Computational Optimization and Applications (2023) 86:745–766
https://doi.org/10.1007/s10589-023-00506-y

Sparse optimization via vector k-norm and DC programming with an application to feature selection for support vector machines

Manlio Gaudioso[1] · **Giovanni Giallombardo**[1] · **Giovanna Miglionico**[1]

Received: 27 February 2023 / Accepted: 24 June 2023 / Published online: 12 July 2023

Abstract
Sparse optimization is about finding minimizers of functions characterized by a number of nonzero components as small as possible, such paradigm being of great practical relevance in Machine Learning, particularly in classification approaches based on support vector machines. By exploiting some properties of the k-norm of a vector, namely, of the sum of its k largest absolute-value components, we formulate a sparse optimization problem as a mixed-integer nonlinear program, whose continuous relaxation is equivalent to the unconstrained minimization of a difference-of-convex function. The approach is applied to Feature Selection in the support vector machine framework, and tested on a set of benchmark instances. Numerical comparisons against both the standard ℓ_1-based support vector machine and a simple version of the Slope method are presented, that demonstrate the effectiveness of our approach in achieving high sparsity level of the solutions without impairing test-correctness.

Keywords Global optimization · Sparse optimization · Cardinality constraint · k-norm · Support vector machine

1 Introduction

The sparse counterpart of a mathematical program is aimed at finding optimal solutions that have as few nonzero components as possible, this feature playing a significant role

✉ Manlio Gaudioso
manlio.gaudioso@unical.it

Giovanni Giallombardo
giovanni.giallombardo@unical.it

Giovanna Miglionico
gmiglionico@dimes.unical.it

[1] Dipartimento di Ingegneria Informatica, Modellistica, Elettronica e Sistemistica (DIMES), Università della Calabria, Via Pietro Bucci, 87036 Rende, CS, Italy

in several applications, mainly (but not solely) arising in the broad fields of data science and machine learning. Indeed, this is one of the reason why *sparse optimization* has recently become a topic of great interest in the field of mathematical optimization.

Accounting for sparsity in a mathematical program amounts to embed into its formulation a measure dependent on the number of zero components of a solution-vector. The mathematical tool that is most naturally suited to such task is the ℓ_0-(pseudo-)norm $\|\cdot\|_0$, often simply referred to as the ℓ_0-norm, defined as the number of nonzero components of a solution, it rather being a measure of vector density. Thus, a sparse optimization problem should also encompass the ℓ_0-norm minimization in order to achieve the largest possible sparsity. Such feature highlights the intrinsic bi-objective nature of sparse optimization, that makes it fit to two commonplace single-objective formulations, the ℓ_0-*regularization* problem and the *cardinality-constrained* problem, depending on whether one aims at pushing for sparsity or at ensuring an assigned sparsity level, respectively.

Given a real-valued function $f : \mathbb{R}^n \to \mathbb{R}$, with $n \geq 2$, the (unconstrained) ℓ_0-regularization problem has the following structure

$$\min \left\{ f(\mathbf{x}) + \sigma \|\mathbf{x}\|_0 : \mathbf{x} \in \mathbb{R}^n \right\}, \tag{SOP}$$

where the ℓ_0-norm penalty term inside the objective function pushes towards sparse solutions, and a fixed penalty-parameter $\sigma > 0$ ensures the trade-off between the two (possibly) conflicting objectives. Problem (SOP) has a nonconvex and discontinuous nature (see, e.g., [1] for a study on complexity issues), that makes it unfit to be faced by means of standard nonlinear optimization approaches. In fact, it has been usually tackled by replacing the ℓ_0-norm with the ℓ_1-norm, thus obtaining an optimization problem, the ℓ_1-regularization one (SOP_1), whose tractability rather depends on the properties of f, the ℓ_1 term being a convex function. It has been proved (see [16, 17, 21]) that, under appropriate conditions the solutions of (SOP) and (SOP_1) coincide. Nevertheless, in many practical applications the equivalence conditions do not hold true and the solutions obtained from the ℓ_1 minimization problem are *less sparse* that those of the ℓ_0 problem.

A different viewpoint in dealing with the bi-objective nature of sparse optimization is given by the cardinality-constrained formulation

$$\min \left\{ f(\mathbf{x}) : \|\mathbf{x}\|_0 \leq t, \ \mathbf{x} \in \mathbb{R}^n \right\}, \tag{SOP^t}$$

where sparsity is ensured by forcing the ℓ_0-norm to be not larger than a given integer $t \in \{1, \ldots, n-1\}$, thus obtaining minimizers with at least $n - t$ zero components. Such a problem has been studied in depth in recent years due to its possible application in areas of relevant interest such as compressed sensing [21] and portfolio selection [10]. We refer the reader to [37] for an up-to-date survey. Here, we only mention [4], on the theoretical side, where necessary optimality conditions have been analyzed, and

[15, 24], on the methodological side, where reformulations as mathematical programs with complementarity constraints have been proposed.

In this paper we focus on sparse optimization as an effective way to deal with Feature Selection (FS) in Machine Learning (ML), see [32], where the problem is to detect the sample parameters which are really significant in applications such as unsupervised learning (see [22]) and regression (see [45]). In particular, we restrict our attention to supervised binary classification, an area where intensive research activity has been performed in the last decades, mainly since the advent of Support Vector Machine (SVM) as the election classification methodology, see [19, 47].

Most of the applications of sparse optimization to feature selection fall into two classes of methods, those where the ℓ_0-norm has been approximated by means of appropriate concave functions (see, e.g., [13, 14, 42, 49]) and those where sparsity of the solution has been enforced by resorting to the definition of appropriate sets of binary variables, giving rise to Mixed Integer Linear (MILP) or even Mixed Integer Non Linear (MINLP) problems. In fact, it is natural to associate to any continuous variable an appropriate binary one according to its property of being zero or nonzero. Mixed integer reformulations have been widely adopted in relevant areas of Machine Learning such as classification, logistic regression, medical scoring etc. (see e.g. [7, 8, 20, 43, 46] and the references therein). As for the SVM literature we cite here [6, 29, 39].

The model we propose is inspired by the possibility of connecting the ℓ_0-norm to the *vector k-norm*, which is defined as the sum of the k largest absolute-value components of any vector, and is a polyhedral norm, thus relatively easy to be treated by standard tools of convex nonsmooth optimization. Properties of the vector k-norm have been investigated in several papers (see, e.g., [30, 41, 50]). We particularly recall the applications to matrix completion, see [40], and to linear system approximation, see [48], as a possible alternative to the use of classic ℓ_1, ℓ_2 and ℓ_∞ norms.

More recently, vector k-norm has been proved to be an effective tool for dealing with sparse optimization, see [31], also with specific applications to feature selection in the SVM framework, see [28]. The use of k-norm is somehow evoked also in the trimmed LASSO model described in [9].

The novelty of this paper is the development of a continuous k-norm-based model which is obtained as the continuous relaxation of a MINLP problem. The objective function comes out to be nonconvex, in particular of the Difference-of-Convex (DC) type, thus our approach is closer to the trimmed LASSO [9] than to the SLOPE method [5, 11]. Other DC-like approaches are presented in [23, 38, 52].

The remainder of the article is organized as follows. In Sect. 2 we summarize the properties of vector k-norms, and we introduce a couple of MINLP reformulations of the ℓ_0-regularization problem (SOP). In Sect. 3 we consider the continuous relaxations of both such reformulations, and we focus in Sect. 4 to the one based on vector k-norm, which consists of a DC (Difference of Convex) optimization problem. In Sect. 5 we cast the SVM-based Feature Selection problem into our sparse optimization setting. Finally, in Sect. 6 we report on the computational experience, obtained on some benchmark datasets for binary classification, of both relaxed MINLP reformulations, highlighting the role of k-norms in increasing sparsity levels of the solutions.

Notation Vectors are represented in lower-case bold letters, with $\mathbf{e}$ and $\mathbf{0}$ representing vectors with all the elements equal to one and to zero, respectively. The inner product of vectors $\mathbf{x}$ and $\mathbf{y}$ is denoted by $\mathbf{x}^\top\mathbf{y}$. Given any vector in $\mathbb{R}^n$, say $\mathbf{x} = (x_1, \ldots, x_n)^\top$, we denote the ℓ_p-norm, with $1 \leq p < +\infty$, and the ℓ_∞-norm, respectively, by

$$\|\mathbf{x}\|_p \triangleq \left(\sum_{j=1}^{n} |x_j|^p\right)^{1/p} \qquad \text{and} \qquad \|\mathbf{x}\|_\infty \triangleq \max\left\{|x_j| : j \in \{1, \ldots, n\}\right\}.$$

Furthermore, we recall the definition of the ℓ_0-norm

$$\|\mathbf{x}\|_0 \triangleq \left|\left\{i : x_i \neq 0,\ i \in \{1, \ldots, n\}\right\}\right|,$$

and some of its relevant properties:

(i) $\|\mathbf{x}\|_0 = 0 \quad \Leftrightarrow \quad \mathbf{x} = \mathbf{0}$;
(ii) $\|\mathbf{x} + \mathbf{y}\|_0 \leq \|\mathbf{x}\|_0 + \|\mathbf{y}\|_0$;
(iii) $\|\alpha\mathbf{x}\|_0 \neq |\alpha|\|\mathbf{x}\|_0$;
(iv) $(\|\cdot\|_p)^p \to \|\cdot\|_0 \quad$ when $\quad p \to 0$;
(v) $\liminf_{\mathbf{x}\to\bar{\mathbf{x}}} \|\mathbf{x}\|_0 \geq \|\bar{\mathbf{x}}\|_0$, i.e., $\|\cdot\|_0$ is *lower semicontinuous*.

2 MINLP-based formulations of the ℓ_0-regularization problem

Introducing an n-dimensional binary decision-vector $\mathbf{z}$, associated to $\mathbf{x}$, whose kth component z_k is set to one if $x_k > 0$ and set to zero otherwise, an obvious and classic reformulation of problem (SOP) is the following

$$\begin{aligned} \text{(MISOP)} \quad \min \quad & f(\mathbf{x}) + \sigma\mathbf{e}^\top\mathbf{z} && (1)\\ \text{s.t.} \quad & \mathbf{x} \geq -M\mathbf{z} && (2)\\ & \mathbf{x} \leq M\mathbf{z} && (3)\\ & \mathbf{z} \in \{0,1\}^n, && (4) \end{aligned}$$

where the positive "big" M parameter denotes a uniform bound on the absolute value of any *single* component of $\mathbf{x}$. It is easy to verify that any minimizer $(\mathbf{x}^*, \mathbf{z}^*)$ of (MISOP) is such that

$$x_k^* \neq 0 \quad \Leftrightarrow \quad z_k^* = 1, \tag{5}$$

and, consequently, the term $\mathbf{e}^\top\mathbf{z}^*$ actually represents the ℓ_0-norm of $\mathbf{x}^*$. It is worth observing that an ideal threshold for M is $M^* \triangleq \|\mathbf{x}^*\|_\infty$, where $\mathbf{x}^*$ is any global minimum of (SOP), as indeed any $M \geq M^*$ guarantees equivalence of (SOP) and (MISOP).

In view of introducing an alternative MINLP-based reformulation of (SOP), we first recall, for $k \in \{1, \ldots, n\}$, the polyhedral *vector k-norm* $\|\mathbf{x}\|_{[k]}$, defined as the sum of the k largest unsigned components of $\mathbf{x}$. In particular, given $\mathbf{x} = (x_1, \ldots, x_n)^\top$, and

adopting the following notation

$$\begin{aligned} j_1 &\triangleq \arg\max\{|x_1|, \ldots, |x_n|\} \\ j_2 &\triangleq \arg\max\{|x_1|, \ldots, |x_n|\} \setminus \{|x_{j_1}|\} \\ j_3 &\triangleq \arg\max\{|x_1|, \ldots, |x_n|\} \setminus \{|x_{j_1}|, |x_{j_2}|\} \\ &\vdots \\ j_n &\triangleq \arg\min\{|x_1|, \ldots, |x_n|\}, \end{aligned}$$

such that

$$|x_{j_1}| \geq |x_{j_2}| \geq \ldots \geq |x_{j_n}|, \tag{6}$$

the vector k-norm of $\mathbf{x}$ can be expressed as

$$\|\mathbf{x}\|_{[k]} \triangleq \sum_{s=1}^{k} |x_{j_s}|. \tag{7}$$

Moreover, it is easy to see that $\|\cdot\|_{[k]}$ fulfills the following properties:

$$\|\mathbf{x}\|_\infty = |x_{j_1}| = \|\mathbf{x}\|_{[1]} \leq \|\mathbf{x}\|_{[2]} \leq \ldots \leq \|\mathbf{x}\|_{[n]} = \|\mathbf{x}\|_1, \tag{8}$$

$$\|\mathbf{x}\|_0 \leq k \iff \|\mathbf{x}\|_1 = \|\mathbf{x}\|_{[s]} \ \forall s \in \{k, \ldots, n\}. \tag{9}$$

Now, focusing in particular on the equivalence (9), and introducing an n-dimensional binary decision-vector $\mathbf{y}$, whose kth component y_k is set to one if $\|\mathbf{x}\|_1 - \|\mathbf{x}\|_{[k]} > 0$ and set to zero otherwise, it is possible to state yet another mixed binary reformulation of problem (SOP) as the following

$$\begin{aligned} (\text{MI}k\text{SOP}) \quad \min \quad & f(\mathbf{x}) + \sigma \mathbf{e}^\top \mathbf{y} && && (10) \\ \text{s.t.} \quad & \|\mathbf{x}\|_1 - \|\mathbf{x}\|_{[k]} \leq M' y_k && \forall k \in \{1, \ldots, n\} && (11) \\ & \mathbf{y} \in \{0, 1\}^n && && (12) \end{aligned}$$

where the sufficiently large parameter $M' > 0$ has been introduced. An equivalence between (MIkSOP) and (SOP) can be obtained in the constrained case where $\mathbf{x}$ is restricted to stay in a compact set $S \subset \mathbb{R}^n$. In fact, letting $D \triangleq \max_{\mathbf{x} \in S} \|\mathbf{x}\|_1$, and observing that

$$\|\mathbf{x}\|_1 - \|\mathbf{x}\|_{[k]} \leq \|\mathbf{x}\|_1 - \|\mathbf{x}\|_\infty \leq \left(1 - \frac{1}{n}\right) \|\mathbf{x}\|_1 \leq \left(1 - \frac{1}{n}\right) D,$$

then any $M' \geq \left(1 - \frac{1}{n}\right) D$ ensures the equivalence. Unlike (MISOP) formulation (1)–(4), the (MIkSOP) formulation (10)–(12) is characterized by the presence of a set of nonconvex constraints (11), whose left hand sides are, in particular, DC functions. Furthermore, as for the objective function, we remark that the only feasible solutions of (MIkSOP) which are candidates to be optimal are those $(\mathbf{x}, \mathbf{y})$ for which the equivalence

$$\|\mathbf{x}\|_1 - \|\mathbf{x}\|_{[k]} = 0 \iff y_k = 0 \tag{13}$$

holds for every $k \in \{1, \ldots, n\}$. Observing that at any of such solutions it is necessarily $y_n = 0$ and that, provided $\mathbf{x} \neq \mathbf{0}$, there holds

$$\mathbf{e}^\top \mathbf{y} = \max \Big\{ s : \ \|\mathbf{x}\|_1 - \|\mathbf{x}\|_{[s]} > 0, \ s \in \{1, \ldots, n\} \Big\},$$

it follows that

$$\mathbf{e}^\top \mathbf{y} = \|\mathbf{x}\|_0 - 1,$$

which implies that if $(\mathbf{x}^*, \mathbf{y}^*)$ is optimal for problem (MIkSOP), then $\mathbf{x}^*$ is optimal for problem (SOP) as well.

In the sequel, we analyze the continuous relaxation of the MINLP-based formulations (MISOP) and (MIkSOP), assuming in particular that f is a convex function, not necessarily differentiable.

3 Continuous relaxation of MINLP-based SOP formulations

We start by studying the continuous relaxation (RSOP) of (MISOP), obtained by replacing the binary constraints $\mathbf{z} \in \{0, 1\}^n$ with $\mathbf{0} \leq \mathbf{z} \leq \mathbf{1}$. A simple contradiction argument ensures that any minimizer $(\overline{\mathbf{x}}, \overline{\mathbf{z}})$ of (RSOP) satisfies by equality at least one constraint of the pair $x_k \leq M z_k$ and $x_k \geq -M z_k$, for every $k \in \{1, \ldots, n\}$. More precisely, if $\overline{x}_k \neq 0$ then either $M\overline{z}_k = \overline{x}_k > 0 > -M\overline{z}_k$, or $-M\overline{z}_k = \overline{x}_k < 0 < M\overline{z}_k$, while if $\overline{x}_k \neq 0$ then $\overline{z}_k = 0$. Summarizing, $(\overline{\mathbf{x}}, \overline{\mathbf{z}})$ is such that $\overline{z}_k = \dfrac{|\overline{x}_k|}{M}$ for every $k \in \{1, \ldots, n\}$, from which it follows that

$$\mathbf{e}^\top \overline{\mathbf{z}} = \frac{1}{M} \|\mathbf{x}\|_1,$$

and the relaxed problems (RSOP) can be written as

$$\min \Big\{ \overline{F}(\mathbf{x}) \triangleq f(\mathbf{x}) + \frac{\sigma}{M} \|\mathbf{x}\|_1 : \ \mathbf{x} \in \mathbb{R}^n \Big\}. \tag{RSOP}$$

In other words, the continuous relaxation of (MISOP) just provides the ℓ_1-regularization of function f, often referred to as LASSO model, a convex, possibly nonsmooth program due to the assumption made on f (for a discussion on LASSO and possible variants see, e.g., [52]).

Consider now the continuous relaxation (RkSOP) of (MIkSOP), where $\mathbf{0} \leq \mathbf{y} \leq \mathbf{1}$ replace $\mathbf{y} \in \{0, 1\}^n$. Note that any minimizer $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ of (RkSOP) satisfies by equality all the constraints (11), since if this were not the case for some k, the corresponding $\widetilde{y}_k$ would be reduced, leading to a reduction in the objective function without impairing feasibility. Consequently,every continuous variable $\widetilde{y}_k$, with $k \in \{1, \ldots, n\}$, is such that

$$\widetilde{y}_k = \frac{1}{M'} \big(\|\widetilde{\mathbf{x}}\|_1 - \|\widetilde{\mathbf{x}}\|_{[k]} \big),$$

implying that

$$\mathbf{e}^\top \widetilde{\mathbf{y}} = \frac{1}{M'}\left(n\|\widetilde{\mathbf{x}}\|_1 - \sum_{k=1}^{n} \|\widetilde{\mathbf{x}}\|_{[k]}\right).$$

As a consequence, (RkSOP) can be stated as

$$\min\left\{\widetilde{F}(\mathbf{x}) \triangleq f(\mathbf{x}) + \frac{\sigma}{M'}\left(n\|\mathbf{x}\|_1 - \sum_{k=1}^{n} \|\mathbf{x}\|_{[k]}\right) : \mathbf{x} \in \mathbb{R}^n\right\}. \qquad \text{(R}k\text{SOP)}$$

We observe that, recalling (7), the sum of the k-norm of $\mathbf{x}$ can be written as

$$\sum_{k=1}^{n} \|x\|_{[k]} = n|x_{j_1}| + (n-1)|x_{j_2}| + \ldots + |x_{j_n}|. \tag{14}$$

Hence, taking into account the definition of $\|\mathbf{x}\|_1$, we obtain that

$$\begin{aligned}\left(n\|\mathbf{x}\|_1 - \sum_{k=1}^{n} \|\mathbf{x}\|_{[k]}\right) &= \left(|x_{j_2}| + 2|x_{j_3}| + \ldots + (n-1)|x_{j_n}|\right) \\ &= \left(\sum_{s=2}^{n}(s-1)|x_{j_s}|\right),\end{aligned} \tag{15}$$

and, consequently, we come out with the problem

$$\min\left\{\widetilde{F}(\mathbf{x}) \triangleq f(\mathbf{x}) + \frac{\sigma}{M'}\left(\sum_{s=2}^{n}(s-1)|x_{j_s}|\right) : \mathbf{x} \in \mathbb{R}^n\right\}. \tag{16}$$

The latter formula highlights that increasing weights are assigned to decreasing absolute values of the components of $\mathbf{x}$, i.e., the smaller is the absolute value the bigger is its weight; this somehow indicates a kind of preference towards reduction of the *small* components, which is, in turn, in favor of reduction of the ℓ_0-norm of $\mathbf{x}$.

Remark 1 Formula (16) clarifies the differences between our approach and the SLOPE method (see [5, 11]), the trimmed LASSO method (see [9]), the approach proposed in [28] and the ℓ_{1-2} method [51]. In SLOPE, a reverse ordering of the weights with respect to our model is adopted, as components $|x_{j_s}|$ are associated to non-increasing penalty weights. In fact, the resulting convex program is

$$\min\left\{f(\mathbf{x}) + \sum_{s=1}^{n} \lambda_s |x_{j_s}| : \mathbf{x} \in \mathbb{R}^n\right\}, \tag{17}$$

with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Given any $\sigma > 0$, a possible choice of parameters is

$$\lambda_s = \sigma \quad \forall s \in \{1, \ldots, t\} \qquad \text{and} \qquad \lambda_s = 0 \quad \forall s \in \{t+1, \ldots, n\}, \tag{18}$$

for some sparsity parameter $t \in \{1, \ldots, n-1\}$. Under such a choice problem (17) becomes the following convex penalized k-norm model:

$$\min \left\{ f(\mathbf{x}) + \sigma \|\mathbf{x}\|_{[t]} : \mathbf{x} \in \mathbb{R}^n \right\}. \tag{19}$$

In trimmed LASSO, given a sparsity parameter $t \in \{1, \ldots, n-1\}$, the following nonconvex problem is addressed

$$\min \left\{ f(\mathbf{x}) + \sigma \sum_{s=t+1}^{n} |x_{j_s}| : \mathbf{x} \in \mathbb{R}^n \right\} \tag{20}$$

where, unlike our approach, components of $\mathbf{x}$ with the highest absolute values are not penalized at all, whereas the $(n-t)$ smallest ones are equally penalized. As for the approach presented in [28], it can be proved that, for a given penalty parameter $\sigma > 0$, letting $J_\sigma(\mathbf{x}) \triangleq \left\{ j \in \{1, \ldots, n\} : x_j > \frac{1}{\sigma} \right\}$, the proposed DC (hence nonconvex) program can be formulated as

$$\min \left\{ f(\mathbf{x}) + |J_\sigma(\mathbf{x})| + \sigma \sum_{j \notin J_\sigma(\mathbf{x})} |x_j| : \mathbf{x} \in \mathbb{R}^n \right\}, \tag{21}$$

which tends to (SOP) as $\sigma \to \infty$. Finally, we remark the substantial difference between the method we propose and the ℓ_{1-2} approach [51], where the ℓ_0 pseudo-norm is approximated by the DC model $\|\mathbf{x}\|_1 - \|\mathbf{x}\|_2$.

A better insight into the differences between the two formulations (RSOP) and (Rk-SOP) can be gathered by appropriate setting of the constants M and M'. Since the former is referred to single vector components and the latter to norms, it is natural to set

$$M' = nM. \tag{22}$$

With such choice the objective function $\widetilde{F}(\cdot)$ of (RkSOP) becomes:

$$\widetilde{F}(\mathbf{x}) = f(\mathbf{x}) + \frac{\sigma}{M}\|\mathbf{x}\|_1 - \frac{\sigma}{nM}\sum_{k=1}^{n} \|\mathbf{x}\|_{[k]}. \tag{23}$$

We observe that function $\widetilde{F}(\cdot)$ is the difference of two convex functions, namely,

$$\widetilde{F}(\mathbf{x}) = f_1(\mathbf{x}) - f_2(\mathbf{x}), \tag{24}$$

with

$$f_1(\mathbf{x}) \triangleq f(\mathbf{x}) + \frac{\sigma}{M}\|\mathbf{x}\|_1 \tag{25}$$

and

$$f_2(\mathbf{x}) = \frac{\sigma}{nM} \sum_{k=1}^{n} \|\mathbf{x}\|_{[k]}. \tag{26}$$

As a consequence, (RkSOP) can be handled by resorting to both the theoretical and the algorithmic tools of DC programming (see, e.g., [2, 33, 44]). Moreover, we observe that f_1 is exactly the objective function of (RSOP), thus it is

$$\overline{F}(\mathbf{x}) - \widetilde{F}(\mathbf{x}) = f_2(\mathbf{x}) \geq 0.$$

Summing up, $\overline{F}(\mathbf{x})$ is convex and majorizes the (nonconvex) $\widetilde{F}(\mathbf{x})$, with $f_2(\mathbf{x})$ being the nonnegative gap, whose value depends on the sum of the k-norms.

4 Properties of problem (RkSOP) and its solutions

We start by analyzing some properties of function $f_2(\cdot)$. First, taking into account (14) and (26), we rewrite f_2 as

$$\begin{aligned} f_2(\mathbf{x}) &= \frac{\sigma}{M}\left(|x_{j_1}| + \left(1 - \frac{1}{n}\right)|x_{j_2}| + \left(1 - \frac{2}{n}\right)|x_{j_3}| + \ldots + \frac{1}{n}|x_{j_n}|\right) \\ &= \frac{\sigma}{M} \sum_{s=1}^{n} c_s |x_{j_s}|, \end{aligned} \tag{27}$$

where $c_s = 1 - \frac{(s-1)}{n}$, for every $s \in \{1, \ldots, n\}$, with $c_1 > c_2 > \cdots > c_n$. Next, in order to study the behavior of $f_2(\cdot)$ at equal ℓ_1-norm value, we consider any ball $B(\mathbf{0}, \rho) \triangleq \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_1 = \rho\}$, centered at the origin with radius $\rho > 0$, and we focus on the following two optimization problems ($P^{\max}$) and ($P^{\min}$), respectively.

$$f_2^{(\max)} = \max\left\{f_2(\mathbf{x}) : \mathbf{x} \in B(\mathbf{0}, \rho)\right\} \tag{$P^{\max}$}$$

and

$$f_2^{(\min)} = \min\left\{f_2(\mathbf{x}) : \mathbf{x} \in B(\mathbf{0}, \rho)\right\}. \tag{$P^{\min}$}$$

We observe that, denoting by Π the set of all permutations of $\{1, \ldots, n\}$, both problems can be decomposed in several problems of the type

$$\max\left\{f_2(\mathbf{x}) : \mathbf{x} \in B(\mathbf{0}, \rho),\ |x_{\pi_s}| \geq |x_{\pi_{s+1}}|\ \ \forall s \in \{1, \ldots, n-1\}\right\} \tag{$P_\pi^{\max}$}$$

and

$$\min\left\{f_2(\mathbf{x}) : \mathbf{x} \in B(\mathbf{0}, \rho),\ |x_{\pi_s}| \geq |x_{\pi_{s+1}}|\ \ \forall s \in \{1, \ldots, n-1\}\right\} \tag{$P_\pi^{\min}$}$$

respectively, for every $\pi \in \Pi$. We remark that ($P_\pi^{\max}$) and ($P_\pi^{\min}$) are nonconvex programs, due to the nonconvex ordering constraints $|x_{\pi_s}| \geq |x_{\pi_{s+1}}|$, and we denote their global solutions by $\mathbf{x}_\pi^{(\max)}$ and $\mathbf{x}_\pi^{(\min)}$, respectively.

Proposition 4.1 *For any given $\pi \in \Pi$, the following properties regarding optimal solutions of ($P_\pi^{\max}$) and ($P_\pi^{\min}$) hold:*

(i) *Problem ($P_\pi^{\max}$) has two maximizers $x_{\pi_1}^{(\max)} = \pm\rho$, $x_{\pi_2}^{(\max)} = \cdots = x_{\pi_n}^{(\max)} = 0$. They are both global, with objective function value $f_2(\mathbf{x}^{(\max)}) = \frac{\rho\sigma}{M}$;*

(ii) *Problem (P^{min}) has 2^n global minimizers $|x_{\pi_1}^{(\min)}| = \cdots = |x_{\pi_n}^{(\min)}| = \frac{\rho}{n}$, with objective function value $f_2(\mathbf{x}^{(\min)}) = \frac{\rho\sigma}{M}\left(\frac{n+1}{2n}\right)$.*

Proof (i) The property follows by observing that the solutions $|x_{\pi_1}^{(\max)}| = \rho$, $x_{\pi_2}^{(\max)} = x_{\pi_3}^{(\max)} = \cdots = x_{\pi_n}^{(\max)} = 0$, due to (27), are optimal for the relaxation obtained by eliminating the ordering constraints from ($P_\pi^{\max}$), and that such solutions are also feasible for ($P_\pi^{\max}$).

(ii) We observe first that the monotonically decreasing structure of the cost coefficients c_s guarantees that at the optimum $|x_{\pi_s}^{(\min)}| > |x_{\pi_{s+1}}^{(\min)}|$ cannot occur for any index $s \in \{1, \ldots, (n-1)\}$. Consequently, it comes out from satisfaction of the ordering constraints in ($P_\pi^{\min}$), and from $\mathbf{x}^{(\min)} \in B(\mathbf{0}, \rho)$, that the optimal solutions satisfy $|x_{\pi_1}^{(\min)}| = \cdots = |x_{\pi_n}^{(\min)}| = \frac{\rho}{n}$ and the property follows.

□

Remark 2 The optimal solutions of ($P_\pi^{\max}$) depend on π, in particular on the choice of the index π_1, but they have all the same optimal value. Thus problem ($P^{\max}$) has a total number of $2n$ global solutions, with $f_2^{\max} = \frac{\rho\sigma}{M}$. As for problem ($P_\pi^{\min}$), the 2^n optimal solutions are independent of π, hence they are also global solution for ($P^{\min}$), with $f_2^{(\min)} = \frac{\rho\sigma}{M}\left(\frac{n+1}{2n}\right)$. As a consequence, the variation $\Delta(\rho)$ of f_2 on $B(\mathbf{0}, \rho)$ is

$$\Delta(\rho) = f_2^{max} - f_2^{min} = \rho\frac{\sigma}{M} - \rho\frac{\sigma}{M}\left(\frac{n+1}{2n}\right) = \rho\frac{\sigma}{M}\left(\frac{n-1}{2n}\right).$$

Remark 3 The ℓ_0-norm of the optimal solutions of problem ($P^{\max}$) is equal to 1, while the ℓ_0-norm of those of ($P^{\min}$) is equal to n. As a consequence, the (subtractive) gap f_2, for a fixed value of the ℓ_1 norm, is maximal when $\|\mathbf{x}\|_0 = 1$ and it is minimal when all components are nonzero and equal in modulus, that is $\|\mathbf{x}\|_0 = n$; in other words, model (RkSOP) exhibits a stronger bias towards reduction of the ℓ_0-norm than (RSOP). Of course, the price to be paid to obtain such an advantage is the need of solving the DC (global) optimization problem (RkSOP) instead of the convex program (RSOP).

An additional insight into properties of the solutions of (RkSOP) can be obtained by focusing on the related necessary conditions for global optimality. In particular, at any global minimizer $\widetilde{\mathbf{x}}$, taking into account the DC decomposition (24), the inclusion

$$\partial f_2(\widetilde{\mathbf{x}}) \subset \partial f_1(\widetilde{\mathbf{x}}) \tag{28}$$

is satisfied, see [34]. Before introducing a property of minimizers in terms of their ℓ_0 norm, we recall some differential properties of the k-norm, see [31]. The subdifferential at any point $\mathbf{x} \in \mathbb{R}^n$ is

$$\partial \|\mathbf{x}\|_{[k]} = \left\{ \mathbf{g}^{[k]} \in \mathbb{R}^n : \mathbf{g}^{[k]} = \arg\max \left\{ \mathbf{x}^\top \mathbf{w} : \|\mathbf{w}\|_1 = k, \ \|\mathbf{w}\|_\infty \leq 1 \right\} \right\}. \quad (29)$$

In particular, given any $\overline{\mathbf{x}} \in \mathbb{R}^n$, and denoting by $J_{[k]}(\overline{\mathbf{x}}) \triangleq \{j_1, \ldots, j_k\}$ the index set of k largest absolute-value components of $\overline{\mathbf{x}}$, a subgradient $\overline{\mathbf{g}}^{[k]} \in \partial \|\overline{\mathbf{x}}\|_{[k]}$ can be obtained as

$$\overline{g}_j^{[k]} = \begin{cases} 1 & \text{if} \quad j \in J_{[k]}(\overline{\mathbf{x}}) \text{ and } \overline{x}_j \geq 0, \\ -1 & \text{if} \quad j \in J_{[k]}(\overline{\mathbf{x}}) \text{ and } \overline{x}_j < 0, \\ 0 & \text{otherwise.} \end{cases} \quad (30)$$

Moreover, we note that the subdifferential $\partial \|\cdot\|_{[k]}$ is a singleton (i.e., the vector k-norm is differentiable) any time the set $J_{[k]}(\cdot)$ is uniquely defined.

As for the subdifferential $\partial \|\cdot\|_1$ of the ℓ_1 norm, we recall that its element are the subgradients $\overline{\mathbf{g}}^{(1)}$ at $\overline{\mathbf{x}}$ whose jth component, for every $j \in \{1, \ldots, n\}$, is obtained as

$$\overline{g}_j^{(1)} = \begin{cases} 1 & \text{if} \quad \overline{x}_j > 0 \\ -1 & \text{if} \quad \overline{x}_j < 0 \\ \alpha_j & \text{if } \overline{x}_j = 0 \end{cases} \quad (31)$$

where $\alpha_j \in [-1, 1]$. Summing up, we get

$$\partial f_1(\mathbf{x}) = \partial f(\mathbf{x}) + \frac{\sigma}{M} \partial \|\mathbf{x}\|_1$$

and

$$\partial f_2(\mathbf{x}) = \frac{\sigma}{nM} \sum_{k=1}^{n} \partial \|\mathbf{x}\|_{[k]}.$$

The following proposition provides a quantitative evaluation of the effect of the penalty-parameter σ on the ℓ_0-norm of a global minimizer.

Proposition 4.2 *Let $\widetilde{\mathbf{x}}$ denote a global minimizer of the k-norm relaxation (RkSOP), and L denote the Lipschitz constant of f. Then, the following inequality holds at $\widetilde{\mathbf{x}}$:*

$$\|\widetilde{\mathbf{x}}\|_0 \leq \frac{nML}{\sigma} + 1.$$

Proof Satisfaction of the inclusion (28) at $\widetilde{\mathbf{x}}$ implies that, for each subgradient $\widetilde{\boldsymbol{\xi}}^{(2)} \in \partial f_2(\widetilde{\mathbf{x}})$ there exists a couple of subgradients $\widetilde{\boldsymbol{\xi}} \in \partial f(\widetilde{\mathbf{x}})$ and $\widetilde{\boldsymbol{\xi}}^{(1)} \in \frac{\sigma}{M} \partial \|\mathbf{x}\|_1$ such that

$$\widetilde{\boldsymbol{\xi}}^{(2)} = \widetilde{\boldsymbol{\xi}} + \widetilde{\boldsymbol{\xi}}^{(1)} \quad (32)$$

Now suppose $\|\widetilde{\mathbf{x}}\|_0 = r$, for some $r \in \{1, \ldots, n\}$ and, w.l.o.g., assume that

$$|\widetilde{x}_1| \geq |\widetilde{x}_2| \geq \ldots |\widetilde{x}_r| > 0$$

with the remaining components, if any, equal to zero, that is $J_{[k]}(\widetilde{\mathbf{x}}) = \{1, 2, \ldots, k\}$, for every $k \in \{1, \ldots, n\}$. Now pick, for $k = 1, \ldots, n$, a subgradient in $\partial \|\widetilde{\mathbf{x}}\|_{[k]}$ defined as in (30) and calculate $\widetilde{\boldsymbol{\xi}}^{(2)} \in \partial f_2(\widetilde{\mathbf{x}})$ as

$$\widetilde{\boldsymbol{\xi}}^{(2)} = \frac{\sigma}{nM} \sum_{k=1}^{n} \widetilde{\mathbf{g}}^{[k]}.$$

By simple calculation we get

$$\widetilde{\xi}_j^{(2)} = \frac{\sigma}{nM}(n - j + 1)s(\widetilde{x}_j), \qquad \forall j \in \{1, \ldots, n\},$$

where

$$s(\widetilde{x}_j) = \begin{cases} 1 & \text{if} \quad \widetilde{x}_j \geq 0, \\ -1 & \text{if} \quad \widetilde{x}_j < 0. \end{cases}$$

By a component-wise rewriting of condition (32), and taking into account only the components $j \in \{1, \ldots, r\}$ such that $\widetilde{x}_j \neq 0$, from (31) we obtain that at the minimizer $\widetilde{\mathbf{x}}$, for some $\widetilde{\boldsymbol{\xi}} \in \partial f(\widetilde{\mathbf{x}})$ it is

$$\widetilde{\xi}_j = \frac{\sigma}{nM}(1 - j)s(\widetilde{x}_j), \ \ \forall j \in \{1, \ldots, r\},$$

that is

$$|\widetilde{\xi}_j| = \frac{\sigma}{nM}(j - 1), \ \ \forall j \in \{1, \ldots, r\}. \tag{33}$$

Then, taking into account $|\widetilde{\xi}_j| \leq L$, the thesis follows by observing that condition (33) implies

$$r \leq \frac{nML}{\sigma} + 1.$$

□

Remark 4 Proposition 4.2 suggests a quantitative way to control sparsity of the solution acting on parameter σ. It is worth pointing out, however, that the bound holds only if a *global* minimizer is available. This is not necessarily the case if a local optimization algorithm is adopted, as we do in Sect. 6.

5 Feature selection in support vector machine (SVM)

In the SVM framework for binary classification two (labeled) point-sets $\mathcal{A} \triangleq \{\mathbf{a}_1, \ldots, \mathbf{a}_{m_1}\}$ and $\mathcal{B} \triangleq \{\mathbf{b}_1, \ldots, \mathbf{b}_{m_2}\}$ in $\mathbb{R}^n$ are given, the objective being to find a hyperplane, associated to a couple $(\mathbf{x}, \gamma) \in \mathbb{R}^n \times \mathbb{R}$, strictly separating them. Thus, it is required that the following inequalities hold true:

$$\mathbf{a}_i^\top \mathbf{x} \leq \gamma - 1, \qquad \forall i \in \{1, \ldots, m_1\}, \tag{34}$$

$$\mathbf{b}_l^\top \mathbf{x} \geq \gamma + 1, \qquad \forall l \in \{1, \ldots, m_2\}. \tag{35}$$

The existence of such a hyperplane is ensured if and only if $\text{conv}\mathcal{A} \cap \text{conv}\mathcal{B} = \emptyset$, a property hard to be checked. A convex, piecewise linear and nonnegative error function of $(\mathbf{x}, \gamma)$ is thus defined. It has the form

$$e(\mathbf{x}, \gamma) = \sum_{i=1}^{m_1} \max\{0, \mathbf{a}_i^\top \mathbf{x} - \gamma + 1\} + \sum_{l=1}^{m_2} \max\{0, -\mathbf{b}_l^\top \mathbf{x} + \gamma + 1\}, \tag{36}$$

being equal to zero if and only if $(\mathbf{x}, \gamma)$ actually defines a (strictly) separating hyperplane satisfying (34–35).

In the SVM approach the following convex problem

$$\min \left\{ Ce(\mathbf{x}, \gamma) + \|\mathbf{x}\| : \mathbf{x} \in \mathbb{R}^n,\ \gamma \in \mathbb{R} \right\}, \tag{37}$$

is solved, where the norm of $\mathbf{x}$ is added to the error function aiming to obtain a maximum-margin separation, C being a positive trade-off parameter.

The ℓ_1 and ℓ_2 norms are in general adopted in model (37), but, in case feature selection is pursued, the ℓ_0-norm looks as the most suitable tool, although the ℓ_1-norm is usually considered as a good approximation. In the following, we will focus on the feature selection ℓ_0-norm problem

$$\min \left\{ Ce(\mathbf{x}, \gamma) + \|\mathbf{x}\|_0 : \mathbf{x} \in \mathbb{R}^n,\ \gamma \in \mathbb{R} \right\}, \tag{38}$$

by applying a relaxed model of the type (RkSOP) and the relative machinery. Note that, to keep the notation as close as possible to that commonly adopted in the SVM framework, the trade-off parameter C, unlike the classical formulation of (SOP), where parameter σ is present, is now equivalently put in front of the error term $e(\mathbf{x}, \gamma)$.

Our MINLP formulation of problem of (38), along the guidelines of the (MIkSOP) model (10–12), letting $M = 1$ and $M' = n$ (see (22)), is

$$\min \quad Ce(\mathbf{x}, \gamma) + \mathbf{e}^\top \mathbf{y} \tag{39}$$

$$\text{s.t.} \quad \|\mathbf{x}\|_1 - \|\mathbf{x}\|_{[k]} \leq n y_k \qquad \forall k \in \{1, \ldots, n\} \tag{40}$$

$$\mathbf{y} \in \{0, 1\}^n \tag{41}$$

By adopting the same relaxation scheme described in §3, we obtain the DC continuous program (SVM-RkSOP)

$$\min\left\{Ce(\mathbf{x},\gamma)+\|\mathbf{x}\|_1-\frac{1}{n}\sum_{k=1}^{n}\|\mathbf{x}\|_{[k]}:\ \mathbf{x}\in\mathbb{R}^n,\ \gamma\in\mathbb{R}\right\},\qquad\text{(SVM-R}k\text{SOP)}$$

an SVM model, tailored for feature selection, as it exploits the continuous relaxation of (MISOP), whose computational behavior will be analyzed in the next section.

6 Computational experience

We have evaluated the computational behavior of our SVM-based feature selection model (SVM-RkSOP) by testing it on 8 well known datasets whose relevant details are listed in Table 1. In particular, we remark that such datasets can be partitioned in two groups depending on the relative proportion between the number of points (m_1+m_2) and the number of features (n). From this viewpoint, datasets 1 to 4 have a large number of points compared to the small number of features, whilst datasets 5 to 8 have a large number of features compared to the small number of points.

We recall that (SVM-RkSOP) is a DC optimization problem, analogous to (24), where

$$f_1(\mathbf{x},\gamma)\triangleq C\sum_{i=1}^{m_1}\max\{0,\mathbf{a}_i^\top\mathbf{x}-\gamma+1\}+C\sum_{l=1}^{m_2}\max\{0,-\mathbf{b}_l^\top\mathbf{x}+\gamma+1\}+\|\mathbf{x}\|_1 \quad (42)$$

and

$$f_2(\mathbf{x},\gamma)\triangleq\frac{1}{n}\sum_{k=1}^{n}\|\mathbf{x}\|_{[k]}, \quad (43)$$

that can be tackled by adopting techniques, like those described in [3, 27, 36], which require to calculate at each iterate-point the linearization of function $f_2(\cdot)$. In fact, such linearization can be easily calculated by applying (30) at any point $\overline{\mathbf{x}}\in\mathbb{R}^n$

Table 1 Details of datasets

#	Name	References	m_1+m_2	n
1	Breast-cancer	[18]	683	10
2	Diabetes	[18]	768	8
3	Heart	[18]	270	13
4	Ionosphere	[18]	351	34
5	Brain_Tumor1	[12]	60	7129
6	Brain_Tumor2	[12]	50	12625
7	DLBCL	[12]	77	7129
8	Leukemia/ALLAML	[12]	72	5327

to a get a subgradient $\overline{\mathbf{g}}^{[k]} \in \partial\|\overline{\mathbf{x}}\|_{[k]}$. In particular, we have implemented a version of the DCA algorithm, see [2], exploiting the fact that at an iterate point $\overline{\mathbf{x}}$, once obtained a linearization $h_2(\overline{\mathbf{x}}, \cdot)$ of $f_2(\cdot)$ at $\overline{\mathbf{x}}$, the new iterate-point can be calculated by minimizing the function

$$C\sum_{i=1}^{m_1}\max\{0, \mathbf{a}_i^\top\mathbf{x} - \gamma + 1\} + C\sum_{l=1}^{m_2}\max\{0, -\mathbf{b}_l^\top\mathbf{x} + \gamma + 1\} + \|\mathbf{x}\|_1 - h_2(\overline{\mathbf{x}}, \mathbf{x})$$

that can be easily turned into an equivalent linear program. Hence, applying DCA to (SVM-RkSOP) amounts to solving a sequence of linear programs as long as a decrease of the objective function is obtained. In order to evaluate the effectiveness of (SVM-RkSOP) solved by DCA, we will also make some comparison against the behavior of the convex standard ℓ_1-norm-based SVM model (SVM-RSOP)

$$\min\left\{Ce(\mathbf{x}, \gamma) + \|\mathbf{x}\|_1 : \mathbf{x} \in \mathbb{R}^n,\ \gamma \in \mathbb{R}\right\} \qquad \text{(SVM-RSOP)}$$

whose solution can be easily obtained by solving just one equivalent linear program, next referred to as LP-SVM-RSOP. The experimental plan for both approaches is based on a two-level cross-validation protocol to tune parameter C and to train the classifier. In fact, the tenfold cross-validation approach has been adopted at the higher level to train the classifier, every dataset being randomly partitioned into 10 groups of equal size. Then, 10 different blocks (the training sets) are built, each containing 9 out of 10 groups. Every block is used to train the classifier, using the left out group as the testing-set that returns the percentage of correctly classified bags (test correctness). Before executing the training phase, at the lower level it is needed to tune parameter C. A grid of 20 values, ranging between 10^{-3} and 10^2, has been selected, and a five-fold cross-validation approach has been adopted on each training set (the model-selection phase). For each training set, we choose the C value as the one returning the highest average test-correctness. As for the selection of an appropriate starting point for DCA applied to (SVM-RkSOP), next referred to as DCA-SVM-RkSOP, denoting by $\mathbf{x}_a$ the barycenter of all the $\mathbf{a}_i$ instances, and by $\mathbf{x}_b$ the barycenter of all the $\mathbf{b}_l$ instances, we have selected the starting point $(\mathbf{x}_0, \gamma_0)$ by setting

$$\mathbf{x}_0 = \mathbf{x}_a - \mathbf{x}_b \tag{44}$$

and choosing γ_0 such that all the $\mathbf{b}_l$ instances are well classified. We have implemented the DCA-SVM-RkSOP algorithm [26] in Python 3.6, and run the computational experiments on a 2.80 GHz Intel(R) Core(TM) i7 computer. The LP solver of IBM ILOG CPLEX 12.8 [35] has been used to solve linear programs. The numerical results regarding DCA-SVM-RkSOP and LP-SVM-RSOP are reported in Table 2 and Table 3, respectively, where we list the percentage correctness averaged over the 10 folds of both the testing and the training phases. Moreover, we report the values ft0, ft-2, ft-4, and ft-9, representing the percentage average of features for which the corresponding component of the minimizer $\mathbf{x}^*$ is larger than 1, 10^{-2} , 10^{-4} , 10^{-9}, respectively.

Table 2 Numerical results: DCA-SVM-R*k*SOP

Name	Test (%)	Train (%)	ft0 (%)	ft-2 (%)	ft-4 (%)	ft-9 (%)	cpu (s)
Breast-Cancer	96.78	97.23	16.00	76.00	76.00	76.00	0.35
Diabetes	76.96	77.52	27.50	86.25	87.50	87.50	0.42
Heart	83.33	85.84	10.00	80.00	80.00	80.77	0.12
Ionosphere	86.05	93.28	31.76	56.47	56.47	56.47	0.20
Brain_Tumor1	63.62	68.40	0.010	0.015	0.015	0.015	3.91
Brain_Tumor2	80.67	97.57	0.036	0.051	0.051	0.051	11.96
DLBCL	92.50	100.00	0.050	0.085	0.085	0.085	9.86
Leukemia/ALLAML	93.81	100.00	0.071	0.090	0.090	0.090	6.17

Table 3 Numerical results: `LP-SVM-RSOP`

Name	Test (%)	Train (%)	ft0 (%)	ft-2 (%)	ft-4 (%)	ft-9 (%)	cpu (s)
Breast-Cancer	96.63	97.23	8.00	86.00	86.00	86.00	0.30
Diabetes	76.83	77.52	25.00	91.25	92.50	92.50	0.40
Heart	84.07	85.05	2.31	85.38	86.92	86.92	0.07
Ionosphere	87.49	93.32	28.53	69.71	70.88	70.88	0.12
Brain_Tumor1	58.38	77.41	0.000	0.192	0.205	0.206	1.49
Brain_Tumor2	82.33	96.02	0.000	0.181	0.188	0.188	2.08
DLBCL	96.25	99.71	0.000	0.397	0.411	0.411	2.71
Leukemia/ALLAML	95.42	99.69	0.000	0.593	0.629	0.629	1.74

Hence, small values of ft-9 denote high sparsity of $\mathbf{x}^*$. Finally we also report the cpu time (measured in seconds) regarding the execution time of DCA-SVM-RkSOP and LP-SVM-RSOP in the training phase, averaged over the 10 training folds.

The results clearly show the benefits of the regularization model obtained by exploiting vector-k-norms. In fact, for all datasets DCA-SVM-RkSOP returns average test-correctness never significantly worse than LP-SVM-RSOP, managing to improve on sparsity levels. If the latter outcome is apparent although not particularly relevant for small-size problems, it turns out very significant for large-size problems for which sparsity increases of an order of magnitude. Take, for an example, the Leukemia/ALLAML dataset, having 5327 features: while LP-SVM-RSOP manages to get an average test-correctness of 95.42% using around 33 features, DCA-SVM-RkSOP returns an average test-correctness of 93.81% using only 5 features. As expected, such improvement is obtained at the expenses of an increased computational time.

Finally, to evaluate (SVM-RkSOP) not only against the ℓ_1-norm-based SVM model (SVM-RSOP), we have implemented the convex k-norm-based model (19), that, as previously remarked, can be seen as a special case of the SLOPE approach (17), based on the weight setting (18) for some $t \in \{1, \ldots, n-1\}$. In fact we have defined the (SVM-kPURE) model as follows:

$$\min \left\{ e(\mathbf{x}, \gamma) + \sigma \|\mathbf{x}\|_{[t]} : \mathbf{x} \in \mathbb{R}^n,\ \gamma \in \mathbb{R} \right\}. \qquad \text{(SVM-}k\text{PURE)}$$

To handle such problem we have used the solver NCVX, a general-purpose code for nonsmooth optimization described in [25].

We have adopted different experimentation strategies for datasets 1–4 and 5–8. In particular for the first group we have set $t = 0.6n$ (with possible rounding) and have tested $\sigma = 1$ and $\sigma = 100$. The results are in Tables 4 and 5, respectively. As for datasets 5–8, characterized by large values of n, we have tested two different settings of t. The first one is the same as for datasets 1–4, that is $t = 0.6n$. The second one comes from the results obtained by (SVM-RSOP). In fact, letting $t_{\text{svm}} = n(\text{ft-9})/100$, where ft-9 is available in the bottom half part of Table 3, we have selected $t = t_{\text{svm}}$ as the average number of significant components of $\mathbf{w}$ at the solution of (SVM-RSOP). Since we have observed that the results are rather insensitive to the parameter σ, we report in Tables 6 and 7 only the ones obtained for $\sigma = 1$.

The results demonstrate that the simplistic use of the k-norm in the convex setting provides poor performance, definitely worse even than the ℓ_1-norm model. Taking in fact ft-9 as a measure of sparsity, we observe that on datasets of the first group (Tables 4 and 5) no significant sparsity is enforced for $\sigma = 1$, while for $\sigma = 100$ the results for the two datasets Diabetes and Heart are sparse, at the expenses of a severe reduction of the classification correctness. As for the datasets of the second group, for both choices of t we observe (Tables 6 and 7) that the solutions are not at all sparse, while the large difference between training and testing correctness indicates the presence of overfitting.

We remark that in our experiments we have not considered the comparison of our model with the plain MILP formulation (MISOP), as the superiority of the k-

Table 4 Numerical results: SVM-kPURE ($\sigma = 1, t = 0.6n$)

Name	Test (%)	Train (%)	ft0 (%)	ft-2 (%)	ft-4 (%)	ft-9 (%)	cpu (s)
Breast-Cancer	96.92	97.23	2.00	98.00	100.00	100.00	1.16
Diabetes	76.96	77.79	26.25	100.00	100.00	100.00	1.38
Heart	83.33	85.59	9.23	99.23	100.00	100.00	1.97
Ionosphere	88.30	93.63	20.29	96.76	97.05	97.05	17.45

Table 5 Numerical results: SVM-kPURE ($\sigma = 100, t = 0.6n$)

Name	Test (%)	Train (%)	ft0 (%)	ft-2 (%)	ft-4 (%)	ft-9 (%)	cpu (s)
Breast-Cancer	95.90	95.98	0.00	99.00	100.00	100.00	9.62
Diabetes	65.10	65.10	0.00	0.00	0.00	60.00	0.79
Heart	73.33	74.21	0.00	7.69	7.69	43.84	1.38
Ionosphere	64.11	64.10	0.00	0.00	0.00	97.05	3.31

Table 6 Numerical results: SVM-kPURE ($\sigma = 1, t = t_{\text{svm}}$)

Name	Test (%)	Train (%)	ft0 (%)	ft-2 (%)	ft-4 (%)	ft-9 (%)	cpu (s)
Brain_Tumor1	48.38	100.00	0.00	59.90	99.58	100.00	77.89
Brain_Tumor2	68.17	100.00	0.0	40.91	99.33	100.00	173.19
DLBCL	93.33	100.00	0.00	37.63	99.26	100.00	64.02
Leukemia/ALLAML	91.13	100.00	0.00	46.75	99.36	100.00	42.51

Table 7 Numerical results: SVM-kPURE ($\sigma = 1, t = 0.6n$))

Name	Test (%)	Train (%)	ft0 (%)	ft-2 (%)	ft-4 (%)	ft-9 (%)	cpu (s)
Brain_Tumor1	47.17	97.78	0.00	44.35	99.57	100.00	66.57
Brain_Tumor2	57.67	100.00	0.00	25.65	99.37	100.00	176.21
DLBCL	86.91	99.71	0.00	24.08	99.32	100.00	70.93
Leukemia/ALLAML	90.06	100.00	0.00	25.60	99.29	10.00	39.23

norm based approaches has been elsewhere demonstrated (see [28, 29]), at least for problems of reasonable size. As for comparison with the k-norm based approach SVM_0 described in [28], it has to be taken into account that the results therein (see Tables 6 and 7) are strongly affected by the penalty parameter choice. Nevertheless, as far as test-correctness is considered, from comparison of Table 2 with [28, Table 5], we observe that each of the two approaches SVM_0 and DCA-SVM-RkSOP prevails on four of the eight datasets considered in both papers. As for sparsity of the solutions, the behavior is comparable on the data sets of the second group, while SVM_0 exhibits stronger sparsity enforcement on the first group of datasets, somehow at expenses of test-correctness.

7 Conclusions

We have introduced a novel continuous nonconvex k-norm-based model for sparse optimization, which is derived as the continuous relaxation of a MINLP problem. We have applied such model to Feature Selection in the SVM setting and the results suggest the superiority of the proposed approach over other attempts to simulate the ℓ_0 pseudo-norm (e.g., the ℓ_1 norm penalization). On the other hand, we have also observed that the mere replacement of the ℓ_0 pseudo-norm with the k-norm, evoked in some methods available in the literature, does not provide satisfactory results, at least in the experimentation area considered in this paper.

Funding Open access funding provided by Universitá della Calabria within the CRUI-CARE Agreement.

Data availability The datasets analyzed during the current study are available in the following repository: https://github.com/GGiallombardo/DCA-SVM-RkSOP.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

1. Amaldi, E., Kann, V.: On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. Theoret. Comput. Sci. **209**(1–2), 237–260 (1998)
2. An, L.T.H., Tao, P.D.: The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. Ann. Oper. Res. **133**, 23–46 (2005)
3. An, L.T.H., Nguyen, V.V., Tao, P.D.: A DC programming approach for feature selection in support vector machines learning. Adv. Data Anal. Classif. **2**, 259–278 (2008)
4. Beck, A., Eldar, Y.C.: Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. SIAM J. Optim. **23**(3), 1480–1509 (2013)
5. Bellec, P.C., Lecué, G., Tsybakov, A.B.: Slope meets lasso: Improved oracle bounds and optimality. Ann. Stat. **46**(6B), 3603–3642 (2018)
6. Bertolazzi, P., Felici, G., Festa, P., Fiscon, G., Weitschek, E.: Integer programming models for feature selection: New extensions and a randomized solution algorithm. Eur. J. Oper. Res. **250**(2), 389–399 (2016)
7. Bertsimas, D., King, A., Mazumder, R., et al.: Best Subset Selection via a Modern Optimization Lens. Ann. Stat. **44**(2), 813–852 (2016)
8. Bertsimas, D., King, A.: Logistic regression: from art to science. Stat. Sci. **32**(3), 367–384 (2017)
9. Bertsimas, D., Copenhaver, M.S., Mazumder, R.: The trimmed Lasso: sparsity and robustness. arXiv preprint (2017b) https://arxiv.org/pdf/1708.04527.pdf
10. Bienstock, D.: Computational study of a family of mixed-integer quadratic programming problems. Math. Programm. Ser. B Part A **74**(2), 121–140 (1996)

11. Bogdan, M., van den Berg, E., Sabatti, C., Su, W., Candès, E.J.: Slope-adaptive variable selection via convex optimization. Ann. Appl. Stat. **9**(3), 1103–1140 (2015)
12. Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J.M., Herrera, F.: A review of microarray datasets and applied feature selection methods. Inf. Sci. **282**, 111–135 (2014)
13. Bradley, P.S., Mangasarian, O.L.: Feature selection via concave minimization and support vector machines. In: Machine Learning proceedings of the fifteenth international conference (ICML '98). Shavlik J editor, Morgan Kaufmann, San Francisco, California, 82–90 (1998)
14. Bradley, P.S., Mangasarian, O.L., Street, W.N.: Feature selection via mathematical programming. INFORMS J. Comput. **10**(2), 209–217 (1998)
15. Burdakov, O.P., Kanzow, C., Schwartz, A.: Mathematical programs with cardinality constraints: Reformulation by complementarity-type conditions and a regularization method. SIAM J. Optim. **26**(1), 397–425 (2016)
16. Candés, E.J., Romberg, J.K., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**, 1207–1223 (2006)
17. Candés, E.J., Tao, T.: Decoding by linear programming. IEEE Trans. Inf. Theory **51**, 4203–4215 (2005)
18. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**(27), 1–27 (2011)
19. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press (2000)
20. Dedieu, A., Hazimeh, H., Mazumder, R.: Learning sparse classifiers: continuous and mixed integer optimization perspectives. (2020) arXiv preprint https://arxiv.org/pdf/2001.06471.pdf
21. Donoho, D.L.: Compressed sensing. IEEE Trans. Inf. Theory **52**, 1289–1306 (2006)
22. Dy, J.G., Brodley, C.E., Wrobel, S.: Feature selection for unsupervised learning. J. Mach. Learn. Res. **5**, 845–889 (2004)
23. Fan, J.Q., Li, R.Z.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**(456), 1348–1360 (2001)
24. Feng, M., Mitchell, J.E., Pang, J.-S., Shen, X., Wächter, A.: Complementarity formulations of ℓ_0-norm optimization problems. Pac. J. Optim. **14**(2), 273–305 (2018)
25. Fuduli, A., Gaudioso, M., Giallombardo, G.: Minimizing nonconvex nonsmooth functions via cutting planes and proximity control. SIAM J. Optim. **14**(3), 743–756 (2004)
26. Gaudioso, M., Giallombardo, G., Miglionico, G.: The DCA-SVM-RkSOP approach (2023) https://github.com/GGiallombardo/DCA-SVM-RkSOP
27. Gaudioso, M., Giallombardo, G., Miglionico, G., Bagirov, A.M.: Minimizing nonsmooth DC functions via successive DC piecewise-affine approximations. J. Global Optim. **71**(1), 37–55 (2018)
28. Gaudioso, M., Gorgone, E., Hiriart-Urruty, J.B.: Feature selection in SVM via polyhedral k-norm. Optim. Lett. **14**, 19–36 (2020)
29. Gaudioso, M., Gorgone, E., Labbé, M., Rodríguez-Chía, A.M.: Lagrangian relaxation for SVM feature selection. Comput. Oper. Res. **87**, 137–145 (2017)
30. Gaudioso, M., Hiriart-Urruty, J.-B.: Deforming $\|\cdot\|_1$ into $\|\cdot\|_\infty$ via polyhedral norms: a pedestrian approach. SIAM Rev. **64**(3), 713–727 (2022)
31. Gotoh, J., Takeda, A., Tono, K.: DC formulations and algorithms for sparse optimization problems. Math. Programm. Ser. B **169**(1), 141–176 (2018)
32. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
33. Hiriart-Urruty, J.-B.: Generalized differentiability/duality and optimization for problems dealing with differences of convex functions. In Convexity and duality in optimization. Lecture Notes in Economics and Mathematical Systems (1985)
34. Hiriart-Urruty, J.-B.: From convex optimization to nonconvex optimization: necessary and sufficient conditions for global optimality. In: Nonsmooth Optimization and Related Topics, pp. 219–240. Plenum, New York/London (1989)
35. IBM ILOG CPLEX 12.8 User Manual (2018) IBM Corp. Accessed 13 May 2023. https://www.ibm.com/docs/SSSA5P_12.8.0/ilog.odms.studio.help/pdf/usrcplex.pdf
36. Joki, K., Bagirov, A.M., Karmitsa, N., Mäkelä, M.M.: A proximal bundle method for nonsmooth DC optimization utilizing nonconvex cutting planes. J. Global Optim. **68**(3), 501–535 (2017)
37. Levato, T.: Algorithms for ℓ_0: norm optimization problems. Doctoral Dissertation, Dipartimento di Ingegneria dell'Informazione, Università di Firenze, Italia (2019)

38. Liu, Y.L., Bi, S.J., Pan, S.H.: Equivalent Lipschitz surrogates for zero-norm and rank optimization problems. J. Glob. Optim. **72**, 679–704 (2018)
39. Maldonado, S., Pérez, J., Weber, R., Labbé, M.: Feature selection for Support Vector Machines via Mixed Integer Linear Programming. Inf. Sci. **279**, 163–175 (2014)
40. Miao, W., Pan, S., Sun, D.: A Rank-Corrected Procedure for Matrix Completion with Fixed Basis Coefficients. Math. Program. **159**, 289–338 (2016)
41. Overton, M.L., Womersley, R.S.: Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices. Math. Program. **62**(1–3), 321–357 (1993)
42. Rinaldi, F., Schoen, F., Sciandrone, M.: Concave programming for minimizing the zero-norm over polyhedral sets. Comput. Optim. Appl. **46**, 467–486 (2010)
43. Sato, T., Takano, Y., Miyashiro, R., Yoshise, A.: Feature subset selection for logistic regression via mixed integer optimization. Comput. Optim. Appl. **64**(3), 865–880 (2016)
44. Strekalovsky, A.S.: Global optimality conditions for nonconvex optimization. J. Global Optim. **12**, 415–434 (1998)
45. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B **58**(1), 267–288 (1996)
46. Ustun, B., Rudin, C.: Supersparse linear integer models for optimized medical scoring systems. Mach. Learn. **102**, 349–391 (2016)
47. Vapnik, V.: The Nature of the Statistical Learning Theory. Springer (1995)
48. Watson, G.A.: Linear best approximation using a class of polyhedral norms. Numer. Algorithms **2**, 321–336 (1992)
49. Weston, J., Elisseeff, A., Schölkopf, B., Tipping, M.: Use of the zero-norm with linear models and kernel methods. J. Mach. Learn. Res. **3**, 1439–1461 (2003)
50. Wu, B., Ding, C., Sun, D., Toh, K.-C.: On the Moreau-Yosida regularization of the vector $k-$norm related functions. SIAM J. Optim. **24**(2), 766–794 (2014)
51. Yin, P., Lou, Y., He, Q., Xin, J.: Minimization of ℓ_{1-2} for compressed sensing. SIAM J. Sci. Comput. **37**(2), 536–563 (2015)
52. Zhang, C.H.: Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. **38**, 894–942 (2010)