ScienceDirect

Information Sciences

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

Double regularization methods for robust feature selection and SVM classification via DC programming

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

Article history:
Received 9 March 2017
Revised 15 September 2017
Accepted 17 November 2017
Available online 21 November 2017

Keywords:
Zero norm
Support vector machines
Second-order cone programming
Dc algorithm

ABSTRACT

In this work, two novel formulations for embedded feature selection are presented. A second-order cone programming approach for Support Vector Machines is extended by adding a second regularizer to encourage feature elimination. The one- and the zero-norm penalties are used in combination with the Tikhonov regularization under a robust setting designed to correctly classify instances, up to a predefined error rate, even for the worst data distribution. The use of the zero norm leads to a nonconvex formulation, which is solved by using Difference of Convex (DC) functions, extending DC programming to second-order cones. Experiments on high-dimensional microarray datasets were performed, and the best performance was obtained with our approaches compared with well-known feature selection methods for Support Vector Machines.

1. Introduction

Support Vector Machines (SVMs) [38] is one of the best-known machine learning tools for classification. This method has proved to be very effective in terms of predictive performance thanks to its ability to decrease the model's complexity [38]. Due to its popularity, a plethora of different SVM extensions has been presented in the literature. In particular, the maximum margin approach proposed by Saketha Nath and Bhattacharyya [32] uses second-order cone programming (SOCP) formulation (see [2] for details) to provide a robust setting for specified false positive and false negative error rates. This robust framework has demonstrated its superior performance in several domains, such as binary [23] and multi-class classification [22].

SVM classifiers have the disadvantage of not being able to perform embedded feature selection, i.e. to exclude irrelevant variables from the hyperplane automatically in the classifier construction [16,25]. Feature selection can be very helpful when facing high-dimensional datasets since it reduces the risk of overfitting and mitigates the "curse of dimensionality", the mathematical issue that arises in high-dimensional settings when the amount of available data is not sufficient to obtain statistically sound results [15,25]. When the number of variables increases, the concepts of distance and proximity become ill defined. This issue is of major relevance in prediction tasks in which the number of features is much larger than the number of observations, a problem known as $p \gg N$ [18]. Under this setting, highly regularized methods are usually the best strategies to overcome overfitting and high variance. Such tasks have become of increasing importance in domains like bioinformatics and genomics.

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

https://doi.org/10.1016/j.ins.2017.11.035
0020-0255/

Several methods have been proposed in the literature to perform feature selection with SVMs [16]. In particular, penalty functions, such as the Least Absolute Shrinkage and Selection Operator (LASSO) or concave approximations of the zero norm, have been adapted to penalize the use of variables in the SVM classifier [6,28]. In this work, we have extended the ideas of double-regularized SVM classifiers to the robust SOCP setting proposed by Saketha Nath and Bhattacharyya, resulting in two novel formulations for simultaneous SVM classification and feature selection. The Difference of Convex (DC) functions Algorithm (DCA) is adapted for SOCP to find optimal solutions for our proposals. DCA was proposed by Pham Dinh Tao and Souad in [10], and has been studied extensively by Le Thi Hoai An, Pham Dinh Tao, and their collaborators, see e.g. [11,12,36,37].

This paper is structured as follows: in Section 2, previous work on SVM classification and feature selection is discussed, including all formulations that are relevant for this study. The proposed double-regularized SOCP formulations for simultaneous feature selection and classification, and the DC framework are described in Section 3. In Section 4, experimental results using high-dimensional datasets are given. Finally, the main conclusions of this work are provided in Section 5, where future developments are also proposed.

2. Literature review

In this section, the standard soft-margin SVM formulation is presented, including also its robust version based on SOCP. Subsequently, the feature selection methods that are relevant for this work are introduced. These methods are the best known feature selection approaches: Fisher Score RFE-SVM, ℓ_1-SVM, and ℓ_0-SVM; and some double-regularized extensions of the latter two methods that are relevant for our work: the ℓ_2-ℓ_1-SVM and the ℓ_2-ℓ_0-SVM methods. Finally, we present a feature selection method for the robust SOCP implementation for SVM: the ℓ_1-SOCP approach.

2.1. Soft-margin SVM

The soft-margin SVM model [8] constructs a hyperplane of the form $\mathbf{w}^\top \mathbf{x} + b = 0$, by solving the following problem:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \ \ i = 1,\ldots,m, \\ & \xi_i \geq 0, \qquad i = 1,\ldots,m, \end{aligned} \tag{1}$$

where $\|\cdot\|_2$ denotes the Euclidean norm, $\mathbf{x}_i \in \Re^n$ are the training samples and $y_i \in \{-1, 1\}$ their respective labels; $C > 0$ is a parameter that balances the trade-off between complexity reduction (minimization of the Euclidean norm) and model fit; and $\boldsymbol{\xi}$ denotes the vector of slack variables related to each training object.

2.2. Robust SVM via SOCP

The SOCP formulation proposed by Saketha Nath and Bhattacharyya [32] provides a robust scheme for classification that has proved to be very effective in terms of predictive performance [5,23]. Let $\mathbf{X}_l$ ($l = 1, 2$) be random variables that generate the samples of classes, with means and covariance matrices given by (μ_l, S_l). The main idea is to construct a robust classifier with which the probability of correct classification for each class l would be higher than $\eta_l \in (0, 1)$. The following chance-constraint problem is formulated:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 \\ \text{s.t.} \quad & \Pr\{\mathbf{w}^\top \mathbf{X}_1 + b \geq 0\} \geq \eta_1, \\ & \Pr\{\mathbf{w}^\top \mathbf{X}_2 + b \leq 0\} \geq \eta_2, \end{aligned} \tag{2}$$

Let $\mathbf{X}_l$ be represented by its mean and covariance (μ_l, S_l) for $l = 1, 2$. The robust framework replaces the chance constraints by deterministic ones in order to classify each class l correctly (at least a class recall of η_l) even for the worst case for data distribution given (μ_l, S_l). Using the multivariate Chebyshev inequality [20, Lemma 1], this worst distribution approach leads to the following quadratic SOCP problem (see [32] for the detailed derivation):

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 \\ \text{s.t.} \quad & \mathbf{w}^\top \mu_1 + b \geq 1 + \kappa_1\sqrt{\mathbf{w}^\top S_1 \mathbf{w}}, \\ & -(\mathbf{w}^\top \mu_2 + b) \geq 1 + \kappa_2\sqrt{\mathbf{w}^\top S_2 \mathbf{w}}, \end{aligned} \tag{3}$$

where $\kappa_l = \sqrt{\frac{\eta_l}{1-\eta_l}}$, for $l = 1, 2$. This problem can be formulated as a linear SOCP by introducing a new variable z and a second-order cone (SOC) constraint $\|\mathbf{w}\|_2 \leq z$. The solutions for both models are the same, but linear SOCP formulations are required by some solvers, such as the one used in this work. These linear SOCP formulations can be solved efficiently by interior point methods [2,4].

Remark 2.1. In practice, the mean and the covariance matrix are not usually available, and therefore, its respective empirical estimations are used instead.

2.3. Fisher Score for SVM

The Fisher Score is a statistical metric that can be used to rank variables according to their degree of dependency with the label vector [13]. This ranking is used further to select the top-ranked features to use in an SVM classifier. The Fisher Score follows:

$$F(j) = \left| \frac{\mu_j^+ - \mu_j^-}{(\sigma_j^+)^2 + (\sigma_j^-)^2} \right|, \tag{4}$$

where μ_j^+ and μ_j^- are the average values for variable j when taking only the positive and negative samples into account, respectively; while σ_j^+ and σ_j^- are their respective standard deviations.

2.4. Recursive Feature Elimination SVM

The RFE-SVM algorithm is a backward elimination strategy designed to remove those variables that contribute less to an SVM classifier [17]. In particular, the margin of the classifier is computed as the Euclidean norm of the weight vector $\mathbf{w}$, and then, the same measure is computed excluding one variable at a time. The feature whose removal leads to the largest margin is removed.

The RFE-SVM algorithm can be very time-consuming in high-dimensional datasets. Therefore, the authors suggested removing 50% of the attributes at each iteration of the algorithm to speed up the process [17]. Despite the higher complexity, RFE-SVM takes both the correlations between variables and their interaction with the classifier into account, being potentially more effective than filter methods like the Fisher Score.

2.5. The ℓ_1-SVM method

The soft-margin SVM method was extended to perform automatic feature selection by Bradley and Mangasarian [6]. In this work, the authors replaced the Euclidean norm as regularizer by the LASSO penalty, which corresponds to the sum of the absolute values of the weight vector, as follows:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \|\mathbf{w}\|_1 + C\sum_{i=1}^{m} \xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \;\; i = 1, \ldots, m, \\ & \xi_i \geq 0, \qquad i = 1, \ldots, m, \end{aligned} \tag{5}$$

where the expression $\|\mathbf{w}\|_1 = \sum_{i=1}^{n} |x_i|$ leads to a non-smooth optimization problem. In order to solve this issue, auxiliary decision variables $\mathbf{t}$ were introduced, with $\mathbf{t} \geq 0$ and $-\mathbf{t} \leq \mathbf{w} \leq \mathbf{t}$.

This strategy has been widely applied due to its computational efficiency and positive predictive performance. Extensions have been made to deal with several data characteristics, such as multiclass and multilabel tasks [7], grouped variables (for example, nominal attributes with multiple categories expressed through a set of dummy variables) [27], or functional data [26]. The algorithm has also been adapted for finding more effective and intuitive ways to select variables based on the amount of penalization, avoiding greedy strategies such as forward selection or backward elimination (see e.g. [27]).

2.6. The ℓ_0-SVM method

In the same work by Bradley and Mangasarian [6], the authors suggest a stronger approach to encouraging feature selection: the ℓ_0-"norm" or the cardinality of the non-zero elements of the weight vector $\mathbf{w}$. Notice that the ℓ_0 penalty is not a norm because the triangle inequality does not hold [6]. Moreover, this penalty function is discontinuous, and therefore the authors proposed the following approximation:

$$\|\mathbf{w}\|_0 \approx \sum_{j=1}^{n} (1 - \exp(-\alpha |w_j|)), \quad \alpha > 0, \tag{6}$$

where the issue of having the absolute values of the weight vector can be tackled again by the inclusion of the new variables $\mathbf{t}$, leading to the following smooth nonconvex optimization problem:

$$\begin{aligned} \min_{\mathbf{t},\mathbf{w},b,\boldsymbol{\xi}} \quad & \sum_{j=1}^{n} (1 - \exp(-\alpha t_j)) + C\mathbf{e}^\top \boldsymbol{\xi} \\ \text{s.t.} \quad & y_i(\mathbf{x}_i^\top \mathbf{w} + b) \geq 1 - \xi_i, \;\; i = 1, \ldots, m, \\ & \boldsymbol{\xi} \geq 0, \end{aligned}$$

$$-\mathbf{t} \leq \mathbf{w} \leq \mathbf{t}. \tag{7}$$

The authors refer to the previous problem as the FSV (Feature Selection ConcaVe) method, which was solved via the Successive Linearization Algorithm (SLA). The parameter α was set to 5 [6]. Alternatively, other approximations for the ℓ_0 penalty have been proposed in the SVM literature [30,40]. For instance, Weston et al. [40] proposed an approximation for this function using the logarithmic function, as follows:

$$\|\mathbf{w}\|_0 \approx \sum_{j=1}^{n} \ln(\epsilon + |w_j|), \quad 0 < \epsilon \ll 1. \tag{8}$$

2.7. Double-regularized SVM approaches

The previously described regularizers ℓ_2, ℓ_1, and ℓ_0 norms can be combined in the SVM formulation in order to balance the different objectives adequately. Neumann et al. [28] proposed the ℓ_2 - ℓ_1-SVM and ℓ_2 - ℓ_0-SVM formulations combining the Euclidean norm with the LASSO penalty and the zero-"norm", respectively. The ℓ_2 - ℓ_1-SVM formulation follows:

$$\begin{aligned} \min_{\mathbf{t},\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C_1 \sum_{j=1}^{n} t_j + C_2 \mathbf{e}^\top \boldsymbol{\xi} \\ \text{s.t.} \quad & y_i(\mathbf{x}_i^\top \mathbf{w} + b) \geq 1 - \xi_i, \ \ i = 1, \ldots, m, \\ & \boldsymbol{\xi} \geq 0, \\ & -\mathbf{t} \leq \mathbf{w} \leq \mathbf{t}. \end{aligned} \tag{9}$$

Notice that the previous formulation requires two positive hyperparameters, C_1 and C_2, in order to balance the three objectives presented in the formulation. The ℓ_2 - ℓ_0-SVM formulation has the following form:

$$\begin{aligned} \min_{\mathbf{t},\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C_1 \sum_{j=1}^{n} (1 - \exp(-\alpha t_j)) + C_2 \mathbf{e}^\top \boldsymbol{\xi} \\ \text{s.t.} \quad & y_i(\mathbf{x}_i^\top \mathbf{w} + b) \geq 1 - \xi_i, \ \ i = 1, \ldots, m, \\ & \boldsymbol{\xi} \geq 0, \\ & -\mathbf{t} \leq \mathbf{w} \leq \mathbf{t}. \end{aligned} \tag{10}$$

This formulation was solved with the DC algorithm (see [28]), which will be described in Section 3.1.

2.8. ℓ_1-SOCP

Bhattacharyya [5] extended the idea of using the LASSO penalty, instead of the Euclidean norm, on the maximum-margin SOCP formulation proposed by Saketha Nath and Bhattacharyya [32], as follows:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & ||\mathbf{w}||_1 \\ \text{s.t.} \quad & \mathbf{w}^\top \boldsymbol{\mu}_1 + b \geq 1 + \kappa_1 \|S_1^{1/2}\mathbf{w}\|_2, \\ & -\mathbf{w}^\top \boldsymbol{\mu}_2 - b \geq 1 + \kappa_2 \|S_2^{1/2}\mathbf{w}\|_2, \end{aligned} \tag{11}$$

where $S_l = S_l^{1/2} S_l^{1/2}$ for $l = 1, 2$. Formulation (11) can be cast into a linear SOCP problem by introducing a new variable $\mathbf{t}$, as follows:

$$\begin{aligned} \min_{\mathbf{w},\mathbf{t},b} \quad & \sum_{j=1}^{n} t_j \\ \text{s.t.} \quad & \mathbf{w}^\top \boldsymbol{\mu}_1 + b \geq 1 + k_1 \|S_1^{1/2}\mathbf{w}\|_2, \\ & -\mathbf{w}^\top \boldsymbol{\mu}_2 - b \geq 1 + k_2 \|S_2^{1/2}\mathbf{w}\|_2, \\ & -\mathbf{t} \leq \mathbf{w} \leq \mathbf{t}. \end{aligned} \tag{12}$$

3. Novel robust SVM approaches for feature selection

In this section, we present two novel maximum-margin approaches for binary classification and feature selection. The main idea is to balance two regularizers in the objective function of a SOCP problem: the l_2 norm for complexity reduction and structural risk minimization, and the l_1 and l_0 norms for feature elimination. The latter two regularizers are used in combination with the l_2 norm in two different formulations. The use of the l_0 norm leads to a nonconvex SOCP formulation which we propose solving via DC programming.

This section is organized as follows: a brief review of DC programming and DC algorithm (DCA) is provided first. Then, our SOCP models for feature selection and simultaneous SVM classification are presented.

3.1. Review of DC programming and DCA

The DC algorithm is a very useful mathematical programming tool used to solve nonconvex optimization problems. Let us denote by $\Gamma_0(\Re^n) := \{f : \Re^n \to \Re \cup \{+\infty\} : f$ is a lower-semicontinuous, proper and convex function$\}$, and let us consider the following optimization problem:

$$\min_{x\in\Re^n}\{f(x) = g(x) - h(x)\} \tag{P_{dc}}$$

where $g, h \in \Gamma_0(\Re^n)$. This type of problem is known as a DC program, and its objective function f is called a DC function.

Note that a DC problem with the convex constraint $x \in \Omega$ can be rewritten in the form (P_{dc}) by adding the indicator function δ_Ω of Ω into g, that is,

$$\min_{x\in\Omega}\{f(x) = g(x) - h(x)\} = \min_{x\in\Re^n}\{g(x) + \delta_\Omega(x) - h(x)\}.$$

where δ_Ω is defined by $\delta_\Omega(x) = 0$ if $x \in \Omega$, and $+\infty$ otherwise.

Since the functions that appear in problem (P_{dc}) can be nonsmooth, the concept of subdifferential is used [31]. The subdifferential of a function $\varphi \in \Gamma_0(\Re^n)$ at $x_0 \in \mathrm{dom}(\varphi)$, denoted by $\partial\varphi(x_0)$, is defined as

$$\partial\varphi(x_0) := \{y \in \Re^n : \varphi(x) \geq \varphi(x_0) + \langle y, x - x_0\rangle,\ \forall x \in \Re^n\}, \tag{13}$$

where $\langle\cdot,\ \cdot\rangle$ denotes the scalar product.

The subdifferential $\partial\varphi(x_0)$ generalizes the concept of derivative in the sense that φ is differentiable at x_0 if and only if $\partial\varphi(x_0) = \{\nabla\varphi(x_0)\}$.

One issue of DC programming is the lack of global optimality conditions. Local optimality conditions are therefore useful in DC programming. The necessary local optimality condition for DC programming (P_{dc}) is given by

$$\emptyset \neq \partial h(x^*) \subset \partial g(x^*). \tag{14}$$

Any point $x^* \in \mathrm{dom}(f)$ such that $\partial h(x^*) \cap \partial g(x^*) \neq \emptyset$ is called a critical point of (P_{dc}). The condition (14) is also sufficient (for local optimality) for many important classes of DC programs, for instance, for DC polyhedral programs, or when function f is locally convex at x^* [35,36].

Philosophy of DCA: DCA is based on local optimality conditions and duality in DC programming. The main idea of DCA is simple: each iteration k of DCA approximates the function h by its affine minorization defined by

$$h_k(x) = h(x^k) + \langle y^k, x - x^k\rangle, \quad y^k \in \partial h(x^k),$$

and then it solves the resulting minimization problem

$$\min_{x\in\Re^n}\{g(x) - h_k(x)\} = \min_{x\in\Re^n}\{g(x) - \langle y^k, x\rangle\} - h(x^k) + \langle y^k, x^k\rangle.$$

The generic DCA scheme is described in Algorithm 1:

Algorithm 1 Generic DCA scheme.

Initialization: Let $x^0 \in \Re^n$ be an initial point. Set $k = 0$.
Repeat

1. Choose some $y^k \in \partial h(x^k)$.
2. Compute
$$x^{k+1} \in \arg\min_{x\in\Re^n}\{g(x) - \langle y^k, x\rangle\}. \tag{P_k}$$
3. $k = k + 1$

Until convergence of x^k

Note that (P_k) is a convex optimization problem and hence, it is assumed to be easier to solve than the original problem.

Convergence properties of DCA and its theoretical basis can be found in [11,12,36]. For instance, it is worth mentioning that

(i) DCA is a descent method without line search but with global convergence (i.e. it converges to a critical point from any starting point).
(ii) If $g(x^{k+1}) - h(x^{k+1}) = g(x^k) - h(x^k)$, then x^k is a critical point of $g - h$. In such a case, DCA terminates at the k-th iteration.
(iii) If the optimal value of problem (P_{dc}) is finite and the infinite sequence $\{x^k\}$ is bounded, then every limit point of $\{x^k\}$ is a critical point of $g - h$.

3.2. Double-regularized SOCP formulations for feature selection

Let us consider the following chance-constrained programming problem:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\rho(\mathbf{w}) \\ \text{s.t.} \quad & \Pr\{\mathbf{w}^\top \mathbf{X}_1 + b \geq 0\} \geq \eta_1, \\ & \Pr\{\mathbf{w}^\top \mathbf{X}_2 + b \leq 0\} \geq \eta_2, \end{aligned} \tag{15}$$

where $C > 0$ denotes a weight parameter and ρ a function, which can be $\rho(\mathbf{w}) = \|\mathbf{w}\|_1$ (ℓ_1-regularized function) or $\rho(\mathbf{w}) = \|\mathbf{w}\|_0$ (ℓ_0-regularized function). The procedure described in Saketha Nath and Bhattacharyya [32] presented in Section 2.2 leads to the following deterministic problem:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\rho(\mathbf{w}) \\ \text{s.t.} \quad & \mathbf{w}^\top \boldsymbol{\mu}_1 + b \geq 1 + k_1 \|S_1^{1/2}\mathbf{w}\|_2, \\ & -\mathbf{w}^\top \boldsymbol{\mu}_2 - b \geq 1 + k_2 \|S_2^{1/2}\mathbf{w}\|_2. \end{aligned} \tag{16}$$

3.2.1. ℓ_2 - ℓ_1-SOCP formulation

Let us consider $\rho(\mathbf{w}) = \|\mathbf{w}\|_1$. Thus, formulation (16) reduces to

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\|\mathbf{w}\|_1 \\ \text{s.t.} \quad & \mathbf{w}^\top \boldsymbol{\mu}_1 + b \geq 1 + k_1 \|S_1^{1/2}\mathbf{w}\|_2, \\ & -\mathbf{w}^\top \boldsymbol{\mu}_2 - b \geq 1 + k_2 \|S_2^{1/2}\mathbf{w}\|_2, \end{aligned} \tag{17}$$

which is equivalent to the following problem

$$\begin{aligned} \min_{\mathbf{w},b,\mathbf{t}} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{j=1}^{n} t_j \\ \text{s.t.} \quad & \mathbf{w}^\top \boldsymbol{\mu}_1 + b \geq 1 + k_1 \|S_1^{1/2}\mathbf{w}\|_2, \\ & -\mathbf{w}^\top \boldsymbol{\mu}_2 - b \geq 1 + k_2 \|S_2^{1/2}\mathbf{w}\|_2, \\ & -\mathbf{t} \leq \mathbf{w} \leq \mathbf{t}. \end{aligned} \tag{18}$$

This problem is an instance of convex quadratic SOCP problem, which contains two SOC constraints and $2n$ linear inequality ones.

We note that Problem (18) can be written as a linear SOCP by introducing a new variable z, with an objective function $z + C\sum_{j=1}^{n} t_j$, and an additional SOC constraint $\|(z-1, \sqrt{2}\mathbf{w})\|_2 \leq z+1$, (see [2] for more details). Thus, it can be solved efficiently by a standard SOCP solver such as SeDuMi [34].

3.2.2. ℓ_2 - ℓ_0-SOCP formulation

Let us consider $\rho(\mathbf{w}) = \|\mathbf{w}\|_0$. By using the approximation (6) and introducing a new variable $\mathbf{t}$, the formulation (16) is reduced to the following problem:

$$\begin{aligned} \min_{\mathbf{t},\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{j=1}^{n}(1 - \exp(-\alpha t_j)) \\ \text{s.t.} \quad & \mathbf{w}^\top \boldsymbol{\mu}_1 + b \geq 1 + k_1 \|S_1^{1/2}\mathbf{w}\|_2, \\ & -\mathbf{w}^\top \boldsymbol{\mu}_2 - b \geq 1 + k_2 \|S_2^{1/2}\mathbf{w}\|_2, \\ & -\mathbf{t} \leq \mathbf{w} \leq \mathbf{t}. \end{aligned} \tag{19}$$

This formulation has a nonconvex objective function, two SOC constraints and $2n$ lineal inequality ones. Let us denote by $\mathbf{u} = (\mathbf{t}, \mathbf{w}, b) \in \Re^{2n+1}$ and

$$\begin{aligned} \Omega = \quad & \{\mathbf{u} \in \Re^{2n+1} : \mathbf{w}^\top \boldsymbol{\mu}_1 + b \geq 1 + k_1 \|S_1^{1/2}\mathbf{w}\|_2, -\mathbf{w}^\top \boldsymbol{\mu}_2 - b \geq 1 + k_2 \|S_2^{1/2}\mathbf{w}\|_2, \\ & -\mathbf{t} \leq \mathbf{w} \leq \mathbf{t}\}. \end{aligned}$$

Then, the problem (19) can be rewritten as

$$\min_{\mathbf{t},\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{j=1}^{n}(1-\exp(-\alpha t_j)) + \delta_\Omega(\mathbf{u}). \tag{20}$$

Let us define

$$g(\mathbf{u}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \delta_\Omega(\mathbf{u}), \quad h(\mathbf{u}) = C\sum_{i=1}^{n}(\exp(-\alpha t_i) - 1).$$

Given that the set Ω is convex, the functions g and h are also convex. Thus, Formulation (19) has the form of Problem (P_{dc}). Hence, we can use the DCA scheme (Algorithm 1) for solving this formulation.

3.3. DCA for solving the ℓ_2-ℓ_0-SOCP formulation

Note that h is a differentiable function. Then, $\nabla h(\mathbf{u}) = (\mathbf{v}, \mathbf{0}, 0)$ with

$$v_i = -C\alpha\exp(-\alpha t_i), \quad i = 1,\ldots,n. \tag{21}$$

In consequence, (P_k) takes the following form:

$$\begin{array}{ll} \min_{\mathbf{t},\mathbf{w},b} & \frac{1}{2}\|\mathbf{w}\|_2^2 - \langle \mathbf{v}^k, \mathbf{t}\rangle \\ \text{s.t.} & \mathbf{w}^\top\boldsymbol{\mu}_1 + b \geq 1 + k_1\|S_1^{1/2}\mathbf{w}\|_2, \\ & -\mathbf{w}^\top\boldsymbol{\mu}_2 - b \geq 1 + k_2\|S_2^{1/2}\mathbf{w}\|_2, \\ & -\mathbf{t} \leq \mathbf{w} \leq \mathbf{t}. \end{array} \tag{22}$$

This problem is an instance of quadratic SOCP problem.

Summarizing, the DCA scheme for solving (19) can be described as follows:

Algorithm 2 DCA scheme applied to (19).

Initialization: Let $\varepsilon > 0$ be a tolerance sufficiently small and $(\mathbf{t}^0, \mathbf{w}^0, b^0) \in \Re^{2n+1}$ a initial point. Set $k = 0$.
Repeat

1. Compute $\mathbf{v}^k$ via (21).
2. Solve the quadratic SOCP problem (22)to obtain $\mathbf{u}^{k+1} = (\mathbf{t}^{k+1}, \mathbf{w}^{k+1}, b^{k+1})$.
3. $k = k + 1$

Until $\|\mathbf{u}^{\mathbf{k+1}} - \mathbf{u}^k\| < \varepsilon$.

Taking into account [12, Lemma 3.6, Theorem 3.7], and the fact that the objective function of problem (20) is bounded from below, we obtain the following theorem:

Theorem 3.1.

(i) *Algorithm 2 generates a sequence* $\{\mathbf{u}^k = (\mathbf{t}^k, \mathbf{w}^k, b^k)\}$ *contained in* Ω *such that the sequence* $\{g(\mathbf{u}^k) - h(\mathbf{u}^k)\}$ *is monotone decreasing.*

(ii) *If the sequence* $\{\mathbf{u}^k\}$ *is bounded, then each limit point* u^* *satisfies the necessary local optimality condition (14).*

We note that the Theorem above only guarantees the local optimality condition, and therefore, multiple starting points in the feasible set should be taken into account to attain a global optimum.

4. Results and discussion

The proposed robust feature selection methods ℓ_2-ℓ_1-SOCP and ℓ_2-ℓ_0-SOCP were applied to the following high-dimensional microarray datasets:

- Alon's colon cancer data (ALON) [3]: This dataset contains 2,000 features (genes) that describe 62 sample tissues.
- Alizadeh's lymphoma data (ALIZADEH) [1]: This dataset contains the expression of 96 samples described by 4,026 variables.
- Gravier's breast cancer data (GRAVIER) [14]: This dataset contains the gene expression of 168 samples described by 2,905 features.
- Pomeroy's central nervous system embryonal tumor data (POMEROY) [29]: This dataset contains 7,128 features and 60 samples.
- Shipp's lymphoma data (SHIPP) [33]: This dataset contains the expression of 77 samples described by 7,129 variables.
- West's breast cancer data (WEST) [39]: This dataset contains the gene expression of 49 samples described by 7,129 features.

Table 1
Average AUC over all subsets of selected attributes. All datasets.

	ALON	ALIZADEH	GRAVIER	POMEROY	SHIPP	WEST
Fisher+SVM	86.1	93	71.8	62.7	92	71.9
RFE-SVM	87	93.3	67.5	59.8	91	68.6
ℓ_1-SVM	88.2	92.5	73.3	71.2	97.1	67.1
ℓ_2 - ℓ_1-SVM	88	93.9	72.7	63.8	96.8	75.6
ℓ_2 - ℓ_0-SVM	88.2	95.3	73.9	72.7	96.1	79
ℓ_1-SOCP	86.9	92.9	**78.3**	57.9	96.9	81.6
ℓ_2 - ℓ_1-SOCP	86.9	96	77.8	67.7	97.4	**89.8**
ℓ_2 - ℓ_0-SOCP	**92.3**	**96.2**	78.0	**73.8**	**98.5**	89.4

Table 2
Maximum AUC for the best subset of selected attributes, and its respective cardinality. All datasets.

	ALON		ALIZADEH		GRAVIER		POMEROY		SHIPP		WEST	
	AUC	n^*	AUC	n^*	AUC	n^*	AUC	n^*	AUC	n^*	AUC	n^*
Fisher+SVM	86.9	20	95.6	1000	75.7	1000	72	50	96.5	1000	89.8	20
RFE-SVM	89.4	20	94.8	250	74	1000	65.9	50	93	20	77.4	20
ℓ_1-SVM	89.4	1000	92.6	1000	74.4	1000	72.2	1000	97.4	50	67.4	1000
ℓ_2 - ℓ_1-SVM	88.2	50	94.1	250	74.8	100	66.1	100	96.6	50	75.3	250
ℓ_2 - ℓ_0-SVM	88.2	100	95.6	50	76.2	50	74.7	500	97.4	20	83.7	250
ℓ_1-SOCP	86.9	20	92.6	20	78.3	50	58.8	20	97.4	20	81.6	20
ℓ_2 - ℓ_1-SOCP	86.9	20	97.1	100	**79.7**	50	69.6	50	97.4	20	89.8	20
ℓ_2 - ℓ_0-SOCP	**93.0**	20	**98.5**	100	79.6	50	**75.8**	500	**100**	500	**91.8**	100

Table 3
Holm's post-hoc test for pairwise comparisons.

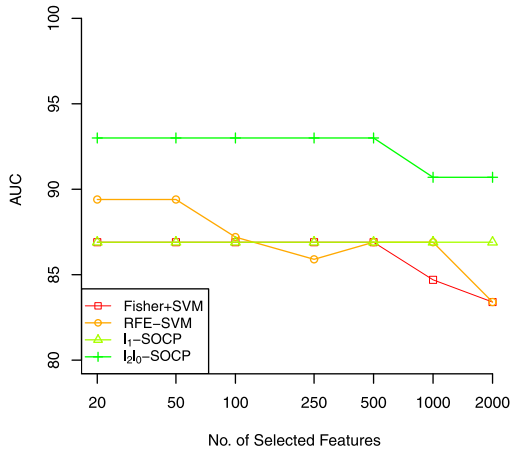
Method	Mean Rank	Mean AUC	p value	$\alpha/(k-i)$	Action
Analysis for ℓ_2 - ℓ_1-SOCP					
ℓ_2 - ℓ_0-SVM	2.58	85.97	-	-	-
ℓ_2 - ℓ_1-SOCP	2.67	86.75	0.95	0.05	not reject
Fisher+SVM	3.83	86.08	0.32	0.025	not reject
ℓ_1-SVM	4.25	82.23	0.18	0.0167	not reject
ℓ_1-SOCP	4.67	82.60	0.09	0.0125	not reject
ℓ_2 - ℓ_1-SVM	4.92	82.52	0.06	0.01	not reject
RFE-SVM	5.08	82.42	0.045	0.0083	not reject
Analysis for ℓ_2 - ℓ_0-SOCP					
ℓ_2 - ℓ_0-SOCP	1.00	89.78	-	-	-
ℓ_2 - ℓ_0-SVM	3.00	85.97	0.108	0.05	not reject
Fisher+SVM	4.17	86.08	0.011	0.025	reject
ℓ_1-SVM	4.67	82.23	0.003	0.0167	reject
ℓ_1-SOCP	4.84	82.60	0.002	0.0125	reject
ℓ_2 - ℓ_1-SVM	5.08	82.52	0.001	0.01	reject
RFE-SVM	5.25	82.42	0.0007	0.0083	reject

The model selection step was performed for all methods using a nested cross-validation (CV) strategy (also referred to as repeated double CV, see e.g. [19,22]): training and test subsets are obtained using leave-one-out cross-validation (LOOCV) for the outer loop, and the training subset is further split into training and validation subsets in order to use 10-fold CV to find the right hyperparameter setting (inner loop).

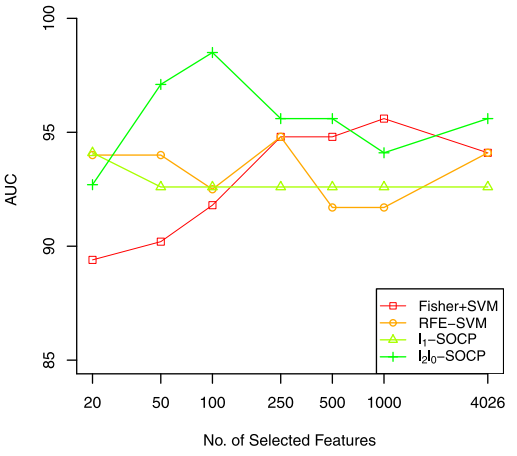
The following values were explored for the various hyperparameters: C_1, C_2, $C \in \{2^{-7}, 2^{-6}, \ldots, 2^6, 2^7\}$ for the methods soft-margin SVM (Formulation (1), base classifier for the approaches Fisher+SVM and RFE-SVM), ℓ_1-SVM, ℓ_2 - ℓ_1-SVM, ℓ_2 - ℓ_0-SVM, and the proposed ℓ_2 - ℓ_1-SOCP and ℓ_2 - ℓ_0-SOCP methods; and parameters η_1, $\eta_2 \in \{0.2, 0.4, 0.6, 0.8\}$ for the robust methods ℓ_1-SOCP, ℓ_2 - ℓ_1-SOCP, and ℓ_2 - ℓ_0-SOCP, as suggested in [5,24]. The Area Under the Curve (AUC) was used as the performance metric for model selection. After this step, a ranking was constructed for all feature selection methods, studying the performance for various subsets of selected attributes of cardinality $n = \{20, 50, 100, 250, 500, 1000\}$. For the embedded methods (all approaches except the Fisher Score and RFE-SVM), the ranking was constructed based on the absolute values of the weight vector.

We used the SeDuMi solver (see [34]) for solving the convex SOCP problems appearing in formulations ℓ_1-SOCP, ℓ_2 - ℓ_1-SOCP and ℓ_2 - ℓ_0-SOCP.

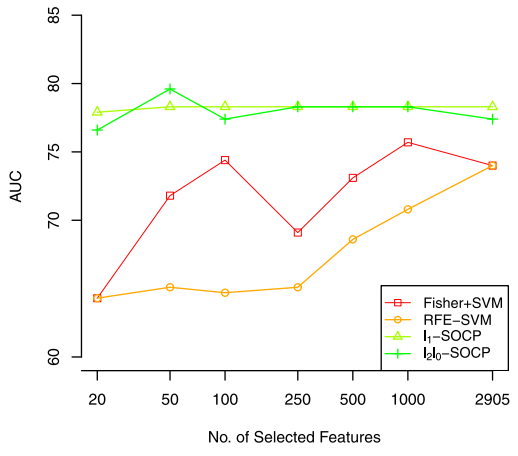
To report a summary of all results, the average and maximum performance across all subsets of cardinality n were computed. Tables 1 and 2 present the average and maximum AUC, respectively, for all feature selection methods and data



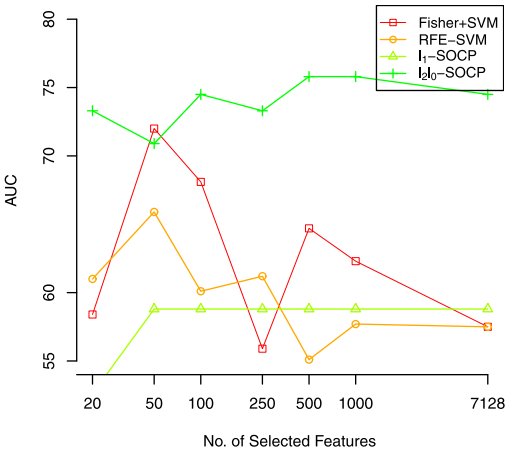
(a) ALON dataset



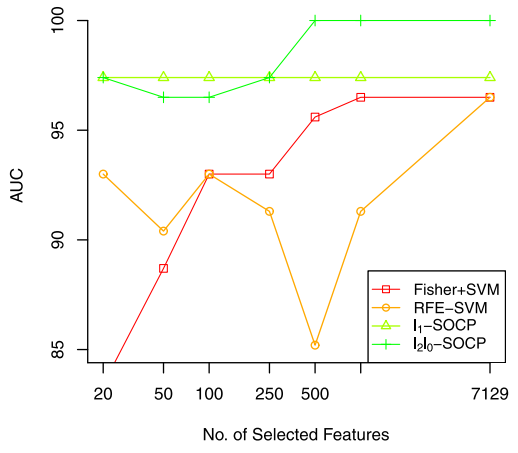
(b) ALIZADEH dataset



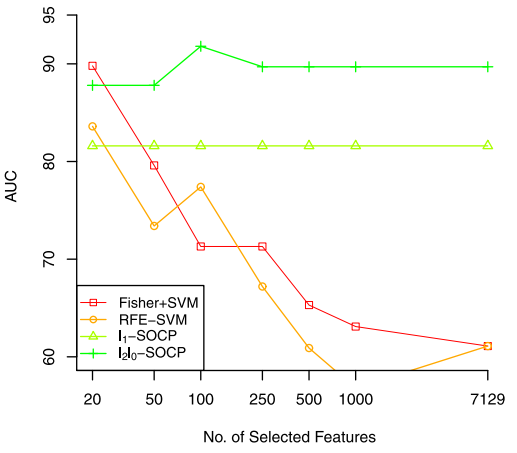
(c) GRAVIER dataset



(d) POMEROY dataset

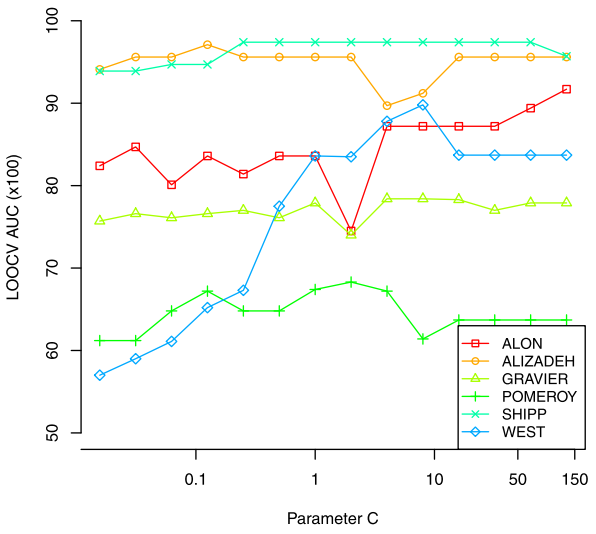


(e) SHIPP dataset

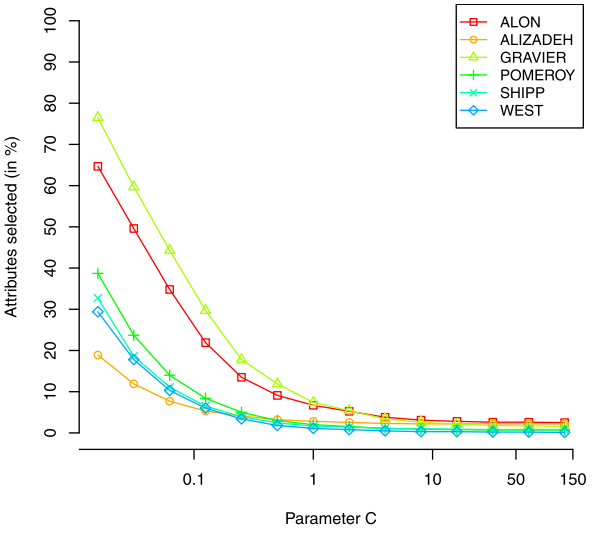


(f) WEST dataset

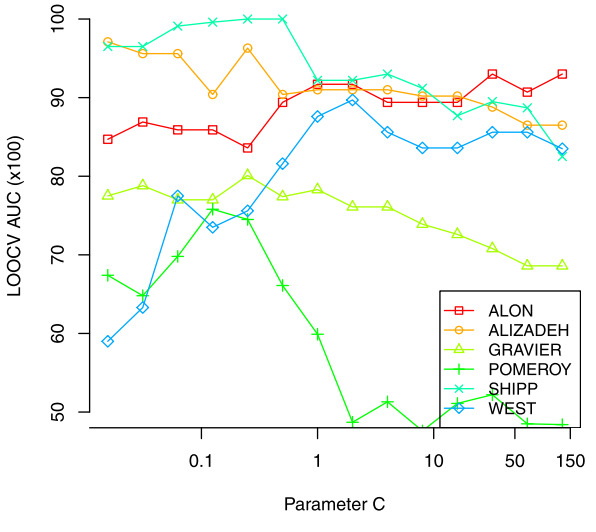
Fig. 1. Performance for an increasing number of features. All datasets.



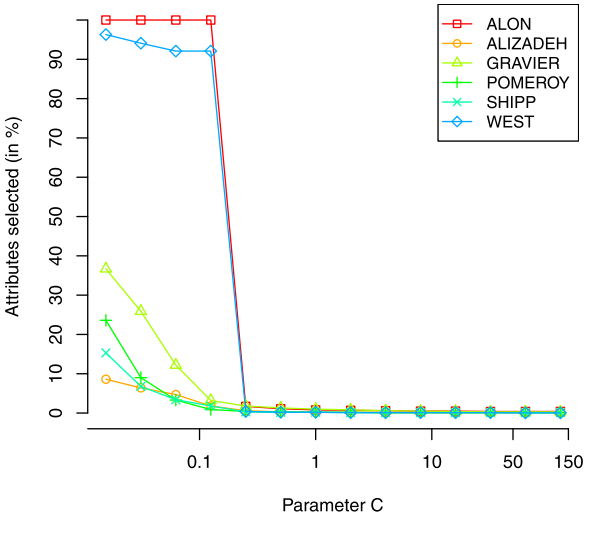
(a) ℓ_2-ℓ_1-SOCP, AUC



(b) ℓ_2-ℓ_1-SOCP, Percentage of attributes selected



(c) ℓ_2-ℓ_0-SOCP, AUC



(d) ℓ_2-ℓ_0-SOCP, Percentage of attributes selected

Fig. 2. Sensitivity analysis for the proposed approaches. Influence of parameter C. All datasets.

sets. For the latter table, the cardinality of the subset of selected variables that leads to the highest AUC (n^*) is also reported. The lowest n^* is presented in case of ties in the highest AUC. For both tables, the highest AUC is highlighted in bold type.

The following results are visualized on Tables 1 and 2:

- Our proposals ℓ_2-ℓ_1-SOCP and ℓ_2-ℓ_0-SOCP achieved the best average performance in five of six datasets (Table 1), and the highest AUC for all six datasets (Table 2). These results confirm the methodological virtues of the methods, where the robustness conferred by the SOCP strategy provides stability and effectiveness, and the double regularization approach finds the best compromise between feature elimination, complexity reduction, and model fit.
- A comparison between double-regularized approaches favors the combination of the ℓ_2 norm with the ℓ_0 regularizer over the ℓ_1 norm: The methods ℓ_2-ℓ_0-SOCP and ℓ_2-ℓ_0-SVM achieve better performance in terms of both average and maximum performance compared with ℓ_2-ℓ_1-SVM and ℓ_2-ℓ_1-SVM, respectively. This fact confirms our assumptions in the sense that the ℓ_0 "norm" is designed exclusively for feature selection, and therefore makes a better combination with

the Tikhonov regularization than the LASSO penalty, which provides a good balance between complexity reduction and sparsity, and provides good results as a single regularizer.
- The maximum performance is usually achieved with 100 attributes or less (see Table 2), demonstrating the importance of feature selection in order to achieve best performance in high-dimensional tasks.

Next, we used the Holm's test [9] to analyze the overall performance statistically. For each feature selection technique, Table 3 presents the average rank based on the maximum AUC reported on Table 2, and the p value for the pairwise test that compares it with the one with the best overall performance. The ℓ_2 - ℓ_1-SOCP and ℓ_2 - ℓ_0-SOCP methods are studied separately.

It can be concluded from Table 3 that the proposed ℓ_2 - ℓ_0-SOCP achieves the best overall performance with an average rank of 1, statistically outperforming all alternative methods but ℓ_2 - ℓ_0-SVM. The ℓ_2 - ℓ_1-SOCP method, however, is not able to outperform the alternative approaches, achieving a slightly below average ranking compared to ℓ_2 - ℓ_0-SVM.

Next, a detailed feature selection performance analysis is presented in Fig. 1, in which the AUC is reported for all the subsets of n variables studied and, all the datasets. Only the methods Fisher Score with SVM, RFE-SVM, ℓ_1-SOCP, and ℓ_2 - ℓ_0-SOCP are presented for visualization purposes. The importance of these four methods rests on the following reasons: Fisher Score and RFE-SVM are the best-known benchmark approaches for feature selection with SVM; ℓ_1-SOCP is the most natural benchmark from the literature given the nature of our work; and finally, ℓ_2 - ℓ_0-SOCP represents our main proposal.

In Fig. 1, it can be observed that our proposed ℓ_2 - ℓ_0-SOCP achieves consistently higher AUC than the other benchmark methods. The Fisher Score and RFE-SVM methods have rather unstable behavior, showing a very large gap between the minimum and the maximum AUC for the different values for n. Finally, although it can be noticed that the gain in terms of performance of using feature selection versus using all variables is not always large, noteworthy gains can be observed in the POMEROY and WEST datasets.

Finally, a sensitivity analysis was performed to analyze the influence of parameter C on the final solution of the proposed approaches ℓ_2 - ℓ_1-SOCP and ℓ_2 - ℓ_0-SOCP in terms of both AUC and feature selection. Feature selection performance is assessed by defining a threshold of 0.00000001 for the absolute values of the weights, and then the percentage of weights that are above this threshold is computed as a measure of parsimony. This analysis is illustrated in Fig. 2.

As expected, a larger penalization given by high values of C leads to aggressive feature selection (see Fig. 2(b) and Fig. 2(d) for ℓ_2 - ℓ_1-SOCP and ℓ_2 - ℓ_0-SOCP, respectively), but it has a rather minor impact on performance (see Fig. 2(a) and Fig. 2(c) for ℓ_2 - ℓ_1-SOCP and ℓ_2 - ℓ_0-SOCP, respectively). Although the AUC remains relatively stable for most datasets, the optimal performance is usually achieved when less than 10% of variables has weights above the threshold, demonstrating the importance of feature selection in high-dimensional settings, such as microarray analysis.

5. Conclusions and future developments

In this work, a robust framework for binary classification and embedded feature selection was developed. Maximum margin classifiers were constructed via second-order cone programming, while the inclusion of a second regularizer based on the LASSO and the zero-norm penalty functions conferred sparsity to the final solution. The use of a concave approximation for the zero-norm regularization leads to a nonconvex SOCP problem, which is solved via a variation of the DC algorithm. A comparison with other SVM-based feature selection approaches using high-dimensional datasets showed the advantages of the proposed methods:

- The proposed ℓ_2 - ℓ_0-SOCP method showed superior performance in terms of both average and maximum performance among various subsets of features, being able to outperform most alternative methods statistically. The proposed ℓ_2 - ℓ_1-SOCP method also reached a very good predictive performance.
- The robust framework has the ability to generalize better by assuming a pessimistic data distribution, while the double-regularized feature selection approach allows an adequate balance between sparsity, model fit, and generalization.
- Our extends the work of Neumann et al. [28] on double-regularized feature selection and SVM classification to SOCP, proposing a robust approach that has proven to be as effective in high-dimensional datasets [5,24].

In our experiments, we observed slightly better performance of the proposed ℓ_2 - ℓ_0-SOCP over the ℓ_2 - ℓ_1-SOCP method, meaning that the Tikhonov regularization seems to work better in company with the zero norm than the 1-norm. Despite the fact that the inclusion of a zero norm approximation increases complexity due to nonconvexity, it is worth making the extra computational effort in order to achieve the best predictive performance. The same phenomenon is observed for standard SVM. A reason that explains this result is that the ℓ_0 focuses merely on feature elimination, and requires an additional regularizer to minimize the structural risk, while the LASSO penalty is a good compromise between complexity reduction and sparsity by itself, and can be used as the sole regularizer without the need of the Euclidean norm.

There are several opportunities for future developments. First, there are several alternatives to zero norm approximations that can be used as alternatives for Equation (6). For instance, SOCP models can be derived from the l_0-SVM method proposed by Weston et al. [40], or the ℓ_0 approximations discussed in Rinaldi et al. [30]. A second line for future research encompasses the development of optimization approaches to solve nonconvex SOCP formulations, which can be used as an alternative to DC programming. Finally, our proposal can be extended to other machine learning tasks that face high-dimensional problems, such as multiclass classification or regression. In this context, wrapper methods have been devel-

oped for maximum margin SOCP formulations for multiclass learning with positive results [21], motivating the deployment of embedded methods that may be more accurate and more efficient computationally.

Acknowledgements

The first author was supported by FONDECYT project 1160894, the second was funded by FONDECYT project 1160738, and third author was supported by FONDECYT project 1130905. This work was partially funded by Complex Engineering Systems Institute (CONICYT, PIA, FB0816). The authors are grateful to the anonymous referees for their careful reading and helpful suggestions that improved the paper greatly.

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, Mathematical Programming 95 (2003) 3–51.
[3] 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, Proceedings of the National Academy of Sciences 96 (12) (1999) 6745–6750.
[4] F. Alvarez, J. López, H. Ramírez C., Interior proximal algorithm with variable metric for second-order cone programming: applications to structural optimization and support vector machines, Optimization Methods Software 25 (6) (2010) 859–881.
[5] C. Bhattacharyya, Second order cone programming formulations for feature selection, Journal of Machine Learning Research 5 (2004) 1417–1433.
[6] P. Bradley, O. Mangasarian, Feature selection va 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.
[7] M. Carrasco, J. López, S. Maldonado, A multi-class svm approach based on the l1-norm minimization of the distances between the reduced convex hulls, Pattern Recognition 48 (5) (2015) 1598–1607.
[8] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (1995) 273–297.
[9] J. Demšar, Statistical comparisons of classifiers over multiple data set, Journal of Machine Learning Research 7 (2006) 1–30.
[10] T.P. Dinh, E. Souad, Algorithms for solving a class of nonconvex optimization problems. methods of subgradients, in: J.-B. Hiriart-Urruty (Ed.), Fermat Days 85: Mathematics for Optimization, North-Holland Mathematics Studies, 129, North-Holland, 1986, pp. 249–271.
[11] T.P. Dinh, H.L. Thi, Convex analysis approaches to dc programming: Theory, algorithms and applications, Acta Mathematica Vietnamica 22 (1) (1997) 287–367.
[12] T.P. Dinh, H.L. Thi, A d.c. optimization algorithm for solving the trust-region subproblem, SIAM Journal on Optimization 8 (2) (1998) 476–505.
[13] R. Duda, P. Hard, D. 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.and Delattre, A prognostic dna signature for t1t2 node-negative breast cancer patients, Genes, Chromosomes and Cancer 49 (12) (2010) 1125.
[15] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning research 3 (2003) 1157–1182.
[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., Machine Learning 46 (1-3) (2002) 389–422.
[18] T. Hastie, R. Tibshirani, J. Friedman, Elements of Statistical Learning, Springer, 2009.
[19] D. Krstajic, L. Buturovic, D. Leahy, S. Thomas, Cross-validation pitfalls when selecting and assessing regression and classification models, Journal of Cheminformatics 6 (10) (2014) 1–15.
[20] G. Lanckriet, L. Ghaoui, C. Bhattacharyya, M. Jordan, A robust minimax approach to classification, Journal of Machine Learning Research 3 (2003) 555–582.
[21] J. López, S. Maldonado, Robust feature selection for multi-class second-order cone programming support vector machines, Intelligent Data Analysis. 19 (S1) (2015) S117–S133.
[22] J. López, S. Maldonado, Multi-class second-order cone programming support vector machines, Information Sciences 330 (2016) 328–341.
[23] S. Maldonado, J. López, Alternative second-order cone programming formulations for support vector classification, Information Sciences 268 (2014) 328–341.
[24] S. Maldonado, J. López, An embedded feature selection approach for support vector classification via second-order cone programming, Intelligent Data Analysis 19 (6) (2015) 1259–1273.
[25] S. Maldonado, R. Weber, J. Basak, Kernel-penalized SVM for feature selection, Information Sciences 181 (1) (2011) 115–128.
[26] B. Martin-Barragan, R. Lillo, J. Romo, Interpretable support vector machines for functional data, European Journal of Operational Research 232 (1) (2014) 146–155.
[27] L. Meier, S. Van De Geer, P. Buehlmann, The group lasso for logistic regression, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70 (1) (2008) 53–71.
[28] J. Neumann, C. Schnorr, G. Steidl, Combined svm-based feature selection and classification, Machine Learning 61 (1-3) (2005) 129–150.
[29] 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.
[30] F. Rinaldi, F. Schoen, M. Sciandrone, Concave programming for minimizing the zero-norm over polyhedral sets, Computational Optimization and Applications 46 (3) (2010) 467–486.
[31] R. Rockafellar, Convex analysis, Princeton Mathematical Series, No. 28, Princeton University Press, Princeton, N.J., 1970.
[32] 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.
[33] 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, Nature Medicine 8 (1) (2002) 68–74.
[34] J. Sturm, Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones, Optimization Methods and Software 11 (12) (1999) 625–653. Special issue on Interior Point Methods (CD supplement with software).
[35] H.L. Thi, T.P. Dinh, Solving a class of linearly constrained indefinite quadratic problems by dc algorithms, Journal of Global Optimization 11 (3) (1997) 253–285.
[36] H.L. Thi, T.P. Dinh, The dc (difference of convex functions) programming and dca revisited with dc models of real world nonconvex optimization problems, Annals of Operations Research 133 (1) (2005) 23–46.

[37] H.L. Thi, T.P. Dinh, L. Muu, Numerical solution for optimization over the efficient set by d.c. optimization algorithms, Operations Research Letters 19 (3) (1996) 117–128.
[38] V. Vapnik, Statistical Learning Theory, John Wiley and Sons, 1998.
[39] 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, Proceedings of the National Academy of Sciences of the United States of America 98 (20) (2001) 11462–11467.
[40] J. Weston, A. Elisseeff, B. Schlkopf, M. Tipping, The use of zero-norm with linear models and kernel methods, Journal of Machine Learning Research 3 (2003) 1439–1461.