HAL
archives-ouvertes.fr

FEATURE SELECTION IN SVM VIA POLYHEDRAL K-NORM

M Gaudioso, E Gorgone, J.-B Hiriart-Urruty

▶ **To cite this version:**

M Gaudioso, E Gorgone, J.-B Hiriart-Urruty. FEATURE SELECTION IN SVM VIA POLYHEDRAL K-NORM. 2018. hal-01977766

HAL Id: hal-01977766
https://hal.archives-ouvertes.fr/hal-01977766

Preprint submitted on 11 Jan 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

FEATURE SELECTION IN SVM VIA POLYHEDRAL K-NORM

M. GAUDIOSO*, E. GORGONE †, AND J.-B. HIRIART–URRUTY ‡

Abstract. We treat the Feature Selection problem in the Support Vector Machine (SVM) framework by adopting an optimization model based on use of the ℓ_0 pseudo–norm. The objective is to control the number of non-zero components of normal vector to the separating hyperplane, while maintaining satisfactory classification accuracy. In our model the polyhedral norm $\|.\|_{[k]}$, intermediate between $\|.\|_1$ and $\|.\|_\infty$, plays a significant role, allowing us to come out with a DC (Difference of Convex) optimization problem that is tackled by means of DCA algorithm.

The results of several numerical experiments on benchmark classification datasets are reported.

Keywords. Sparse optimization, Cardinality constraint, k-norm, Support Vector Machine, DC optimization.

1. Introduction. A relevant problem in binary classification is to design good quality classifiers by resorting to a minimal number of sample parameters. One of the main motivations is to gather a more clear interpretation of phenomena underlying the class membership distribution of the samples. In the more general setting of Machine Learning, such problem falls in the area of Feature Selection (FS), which has been the object of intensive research in recent decades (see, e.g., the survey [10]).

We focus, in particular, on the Support Vector Machine (SVM) framework [24], [4], where binary classification is pursued by finding an "optimal" two–class separating hyperplane, either in the original parameter (or "feature") space or upon appropriate kernel transformation.

Numerical optimization algorithms play a relevant role in SVM area and, more specifically, in FS. The problem is to guarantee a reasonable trade–off between classification accuracy and the number of features actually used. Controlling the latter consists basically in minimizing the number of non–zero components of the normal vector to the separating hyperplane.

The literature offers several contributions. In [1] NP-hardness of the problem has been assessed.

In [3] the model adopted is based on considering the step function for each component of the normal vector; discontinuity is handled by introducing two different approximations, the standard sigmoid and a concave exponential, respectively. In particular, by adopting the concave approximation, FS problem is tackled by solving a finite sequence of linear programs. Different approximations are given in [26] and in [21], where, in particular, the concave and separable objective functions, derived by the approximations, are handled via variants of Frank-Wolfe method.

It is interesting to note that in [16] an approximation scheme of the step function is cast into a DC (Difference of Convex) framework, providing thus the opportunity of resorting to the algorithmic machinery for dealing with such class of nonconvex problems. An early survey on properties and relevance of such class of functions is in [12] (see also [22]).

Parallel to treatment of FS via approximation methods, Mixed Integer Programming formulations have been successfully adopted. The idea is to introduce binary variables, one for each component of the normal vector to the supporting hyperplane, that are switched to "1" if and only if the corresponding component is non-zero. From among the proposed approaches, we recall here [17], [2], [7].

In a more general setting, FS falls into the wide area of sparse optimization, where one is faced to

*Dipartimento di Ingegneria Informatica, Modellistica, Elettronica e Sistemistica, Università della Calabria, 87036 Rende (CS), Italia. E-mail: gaudioso@dimes.unical.it

†Dipartimento di Matematica e Informatica, Università di Cagliari, 09124 Cagliari, Italia. E-mail: egorgone@unica.it

‡Institut de mathématiques, Université Paul Sabatier, 31062 Toulouse, France. E-mail: jbhu@math.univ-toulouse.fr

the (regularized) problem:

$$\min_{x\in\mathbb{R}^n} f(x) + \|x\|_0 \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and $\|.\|_0$ is the ℓ_0 pseudo–norm, which counts the number of non-zero component of any vector. Sometimes sparsity of the solution, instead of acting on the objective function, is enforced by introducing a constraint on the ℓ_0 pseudo–norm of the solution, thus defining a cardinality–constraint problem [20].

In many applications, the ℓ_0 pseudo–norm in (1.1) is replaced by the ℓ_1–norm, which is definitely more tractable from the computational point of view, yet ensuring sparsity, to a certain extent (see [27] for a discussion on a general regularization scheme).

In the seminal paper [25], a class of polyhedral norms (the k-norms), intermediate between $\|.\|_1$ and $\|.\|_\infty$, is introduced to obtain sparse approximation solutions to systems of linear equations. The use of other norms to recover sparsity is described in [6]. In more recent years the use of k-norms has received much attention and has led to several proposals for dealing with ℓ_0 pseudo–norm cardinality constrained problem [11], [28], [23], [9].

In this paper we cast the classic SVM approach into the sparse optimization framework (1.1). Our work is inspired by [9], the main difference being in the explicit (and not parametric) minimization of the ℓ_0 pseudo–norm. We formulate our SVM–ℓ_0 pseudo–norm problem (SVM_0, for short) and we tackle it by means of a penalization approach which allows us to put the problem in DC form. The algorithm adopted is of DCA type [15].

The paper is organized as follows. In Section 2 we recall first the standard SVM model and introduce the FS problem as a sparse optimization one. Then we discuss the use of the polyhedral k-norm in sparse optimization, coming out with a DC formulation. On such basis we formulate in Section 3 the SVM_0 problem. The results of several numerical experiments on benchmark datasets are in Section 4. Some conclusions are finally drawn in Section 5.

2. Feature Selection and ℓ_0 pseudo–norm minimization. The binary classification problem in the SVM setting is usually put in the following form. Given two point-sets $\mathcal{A} \stackrel{\triangle}{=} \{a_1, \ldots, a_{m_1}\}$ and $\mathcal{B} \stackrel{\triangle}{=} \{b_1, \ldots, b_{m_2}\}$ in $\mathbb{R}^n$, we look for *linear separation* of the two sets, that is for a hyperplane $\{x | x \in \mathbb{R}^n,\, x^\top w' = \gamma'\}$, $(w' \in \mathbb{R}^n, \gamma' \in \mathbb{R})$, strictly separating $\mathcal{A}$ and $\mathcal{B}$, thus ensuring

$$a_i^\top w' < \gamma', \quad i = 1, \ldots, m_1 \quad \textit{and} \quad b_l^\top w' > \gamma', \quad l = 1, \ldots, m_2.$$

It is easy to verify that such a hyperplane exists if and only if there exists a hyperplane $\{x | x \in \mathbb{R}^n,\, x^\top w = \gamma\}$, $(w \in \mathbb{R}^n, \gamma \in \mathbb{R})$, such that

$$a_i^\top w \leq \gamma - 1, \quad i = 1, \ldots, m_1 \quad \textit{and} \quad b_l^\top w \geq \gamma + 1, \quad l = 1, \ldots, m_2.$$

Gordan's theorem of the alternative [18] guarantees linear separation if and only if $\mathrm{conv}\mathcal{A} \cap \mathrm{conv}\mathcal{B} = \emptyset$, a property which is not usually known in advance to hold.

Consequently, an *error function* of (w, γ), which is convex, piecewise linear and nonnegative, is introduced. It assumes zero value if and only if (w, γ) actually defines a (strictly) separating hyperplane and it has the form:

$$e(w, \gamma) = \sum_{i=1}^{m_1} \max\{0, a_i^\top w - \gamma + 1\} + \sum_{l=1}^{m_2} \max\{0, -b_l^\top w + \gamma + 1\} \tag{2.1}$$

The SVM approach consists in solving the following convex problem:

$$\min_{w,\gamma} \|w\| + Ce(w,\gamma) \tag{2.2}$$

where the addition of the norm of w to the error function is aimed at obtaining a maximum-margin separation, C being a positive trade-off parameter [4].

In the standard approach ℓ_1 or ℓ_2 norms are usually adopted in the definition of problem (2.2), while for feature selection purposes the ℓ_0 pseudo-norm, which counts the number of non-zero components of any vector, is introduced.

The usual notation $\|.\|_0$ for indicating the ℓ_0 pseudo-norm is motivated by the observation

$$(\|.\|_p)^p \to \|.\|_0 \quad \text{when} \quad p \to 0.$$

Relevant properties of function $x \mapsto \|x\|_0$ are:

i) it is *lower–semicontinuous*, that is to say

$$\lim_{k\to+\infty} \inf \|x_k\|_0 \geq \|x\|_0 \quad \text{whenever} \quad x_k \to x,$$

a property which is fundamental in view of using *descent* algorithms;

ii) it is homogeneous of degree 0, ($\|\lambda x\|_0 = \|x\|_0$, for $\lambda \neq 0$); constancy along rays makes difficult the design of minimization algorithms;

iii) the convex hull of $\|.\|_0$ on the ball $\{x|\, \|x\|_\infty \leq r\}$ is exactly the function $\frac{1}{r}\|\|_1$. This is a mathematical justification for frequent substitution of $\|.\|_0$ by $\|.\|_1$, which in fact ensures, in practical applications, attractive *sparsity* properties of the solution.

In our approach we fix $\|w\|_0$ in formulation (2.2), thus we consider a sparse optimization problem of the type

$$f_0^* = \min_{x\in\mathbb{R}^n} f(x) + \|x\|_0, \tag{2.3}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $n \geq 2$, is convex, not necessarily differentiable. We observe, in passing, the significant parallelism between sparse optimization and certain problems in matrix optimization [19], [13]. See, in particular, the recent approach to the rank function minimization described in [9].

In the sequel, we resort to the use of $\|x\|_{[k]}$, the vector k-norm of x, which is defined as the sum of k largest components (in modulus) of x, $k = 1, \ldots, n$. In fact $\|.\|_{[k]}$ is a *polyhedral norm*, intermediate between $\|.\|_1$ and $\|.\|_\infty$.

The following properties hold:

i) $\|x\|_\infty = \|x\|_{[1]} \leq \ldots \leq \|x\|_{[k]} \leq \ldots \|x\|_{[n]} = \|x\|_1$;

ii) $\|x\|_0 \leq k \;\; \Rightarrow \;\; \|x\|_1 - \|x\|_{[s]} = 0, \;\; k \leq s \leq n.$

Moreover, it is easy to prove the equivalence, valid for $1 \leq k \leq n$,

$$\|x\|_0 \leq k \;\; \Leftrightarrow \;\; \|x\|_1 - \|x\|_{[k]} = 0, \tag{2.4}$$

which allows to replace any constraint of the type $\|x\|_0 \leq k$ with a *difference of norms*, that is a DC constraint.

Taking any point $\bar{x} \in \mathbb{R}^n$ and letting $I_{[k]} \stackrel{\triangle}{=} \{i_1, \ldots, i_k\}$ be the index set of k largest in modulus components of $\bar{x}$, a subgradient $\bar{g}^{[k]}$ of the vector k-norm at $\bar{x}$ can be calculated by setting:

$$\bar{g}_i^{[k]} = \begin{cases} 1 & \text{if} \quad i \in I_{[k]} \;\text{ and }\; \bar{x}_i \geq 0 \\ -1 & \text{if} \quad i \in I_{[k]} \;\text{ and }\; \bar{x}_i < 0 \\ 0 & \text{otherwise} \end{cases}$$

To tackle problem 2.3, we start from the observation (see [19], [28]):

$$\|x\|_{[k]} = \max_{y \in \psi_k} y^\top x,$$

where ψ_k is the subdifferential of $\|.\|_0$ at point 0,

$$\psi_k = \{y \in \mathbb{R}^n \,|\, y = u - v,\ 0 \le u,\, v \le e,\ (u+v)^\top e = k\},$$

with e being the vector of n ones. Then we formulate the following problem:

$$f_c^* = \min_{x,u,v,z} f(x) + z \tag{2.5}$$

$$e^\top (u+v) = z \tag{2.6}$$

$$(u-v)^\top x \ge \|x\|_1 \tag{2.7}$$

$$0 \le u,\, v \le e, \quad x \in \mathbb{R}^n. \tag{2.8}$$

REMARK 2.1. *At any feasible point constraint (2.7) is satisfied by equality.*

PROPOSITION 2.2. *At any optimal (local) solution (x^*, u^*, v^*, z^*) of problem 2.5–2.8, the following relations hold for $j = 1, \ldots, d$:*

$$\begin{cases} x_j^* > 0 & \Leftrightarrow \quad u_j^* = 1 \\ x_j^* < 0 & \Leftrightarrow \quad v_j^* = 1 \\ x_j^* = 0 & \Leftrightarrow \quad u_j^* = v_j^* = 0 \end{cases} \tag{2.9}$$

Proof. Note first that u_j^* and v_j^* cannot be both positive. In fact, in such case the solution obtained by replacing u_j^* and v_j^* by $u_j^* - \delta_j$ and $v_j^* - \delta_j$, with $\delta_j \stackrel{\triangle}{=} \min\{u_j^*, v_j^*\} > 0$, would be still feasible and would reduce the objective function value.

Now observe that, while constraint 2.8 ensures

$$-|x_j| \le x_j(u_j - v_j) \le |x_j|, \quad j = 1, \ldots, d,$$

satisfaction of constraint 2.7 guarantees $x_j(u_j - v_j) = |x_j|$, $j = 1, \ldots, d$. Thus it is proved, in particular, $x_j^* > 0 \Rightarrow u_j^* = 1$ ($x_j^* < 0 \Rightarrow v_j^* = 1$).

The implication $x_j^* = 0 \Rightarrow u_j^* = v_j^* = 0$ can be proved by a simple contradiction argument, taking into account optimality of the solution. The same contradiction argument guarantees that the implications $x_j^* > 0 \Leftarrow u_j^* = 1$ ($x_j^* < 0 \Leftarrow v_j^* = 1$) hold true, while the last implication $x_j^* = 0 \Leftarrow u_j^* = v_j^* = 0$ can be proved by observing that $u_j^* = v_j^* = 0$ and $x_j^* \neq 0$ would lead to violation of constraint 2.7. □

REMARK 2.3. *Implications 2.9 ensure $z^* = \|x^*\|_0$. Moreover, letting (x^*, u^*, v^*, z^*) be any local minimum of problem 2.5–2.8, then x^* is a local minimum for problem 2.3.*

By eliminating the scalar variable z in problem 2.5–2.8 we come out with the reformulation

$$f_c^* = \min_{x,u,v} f(x) + e^\top (u+v) \tag{2.10}$$

$$(u-v)^\top x \ge \|x\|_1 \tag{2.11}$$

$$0 \le u,\, v \le e, \quad x \in \mathbb{R}^n, \tag{2.12}$$

which we approach by penalizing the nonlinear nonconvex constraint 2.11 through the scalar penalty parameter $\sigma > 0$. We obtain

$$f_c(\sigma) = \min_{x,u,v} f(x) + e^\top(u+v) + \sigma\big(\|x\|_1 - (u-v)^\top x\big) \tag{2.13}$$

$$0 \le u,\, v \le e, \quad x \in \mathbb{R}^n. \tag{2.14}$$

The objective function 2.13 is suitable for a DC (Difference of Convex) decomposition.

Observe, in fact, that, letting $p^\top \triangleq (x^\top, u^\top, v^\top)$, the function

$$q(p) \triangleq (v-u)^\top x,$$

can be written in DC form as

$$q(p) = p^\top Q_1 p - p^\top Q_2 p,$$

where the symmetric positive semidefinite matrices Q_1 and Q_2 of dimension $(3n, 3n)$ are defined as follows:

$$Q_1 = \frac{1}{4}\begin{pmatrix} 2I & -I & I \\ -I & I & 0 \\ I & 0 & I \end{pmatrix} \tag{2.15}$$

and

$$Q_2 = \frac{1}{4}\begin{pmatrix} 2I & I & -I \\ I & I & 0 \\ -I & 0 & I \end{pmatrix}, \tag{2.16}$$

with I and 0 being, respectively, the identity matrix and the zero matrix of dimension (n, n).

Summing up, the objective function of problem 2.13–2.14 is decomposed as follows:

$$f(x) + e^\top(u+v) + \sigma\big(\|x\|_1 - (u-v)^\top x\big) = h_1(p) - h_2(p),$$

with

$$h_1(p) \triangleq f(x) + e^\top(u+v) + \sigma\big(\|x\|_1 + p^\top Q_1 p\big),$$

and

$$h_2(p) \triangleq \sigma p^\top Q_2 p.$$

Before concluding the Section, we state a property related to parametric normalization of a convex function via the ℓ_1 norm. It will be useful in explaining the role of parameter σ in the numerical experiments of Section 4.

PROPOSITION 2.4. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and define, for $\sigma > 0$, $f_\sigma(x) = f(x) + \sigma\|x\|_1$. If*

$$\sigma > \min_{g \in \partial f(0)} \|g\|_\infty,$$

then $x^* = 0$ is the unique minimum of f_σ.

Proof. Our aim is to prove

$$f(x) + \sigma\|x\|_1 > f(0) \quad \forall x \in \mathbb{R}^n, \quad x \neq 0. \tag{2.17}$$

According to the definition of $g \in \partial f(0)$,

$$f(x) \geq f(0) + g^\top x, \quad \forall x \in \mathbb{R}^n,$$

whence

$$f(x) + \sigma\|x\|_1 \geq f(0) + \sigma\|x\|_1 + g^\top x, \quad \forall x \in \mathbb{R}^n.$$

Since $g^\top x \geq -\|g\|_\infty \|x\|_1$, we infer from the above

$$f(x) + \sigma\|x\|_1 \geq f(0) + (\sigma - \|g\|_\infty)\|x\|_1, \quad \forall x \in \mathbb{R}^n.$$

To ensure (2.17), it suffices fo have $\sigma - \|g\|_\infty > 0$ for one $g \in \partial f(0)$. This is secured with the assumption $\sigma > \min\limits_{g \in \partial f(0)} \|g\|_\infty$.
□

3. The SVM_0 problem. We rewrite first the SVM model (2.2), taking into account the definition (2.1) of the classification error and adopting the ℓ_1 norm. We obtain problem SVM_1.

$$z^* = \min_{w,\gamma,\xi,\zeta} \|w\|_1 + C(\sum_{i=1}^{m_1} \xi_i + \sum_{l=1}^{m_2} \zeta_l) \tag{3.1}$$

subject to

$$a_i^\top w - \gamma + 1 \leq \xi_i, \qquad i = 1, \ldots, m_1 \tag{3.2}$$

$$-b_l^\top w + \gamma + 1 \leq \zeta_l, \qquad l = 1, \ldots, m_2 \tag{3.3}$$

$$\xi_i \geq 0, \qquad i = 1, \ldots, m_1 \tag{3.4}$$

$$\zeta_l \geq 0, \qquad l = 1, \ldots, m_2, \tag{3.5}$$

where the nonnegative auxiliary variables ξ_i, $i = 1, \ldots, m_1$ and ζ_l, $l = 1, \ldots, m_2$ have been introduced to eliminate nonsmoothness in the definition (2.1).

Moreover, by letting

$$w = w^+ - w^-, \;\; w^+, w^- \geq 0,$$

and indicating by e the vector of ones of dimension n, the above problem can be rewritten in a Linear

Programming form as follows

$$z^* = \min_{w^+, w^-, \gamma, \xi, \zeta} e^\top (w^+ + w^-) + C(\sum_{i=1}^{m_1} \xi_i + \sum_{l=1}^{m_2} \zeta_l) \tag{3.6}$$

subject to

$$a_i^\top (w^+ - w^-) - \gamma + 1 \leq \xi_i, \qquad i = 1, \ldots, m_1 \tag{3.7}$$

$$-b_l^\top (w^+ - w^-) + \gamma + 1 \leq \zeta_l, \qquad l = 1, \ldots, m_2 \tag{3.8}$$

$$\xi_i \geq 0, \qquad i = 1, \ldots, m_1 \tag{3.9}$$

$$\zeta_l \geq 0, \qquad l = 1, \ldots, m_2, \tag{3.10}$$

$$w^+ \geq 0, \;\; w^- \geq 0 \tag{3.11}$$

Of course 3.1-3.5 or 3.6-3.11 are equivalent formulations of SVM_1.

We remark that choice of $\|.\|_1$ in (2.2), instead of $\|.\|_2$, has a beneficial effect in terms of feature selection (see [3]).

To guarantee, however, a better control on the number of features actually entering the classification process, we replace $\|.\|_1$ with $\|.\|_0$ and adapt to SVM the sparse optimization approach described in the previous section. We obtain the SVM_0 problem

$$z^* = \min_{w^+, w^-, \gamma, \xi, \zeta, u, v} C(\sum_{i=1}^{m_1} \xi_i + \sum_{l=1}^{m_2} \zeta_l) + e^\top (u + v) \tag{3.12}$$

subject to

$$(u - v)^\top (w^+ - w^-) \geq e^\top (w^+ + w^-) \tag{3.13}$$

$$a_i^\top (w^+ - w^-) - \gamma + 1 \leq \xi_i, \qquad i = 1, \ldots, m_1 \tag{3.14}$$

$$-b_l^\top (w^+ - w^-) + \gamma + 1 \leq \zeta_l, \qquad l = 1, \ldots, m_2 \tag{3.15}$$

$$\xi_i \geq 0, \qquad i = 1, \ldots, m_1 \tag{3.16}$$

$$\zeta_l \geq 0, \qquad l = 1, \ldots, m_2, \tag{3.17}$$

$$w^+ \geq 0, \;\; w^- \geq 0 \tag{3.18}$$

$$0 \leq u, \, v \leq e \tag{3.19}$$

Penalizing the (nonlinear) constraint (3.13) we obtain

$$z^* = \min_{w^+,w^-,\gamma,\xi,\zeta,u,v} e^\top(u+v) + C(\sum_{i=1}^{m_1}\xi_i + \sum_{l=1}^{m_2}\zeta_l) + \sigma\big(e^\top(w^+ + w^-) - (u-v)^\top(w^+ - w^-)\big) \tag{3.20}$$

subject to

$$a_i^\top(w^+ - w^-) - \gamma + 1 \le \xi_i, \quad i = 1,\dots,m_1 \tag{3.21}$$

$$-b_l^\top(w^+ - w^-) + \gamma + 1 \le \zeta_l, l \quad = 1,\dots,m_2 \tag{3.22}$$

$$\xi_i \ge 0, \quad i = 1,\dots,m_1 \tag{3.23}$$

$$\zeta_l \ge 0, \quad l = 1,\dots,m_2, \tag{3.24}$$

$$w^+ \ge 0, \ \ w^- \ge 0 \tag{3.25}$$

$$0 \le u,\, v \le e \tag{3.26}$$

REMARK 3.1. *Definition of an exact nondifferentiable penalty function of the type described in* [5] *would require the introduction into the objective function of the term*

$$\sigma \max\big(0, e^\top(w^+ + w^-) - (u-v)^\top(w^+ - w^-)\big),$$

which in our case, according to remark (2.1), is simply replaced by

$$\sigma\big(e^\top(w^+ + w^-) - (u-v)^\top(w^+ - w^-)\big),$$

giving rise to a differentiable exact penalty function

Problem (3.20)-(3.26) can be put in DC form. In fact, letting $s^\top \stackrel{\triangle}{=} (w^{+T}, w^{-T}, u^\top, v^\top)$, the function

$$r(s) \stackrel{\triangle}{=} (v-u)^\top(w^+ - w^-)$$

can be rewritten as

$$r(s) = s^\top \hat{Q}_1 s - s^\top \hat{Q}_2 s,$$

where the symmetric positive semidefinite matrices $\hat{Q}_1$ and $\hat{Q}_1$ of dimension $(4d, 4d)$ are defined as follows

$$\hat{Q}_1 = \frac{1}{4}\begin{pmatrix} 2I & -2I & -I & I \\ -2I & +2I & I & -I \\ -I & I & I & 0 \\ I & -I & 0 & I \end{pmatrix} \tag{3.27}$$

and

$$\hat{Q}_2 = \frac{1}{4}\begin{pmatrix} 2I & -2I & I & -I \\ -2I & +2I & -I & I \\ I & -I & I & 0 \\ -I & I & 0 & I \end{pmatrix}, \tag{3.28}$$

with I and 0 being, respectively, the identity and the zero matrix of dimension (d, d).

The objective function is then decomposed in DC form $f_1(w^+, w^-, \gamma, \xi, \zeta, u, v) - f_2(w^+, w^-, u, v)$ with

$$f_1(w^+, w^-, \gamma, \xi, \zeta, u, v) = e^\top(u+v) + C(\sum_{i=1}^{m_1} \xi_i + \sum_{l=1}^{m_2} \zeta_l) + \sigma\big(e^\top(w^+ + w^-) + s^\top \hat{Q}_1 s\big),$$

and

$$f_2(w^+, w^-, u, v) = \sigma s^\top \hat{Q}_1 s$$

We apply to problem above the DCA method [15], which tackles the unconstrained minimization of a function $q : \mathbb{R}^n \to \mathbb{R}$

$$q(x) = q_1(x) - q_2(x),$$

with q_1 and q_2 convex, by solving a sequence of linearized convex problems. In particular, letting $x^{(k)}$ be the estimate of a (local) minimum of g at iteration k, the next iterate $x^{(k+1)}$ is calculated as

$$x^{(k+1)} = \arg\min_{x \in \mathbb{R}^n} \{q_1(x) - [q_2(x^{(k)}) + g^{(k)\top}(x - x^{(k)})]\},$$

with $g^{(k)} \in \partial q_2(x^{(k)})$. For other methods to solve DC problems see [14], [8] and the references therein.

We remark that, in applying the successive linearization method to our case, a convex quadratic minimization problem is solved at each iteration.

4. Numerical experiments. We have performed our experiments on two groups of five datasets each. They are the same datasets adopted as benchmark for the feature selection method described in [7]. In particular in datasets 1–5 (Group 1), available at `http://www.tech.plym.ac.uk/spmc/`, the number of samples is small with respect to the number of features. The opposite happens for Datasets 6–10 (Group 2), which are available at `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`. Thus for the latter ones a certain class–overlap is expected.

The datasets are listed in Table 4.1, where $m = m_1 + m_2$ is the total number of samples.

Datasets of Group 1	m	n	Datasets of Group 2	m	n
Carcinoma (CARC)	36	7457	Breast Cancer (BC)	683	10
DLBCL	77	7129	PIMA Indians Diabetes (PIMA)	768	8
Leukemia (LEK)	72	5327	HEART	270	13
Tumor1 (TUM1)	60	7129	Ionosphere (IONO)	351	34
Tumor2 (TUM2)	50	12625	Liver Disorders (LIVER)	145	5

TABLE 4.1
Description of the datasets

As a possible reference, we report first the results of the SVM_1 problem provided by CPLEX solver. A standard tenfold cross validation has been performed. The results are in Table 4.2, where the columns "Test" and "Train" indicate the average testing and training correctness, respectively, expressed as percentage of samples correctly classified. Columns "$\|w\|_1$" and"Time" report the average ℓ_1 norm of w and the average execution time (in seconds). Finally columns "%ft(0)"–"%ft(-9)" report the average percentage of components of w whose modulus is greater than or equal to 10^0–10^{-9}, respectively. Note that, assuming, conventionally, to be equal to "zero" any component w_j of w such that $|w_j| < 10^{-9}$, the percentage of zero-components is $(100 - \%ft(-9))$.

	$C=1$							
Dataset	Test	Train	$\|w\|_1$	Time	%ft(0)	%ft(−2)	%ft(−4)	%ft(−9)
CARC	97.50	100.00	2.63	0,27	0.00	0.26	0.28	0.28
DLBCL	94,29	100.00	5.07	0.54	0.00	0.42	0.44	0.44
LEUK	96.67	100.00	4.45	0.37	0.00	0.54	0.58	0.58
TUM1	76.67	100.00	7.09	0.52	0.00	0.53	0.55	0.55
TUM2	80.50	100.00	5.12	0.69	0.00	0.23	0.24	0.24
	$C=10$							
BC	97.24	94.00	6.01	0,01	9.00	94.00	94.00	94.00
PIMA	77.62	100.00	8.30	0,02	26.25	100.00	100.00	100.00
HEART	86.83	100.00	7.27	0,01	12.31	100.00	100.00	100.00
IONO	95.46	95.59	43.43	0,01	50.59	95.00	95.59	95.59
LIVER	74.39	100.00	5.14	0,00	48.00	100.00	100.00	100.00

TABLE 4.2
SVM_1 - Cplex implementation - Tenfold Cross Validation

Two different values of parameter C, obtained through the so called "Model selection" phase (see [7]), have been adopted for the two dataset groups.

Before reporting the results of the implementation of our algorithm to solve SVM_0, we illustrate in Tables 4.3 and 4.4 the results provided by the following mixed binary programming problem MBP [7] for feature selection, where ℓ_0 pseudo–norm minimization is pursued by introducing the set of binary variables y_j, $j = 1, \ldots, n$.

$$z^* = \min_{w,\gamma,\xi,\zeta,y} \|w\|_1 + C(\sum_{i=1}^{m_1} \xi_i + \sum_{l=1}^{m_2} \zeta_l) + D \sum_{j=1}^{n} y_j$$

subject to

$$\begin{aligned}
& a_i^\top w + \gamma + 1 \leq \xi_i, && i = 1, \ldots, m_1 \\
& -b_l^\top w - \gamma + 1 \leq \zeta_l, && l = 1, \ldots, m_2 \\
& -u_j y_j \leq w_j \leq u_j y_j, && j = 1, \ldots, n \\
& \xi_i \geq 0, && i = 1, \ldots, m_1 \\
& \zeta_l \geq 0, && l = 1, \ldots, m_2 \\
& y_j \in \{0, 1\}, && j = 1, \ldots, n,
\end{aligned}$$

where $u_j > 0$, $j = 1, \ldots, n$, is a given bound on the modulus of the j-th component of w (see [7] for a discussion on setting the u_j's) . The binary variable y_j, $j = 1, \ldots, n$ is equal to 1 at the optimum if and only if $w_j \neq 0$, $j = 1, \ldots, n$. Consequently the term $\sum_{j=1}^{n} y_j$ in the objective function represents the ℓ_0 pseudo–norm of w. The positive parameter D provides the tradeoff between the ℓ_0 pseudo–norm objective and that of the SVM_1 problem (3.6)–(3.11).

The results are those obtained by CPLEX, with maximum running time of 1000 seconds. Note that on some test problems, the maximum running time has been achieved with no optimality certification. The results we provide in such cases are related to the best solution found.

Some comments are in order. We observe first that the average computation time for solving the

	$C = 1\,,\ D = 0.01$							
Dataset	Test	Train	$\|w\|_1$	Time	%ft(0)	%ft(−2)	%ft(−4)	%ft(−9)
CARC	92.50	100	2.65	302.07	0.00	0.16	0.16	0.16
DLBCL	92.86	100	5.11	999.98	0.00	0.28	0.28	0.28
LEUK	93.57	100	4.48	999.74	0.00	0.35	0.35	0.35
TUM1	76.67	100	7.12	998.91	0.00	0.39	0.39	0.39
TUM2	83.00	100	5.16	999.89	0.00	0.16	0.16	0.16

TABLE 4.3
MBP - Cplex implementation - Tenfold cross validation - Group 1

	$C = 10\,,\ D = 10$							
Dataset	Test	Train	$\|w\|_1$	Time	%ft(0)	%ft(−2)	%ft(−4)	%ft(−9)
BC	96.42	97.15	5.73	0.06	17.00	68.00	68.00	68.00
PIMA	75.87	77.26	7.75	0.06	26.25	78.75	78.75	78.75
HEART	83.95	86.97	5.16	0.07	9.23	58.46	58.46	58.46
IONO	88.29	94.62	27.06	2.03	36.18	45.29	45.29	45.29
LIVER	71.10	74.05	4.45	0.02	52.00	68.00	68.00	68.00

TABLE 4.4
MBP - Cplex implementation - Tenfold cross validation - Group 2

Linear Program SVM_1 is negligible.

As for classification performance, comparison of Table 4.2 with Tables 4.3 and 4.4 highlights that the use of an explicit feature selection mechanism results in a mild downgrading of the classification correctness. Such phenomenon is compensated by the reduction in the percentage of the numerically significant features (columns"ft(0)"–"ft(-9)").

Coming now to our method (referred to, in the sequel, as the "SVM_0 Algorithm"), we report in Table 4.5 the results on the two groups of datasets.

We have added the two columns "%Viol." and "$e^\top(u+v)$". In particular, column "%Viol." reports the percentage ratio between the average violation of the relaxed constraint 3.13 and the average norm $\|w\|_1$. Column "$e^\top(u+v)$" reports the average value of the scalar product $e^\top(u+v)$, which, for small values of the companion parameter "%Viol.", reasonably approximates $\|w\|_0$. ‡

		$C = 1$									
Dataset	σ	Test	Train	$\|w\|_1$	Time	%Viol.	$e^\top(u+v)$	%ft(0)	%ft(−2)	%ft(−4)	%ft(−9)
CARC	4	97.50	100,00	6.90	39.40	3	2.10	0.03	0.05	0.06	0.06
DLBCL	4	91.43	98.57	12.83	260.48	0.5	3.51	0.05	0.06	0.06	0.06
LEUK	5	98.33	99.83	16.02	188.70	0.0	3.70	0.07	0.07	0.07	0.07
TUM1*	5	76.67	100.00	15.27	50.74	4.7	7.23	0.08	0.14	0.16	0.16
TUM2	4	81.00	96.03	7.40	69.68	7.29	3.05	0.02	0.05	0.06	0.06
		$C = 10$									
Dataset	σ	Test	Train	$\|w\|_1$	Time	%Viol.	$e^\top(u+v)$	%ft(0)	%ft(−2)	%ft(−4)	%ft(−9)
BC	63	93.17	92,48	1.36	0.59	0.29	1.32	2.00	34.00	34.00	34.00
PIMA	17	75.57	76.19	5.89	3.80	0.0	3.50	26.25	43.75	43.75	43.75
HEART**	17	83.95	86.92	7.28	0.1	29.35	6.95	13.85	99.23	100.00	100.00
IONO	17	86.35	89,31	3.14	0.27	0.47	2.56	5.88	13.82	13.82	13.82
LIVER**	6	74.89	74.39	5.77	0.02	10.19	3.93	64.00	100.00	100.00	100.00

TABLE 4.5
SVM_0 Algorithm - Tenfold cross validation

We have run the algorithm on each dataset for different values of the penalty parameter σ, and we indicate in Table 4.5 the specific value of σ the results refer to.

Some comments follow.

- In terms of classification correctness, the results of the SVM_0 Algorithm are comparable with those of MBP.
- The SVM_0 Algorithm provides better results in terms of number of zero-components of w. In fact, the percentage of components conventionally assumed equal to zero, that is $(100-\%ft(-9))$, is significantly bigger, except that in two cases, in SVM_0 Algorithm than in MBP.
- The computation time is, for both algorithms, negligible on the datasets of Group 2 while it is remarkably smaller for SVM_0 Algorithm as far as datasets of Group 1 are concerned.

In running the SVM_0 Algorithm, the most relevant issue is the appropriate tuning of the penalty parameter $\sigma > 0$, which is, of course, dataset-specific. Two aspects are to be taken into account.

- "Small" values of σ may lead to significant violation of the penalized constraint at the optimum of problem 3.20–3.23. Note that in such case, variables u and v may loose their "marker" role highlighted in Proposition 2.2.
- "Large" values of σ may result in trivial solutions ($w = 0$) to the penalized problem (see Proposition 2.4). .

To illustrate in details the impact on the solution of parameter σ, we analyze the results for increasing values of σ. In particular we focus on the datasets CARC and IONO, from the first and second group, respectively. The results are reported in Tables 4.6 and 4.7.

$Dataset: CARC; C = 1$										
σ	Test	Train	$\|w\|_1$	Time	% Viol.	$e^\top(u+v)$	%ft(0)	%ft(−2)	%ft(−4)	%ft(−9)
1	95.00	100,00	2.62	14.45	100	0.00	0.00	0.23	0.25	0.25
2	94.17	100.00	3.50	73.13	44	0.87	0.01	0.14	0.15	0.15
4	97.50	100,00	6.90	39,40	3	2.10	0.03	0.05	0.06	0.06
5	91.67	100.00	6.26	41.41	1	2.07	0.02	0.03	0.03	0.03
6	85.00	98.33	7.23	68.53	0.1	2.18	0.03	0.03	0.03	0.03
7	75.83	96.99	6.08	93.80	0	2.20	0.03	0.03	0.03	0.03
12	65.83	81,40	2.32	81.04	0	0.80	0.01	0.01	0.01	0.01
16	58.33	57.59	0.0	1.81	0	0.0	0.0	0.0	0.0	0.0

TABLE 4.6
Results on dataset CARC for different values of σ

$Dataset: IONO; C = 10$										
σ	Test	Train	$\|w\|_1$	Time	% Viol.	$e^\top(u+v)$	%ft(0)	%ft(−2)	%ft(−4)	%ft(−9)
1	87.61	94.23	19.39	0.36	52	4.16	18.82	72.94	73.24	73.24
2	87.94	93.53	11.21	0.22	50	4.34	9.41	62.06	63.53	63.53
4	87.96	92.55	7.58	0.26	42	4.31	6.18	47.35	47.35	47.35
10	87.63	89.38	4.23	0.24	19	2.90	5.88	18.53	19.41	19.41
17	86.35	89.31	3.14	0.29	15	2.56	5.88	13.82	13.82	13.82
30	80.80	83.47	1.83	0.23	6	1.77	3.82	7.65	7.94	7.04
50	74.77	74.68	1.00	0.26	3	0.97	2.94	2.94	2.94	2.94
64	64.33	64.24	0.0001	1.84	0.0	0.0001	0.0	0.0	0.29	0.29
93	64.33	64.24	0.0	0.16	0	0.0	0.0	0.0	0.0	0.0

TABLE 4.7
Results on dataset IONO for different values of σ

[‡]Symbols "*" and "**" in Dataset column of Table 4.5 indicate that parameter C has been set to 2 and 100, respectively.

We observe that on both datasets increasing values of σ results, as expected, in deterioration of the classification correctness (under such point of view, significant data are those obtained in the training phase). On the other hand, the violation of the relaxed constraint gets smaller and smaller, until, for too large values of σ, the results become meaningless, as w gets close to zero.

5. Conclusions. We have tackled the Feature Selection problem within SVM binary classification by using the polyhedral k-norm, in the sparse optimization context. The numerical experiments show that the approach is a promising alternative to continuous approximations of the ℓ_0 pseudo–norm and to integer programming-based methods.

REFERENCES

[1] E. Amaldi, V. Kann, On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems, Theoretical Computer Science, 1998, 209, 12, pp. 237–260.

[2] P. Bertolazzi, G. Felici, P. Festa, G. Fiscon, E. Weitschek, Integer programming models for feature selection: New extensions and a randomized solution algorithm, European Journal of Operational Research, 2016, 250, 2, pp. 389–399.

[3] P.S. Bradley, O.L. Mangasarian, W.N. Street, Feature selection via mathematical programming, INFORMS Journal on Computing, 1998, 10, 2, pp. 209-217.

[4] N. Cristianini, J. Shawe–Taylor, 20 0 0. An introduction to support vector machines and other kernel-based learning methods, 2000, Cambridge University Press.

[5] G. Di Pillo, L. Grippo, Exact penalty functions in constrained optimization, SIAM Journal on Control and Optimization, 1989, 27, 6, pp. 1333-1360.

[6] G. Gasso, A. Rakotomamonjy, S. Canu, Recovering Sparse Signals With a Certain Family of Nonconvex Penalties and DC Programming, IEEE Transactions on Signal Processing, 2009, 57, 12, pp. 4686–4698.

[7] M. Gaudioso, E. Gorgone, M. Labbé, A. M. Rodríguez-Chía, Lagrangian relaxation for SVM feature selection, Computers and Operations Research 2017, 87, pp. 137–145.

[8] M. Gaudioso, G. Giallombardo, G. Miglionico, A. M. Bagirov, Minimizing nonsmooth DC functions via successive DC piecewise-affine approximations, Journal of Global Optimization, 2018, 71, 1, pp.37–55.

[9] J.Gotoh, A. Takeda, K.Tono, DC formulations and algorithms for sparse optimization problems, Mathematical Programming, Ser. B, 2018, 169, 1, pp. 141–176.

[10] I. Guyon, A. Elisseeff, 2003. An introduction to variable and feature selection, Journal of Machine Learning Research, 2003, 3, pp. 11571182.

[11] A. B. Hempel, P. J. Goulart, A Novel Method for Modelling Cardinality and Rank Constraints, 53rd IEEE Conference on Decision and Control, Los Angeles, Cal., USA Dec. 15–17, 2014, pp. 4322–4327.

[12] J.-B. Hiriart-Urruty, Generalized differentiability/ duality and optimization for problems dealing with differences of convex functions, Lecture Notes in Economic and Mathematical Systems, 1986, 256, Springer Verlag, pp. 37–70.

[13] J.-B. Hiriart-Urruty, D. Ye, Sensitivity analysis of all eigevalues of a symmetric matrix, Numerische Mathematik, 1995, 70, 1, pp. 45–72.

[14] K. Joki, A.M. Bagirov, N. Karmitsa, and M.M. Mäkelä, A proximal bundle method for nonsmooth DC optimization utilizing nonconvex cutting planes, Journal of Global Optimization, 2017, 68, 3, pp. 501-535. —

[15] H.A. Le Thi, T. Pham Dinh, The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems, Journal of Global Optimization, 2005, 133, pp. 23–46.

[16] H.A. Le Thi, H.M. Le, V.V. Nguyen, T. Pham Dinh, A DC programming approach for feature selection in support vector machines learning, Advances in Data Analysis and Classification, 2008, 2, pp. 259278.

[17] S. Maldonado, J. Pérez, R. Weber, M. Labbé, Feature selection for Support Vector Machines via Mixed Integer Linear Programming, Information Sciences, 2014, 279, pp. 163–175.

[18] O.L. Mangasarian, Nonlinear Programming, 1969, McGraw-Hill.

[19] M.L. Overton, R.S. Womersley, Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices, Mathematical Programming, 1993, 62, 1-3, pp. 321-357.

[20] M. Pilanci, M. J. Wainwright, L. El Ghaoui, Sparse learning via Boolean relaxations, Mathematical Programming, Ser. B, 2015, 151, pp. 63–87.

[21] F. Rinaldi, F. Schoen, M. Sciandrone, Concave programming for minimizing the zero-norm over polyhedral sets, Computational Optimization and Applications, 2010, 46, pp. 467486.

[22] A.S. Strekalovsky, Global optimality conditions for nonconvex optimization, Journal of Global Optimization, 1998, 12, pp. 415–434.
[23] E. Soubies, L. Blanc–Féraud, G. Aubert, A Unified View of Exact Continuous Penalties for ℓ_2-ℓ_0 Minimization, SIAM Journal on Optimization, 2017, 27, 3, pp. 2034-2060.
[24] V. Vapnik, The nature of the statistical learning theory, 1995, Springer Verlag.
[25] G.A. Watson, Linear best approximation using a class of polyhedral norms, Numerical Algorithms, 1992, 2 pp. 321–336.
[26] J. Weston, A. Elisseeff, B. Schölkopf, M. Tipping, Use of the zero-norm with linear models and kernel methods, Journal of Machine Learning Research, 2003, 3, pp. 1439-1461.
[27] S. J. Wright, Accelerated block–cordinate relaxation for regularized optimization, SIAM Journal on Optimization, 2012, 22, 1, pp. 159–186.
[28] B. Wu, C. Ding, D. Sun, K-C. Toh, On the Moreau-Yosida regularization of the vector $k-$norm related functions, SIAM Journal on Optimization, 2014, 24. 2, pp. 766-794.