# LAGRANGIAN RELAXATION FOR SVM FEATURE SELECTION

M. GAUDIOSO*, E. GORGONE†, M. LABBÉ†, AND A. M. RODRGUEZ-CHÍA‡

**Abstract.** We discuss a Lagrangian-relaxation-based heuristics for dealing with feature selection in a standard $L_1$ norm Support Vector Machine (SVM) framework for binary classification. The feature selection model we adopt is a Mixed Binary Linear Programming problem and it is suitable for a Lagrangian relaxation approach.

Based on a property of the optimal multiplier setting, we apply a consolidated nonsmooth optimization ascent algorithm to solve the resulting Lagrangian dual. In the proposed approach we get, at every ascent step, both a lower bound on the optimal solution as well as a feasible solution at low computational cost.

We present the results of our numerical experiments on some benchmark datasets characterized by a high number of features and a relatively small number of samples.

**Key words.** SVM classification, feature selection, Lagrangian relaxation, nonsmooth optimization

**1. Introduction.** The focus of pattern classification is to recognize similarities in the data, categorizing them in different subsets [3, 4, 15]. In many fields, such as the financial and the medical ones [11], the data (samples in classification language) require to be arranged in different classes.

Usually, samples have a small number of elements and these are characterized by a huge number of parameters (features). The handling of the entire feature set would be computationally very expensive and its outcome would lack from insight. For this reason, it is convenient to reduce the set of features which is expected to be easier to interpret and also easy to evaluate. However, it is not always easy to predict which of those are significant for classification purposes.

Hence it is necessary to screen off the significant features from those which are irrelevant.

The process that selects the features entering the subset of the relevant ones is known as *Feature Selection* (FS) and the related literature is extremely rich (see e.g. [10], [8], [13]). As far as mathematical programming-based approaches are concerned, we cite here [16] and, more specifically, [2], where (FS) is pursued by formulating a mathematical program with a parametric objective function which can be tackled in different ways. Yet another approach is presented in [14] where concave minimization is employed. In a recent paper [1] a model based on penalization of the number of features entering into the classification process is treated by means of generalized Benders decomposition.

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

†Département d'Informatique, Université Libre de Bruxelles, CP 212 Boulevard du Triomphe, B-1050 Brussels, Belgium. E-mail: egorgone@deis.unical.it, mlabbe@ulb.ac.be

‡Faculty of Sciences, Universidad de Cdiz Avenida Repblica Saharaui s/n, 11510. Puerto Real (Cdiz), Spain. E-mail: antonio.rodriguezchia@uca.es

The objective of this paper is to treat explicitly the (FS) problem as a Mixed Binary Linear Programming (MBLP) one, in the framework of the SVM (Support Vector Machine) [3, 15] approach. Consequently the objective of our model is three-fold: to minimize the classification error, to maximize the separation margin and to minimize the number features playing a role in the classification process.

The main novelty of our approach relies in the application of the Lagrangian Relaxation approach to our model, which is closely related to the one presented in [12], and in the use of the method described in [5] which belongs to the class of the bundle ones [9]. It implements an ascent procedure for solving the related Lagrangian Dual problem, which is, of course, a nonsmooth one [7]. A useful property of the proposed method is the possibility of getting, at each iteration of the ascent algorithm, a solution of the relaxed problem from which it is possible to get a feasible solution for the original problem at a quite low computational cost. This fact allows us to test goodness in terms of "primal" objective function of many solution as the ascent process goes.

We remark that the proposed relaxation satisfies the integrality property and consequently the bound obtained is as good as the one attained by using standard Linear Programming (LP) relaxation. Nonetheless we implement the dual ascent specifically with the aim at getting a good upper bound too.

The paper is organized as follows. In section 2 we present the model and the Lagrangian relaxation, introducing an appropriate decomposition technique. In section 3 we describe the method and in section 4 we discuss the numerical results obtained on some benchmark problems from cancer classification [13]. Some conclusions are finally drawn in section 5.

**2. The model.** 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 seek a hyperplane $(w, \gamma)$ that separates $\mathcal{A}$ and $\mathcal{B}$. We define the *classification error variables* $\xi_i$ and $\zeta_j$ which account, respectively, for the error related to point $i \in \mathcal{A}$ and $j \in \mathcal{B}$; moreover we introduce the binary *feature variable* vector $y \in \mathbb{R}^n$ with $y_k$ indicating whether or not feature $k$ is active, that is enters into calculation of the classifier. The model we propose may be considered as a variant of the one presented in [12], as in our approach the limitation in the number of features is pursued by means of an appropriate setting of the objective function, while in [12] budget-type constraints are in action. In particular we come out with the following Mixed Binary Programming (MBP) formulation of our SVM-feature-selection problem, which :

$$\min\ z^* = \min_{w,\gamma,\xi,\zeta,y} \|w\| + C(\sum_{i=1}^{m_1}\xi_i + \sum_{l=1}^{m_2}\zeta_l) + D\sum_{k=1}^{n} y_k \tag{2.1}$$

subject to

$$a_i^\top w + \gamma \le \xi_i - 1, \qquad i = 1,\dots,m_1 \tag{2.2}$$

$$-b_l^\top w - \gamma \le \zeta_l - 1 \qquad l = 1,\dots,m_2 \tag{2.3}$$

$$-u_k y_k \le w_k \le u_k y_k, \qquad k = 1,\dots,n \tag{2.4}$$

$$-u_k \le w_k \le u_k, \qquad k = 1,\dots,n \tag{2.5}$$

$$\xi_i \ge 0, \qquad i = 1,\dots,m_1 \tag{2.6}$$

$$\zeta_l \ge 0, \qquad l = 1,\dots,m_2 \tag{2.7}$$

$$y_k \in \{0,1\}, \qquad k = 1,\dots,n, \tag{2.8}$$

where $u_k > 0$ is the bound on the modulus of the $k$-th component of $w$. The objective function (2.1), we minimise, consists of the sum of three parts. In sequel there are (i) the norm of $w$ (we adopt, in particular, the $L_1$ norm), (ii) the classification error, (iii) the number of active features. The positive parameters $C$ and $D$ weight the different objectives. We remark that in the SVM approach minimization of $\|w\|$ corresponds to maximize the separation margin and that nonlinearity in the model can be easily eliminated by letting:

$$w_k = w_k^+ - w_k^-,\ \ w_k^+ \ge 0,\ \ w_k^- \ge 0,\ \ k = 1,\dots,n \text{ and } \|w\| = \sum_{k=1}^{n}(w_k^+ + w_k^-) \tag{2.9}$$

Note that constraints (2.5) are redundant, but we keep them as the approach we adopt is based on relaxation of the constraints (2.4).

In fact by introducing the multiplier vectors of appropriate dimension $\lambda \ge 0$ and $\mu \ge 0$ we obtain the following relaxation:

$$LR(\lambda,\mu) = \left| \begin{array}{l} z(\lambda,\mu) = \min\limits_{w,\gamma,\xi,\zeta,y} \|w\| + C\left(\sum\limits_{i=1}^{m_1}\xi_i + \sum\limits_{l=1}^{m_2}\zeta_l\right) + D\sum\limits_{k=1}^{n} y_k + \\ \qquad\qquad \sum\limits_{k=1}^{n}\lambda_k(w_k - u_k y_k) - \sum\limits_{k=1}^{n}\mu_k(w_k + u_k y_k) \\ \qquad\qquad \text{subject to } (2.2-2.3), (2.5-2.8)\,. \end{array} \right.$$

The objective function of the relaxed problem can be rearranged and consequently $LR(\lambda,\mu)$ is decomposed into two problems, $LR_1(\lambda,\mu)$ and $LR_2(\lambda,\mu)$ respectively, so that the first one involves the original variables $w, \gamma, \xi$ and $\zeta$ while the binary variables

$y_k$ are confined to the latter one. In details, one comes out with:

$$LR_1(\lambda,\mu) = \left| \begin{array}{ll} z_1(\lambda,\mu) = & \min\limits_{w,\gamma,\xi,\zeta} \|w\| + C\left(\sum\limits_{i=1}^{m_1}\xi_i + \sum\limits_{l=1}^{m_2}\zeta_l\right) + \sum\limits_{k=1}^{n}(\lambda_k - \mu_k)w_k \\ & \text{subject to } (2.2-2.3), (2.5-2.7)\,. \end{array} \right.$$

and

$$LR_2(\lambda,\mu) = \left| \begin{array}{ll} z_2(\lambda,\mu) = & \min\limits_{y} \sum\limits_{k=1}^{n}(D - u_k(\lambda_k + \mu_k))y_k \\ & \text{subject to } (2.8). \end{array} \right.$$

In the sequel we indicate by $(w(\lambda,\mu), \gamma(\lambda,\mu), \xi(\lambda,\mu), \zeta(\lambda,\mu), y(\lambda,\mu))$ the optimal solution to $LR(\lambda,\mu)$. Note that, since norm $L_1$ is adopted, $LR_1(\lambda,\mu)$ can be put in a standard $LP$ form thanks to (2.9), while $LR_2(\lambda,\mu)$ can be solved by simply inspecting the sign of the objective function coefficients. In fact we have:

$$y_k(\lambda,\mu) = \begin{cases} 1 & \text{if} \quad u_k(\lambda_k + \mu_k) > D \\ 0 \text{ or } 1 & \text{if} \quad u_k(\lambda_k + \mu_k) = D \\ 0 & \text{if} \quad u_k(\lambda_k + \mu_k) < D, \end{cases}$$

We indicate by $z_{LD}$ the optimal value of the Lagrangian dual, that is

$$z_{LD} = \max_{(\lambda,\mu)\geq 0} z(\lambda,\mu),$$

and remark that in our relaxation the integrality property holds.

We state the following proposition.

PROPOSITION 2.1. *There exists an optimal solution to the Lagrangian dual satisfying the condition*

$$u_k(\lambda_k + \mu_k) = D, \quad k = 1,\ldots,n. \tag{2.10}$$

*Proof.* Let $(\lambda^*, \mu^*)$ be any optimal solution to the Lagrangian dual, with

$$(w(\lambda^*,\mu^*), \gamma(\lambda^*,\mu^*), \xi(\lambda^*,\mu^*), \zeta(\lambda^*,\mu^*), y(\lambda^*,\mu^*))$$

being the corresponding optimal solution to $LR(\lambda^*,\mu^*)$. For short we will refer to such solution as to $(w^*, \gamma^*, \xi^*, \zeta^*, y^*)$.

We consider first the case that for some index $\bar{k}$ it is

$$u_{\bar{k}}(\lambda^*_{\bar{k}} + \mu^*_{\bar{k}}) - D > 0,$$

with of course $y^*_{\bar{k}} = 1$ and define a feasible solution $(\hat{\lambda}, \hat{\mu})$ for the Lagrangian dual as

$$\hat{\lambda}_k = \lambda^*_k, \quad \hat{\mu}_k = \mu^*_k \quad \text{for} \quad k \neq \bar{k},$$

and $\hat{\lambda}_{\bar{k}} \geq 0$, $\hat{\mu}_{\bar{k}} \geq 0$ satisfying the following condition

$$u_{\bar{k}}(\hat{\lambda}_{\bar{k}} + \hat{\mu}_{\bar{k}}) - D = 0. \tag{2.11}$$

The setting:

$$\begin{cases} \hat{\lambda}_{\bar{k}} = \min\{\lambda^*_{\bar{k}}, D/u_{\bar{k}}\} \\ \hat{\mu}_{\bar{k}} = D/u_{\bar{k}} - \min\{\lambda^*_{\bar{k}}, D/u_{\bar{k}}\}, \end{cases} \tag{2.12}$$

satisfies condition (2.11), moreover, letting

$$\Delta_{\bar{k}} \stackrel{\triangle}{=} (\lambda^*_{\bar{k}} + \mu^*_{\bar{k}}) - D/u_{\bar{k}} > 0,$$

we prove it is:

$$|(\hat{\lambda}_{\bar{k}} - \hat{\mu}_{\bar{k}}) - (\lambda^*_{\bar{k}} - \mu^*_{\bar{k}})| \leq \Delta_{\bar{k}}. \tag{2.13}$$

Consider in fact the two cases in (2.12):

- $\lambda^*_{\bar{k}} = \min\{\lambda^*_{\bar{k}}, D/u_{\bar{k}}\}$
  In this case it is $\hat{\lambda}_{\bar{k}} = \lambda^*_{\bar{k}}$ and $\hat{\mu}_{\bar{k}} = D/u_{\bar{k}} - \lambda^*_{\bar{k}}$. Thus it is

  $$|(\hat{\lambda}_{\bar{k}} - \hat{\mu}_{\bar{k}}) - (\lambda^*_{\bar{k}} - \mu^*_{\bar{k}})| = |\lambda^*_{\bar{k}} - D/u_{\bar{k}} + \lambda^*_{\bar{k}} - \lambda^*_{\bar{k}} + \mu^*_{\bar{k}})| = \Delta_{\bar{k}}$$

- $D/u_{\bar{k}} = \min\{\lambda^*_{\bar{k}}, D/u_{\bar{k}}\}$. Note that it is

  $$D/u_{\bar{k}} - \lambda^*_{\bar{k}} \leq 0 \tag{2.14}$$

  In this case it is $\hat{\lambda}_{\bar{k}} = D/u_{\bar{k}}$ and $\hat{\mu}_{\bar{k}} = 0$. Thus it is:

  $$|(\hat{\lambda}_{\bar{k}} - \hat{\mu}_{\bar{k}}) - (\lambda^*_{\bar{k}} - \mu^*_{\bar{k}})| = |D/u_{\bar{k}} - \lambda^*_{\bar{k}} + \mu^*_{\bar{k}})|.$$

  Taking into account (2.14) we obtain:

  $$\begin{aligned} -\Delta_{\bar{k}} = &\quad D/u_{\bar{k}} - \lambda^*_{\bar{k}} - \mu^*_{\bar{k}} \leq D/u_{\bar{k}} - \lambda^*_{\bar{k}} + \mu^*_{\bar{k}} = \\ = &\quad D/u_{\bar{k}} - \lambda^*_{\bar{k}} + \Delta_{\bar{k}} - \lambda^*_{\bar{k}} + D/u_{\bar{k}} = \Delta_{\bar{k}} + 2(D/u_{\bar{k}} - \lambda^*_{\bar{k}}) \leq \Delta_{\bar{k}}, \end{aligned}$$

  and we conclude that

  $$|(\hat{\lambda}_{\bar{k}} - \hat{\mu}_{\bar{k}}) - (\lambda^*_{\bar{k}} - \mu^*_{\bar{k}})| \leq \Delta_{\bar{k}}.$$

Observe now that $(w^*, \gamma^*, \xi^*, \zeta^*, y^*)$ is feasible for $LR(\hat{\lambda}, \hat{\mu})$ and let $(\hat{w}, \hat{\gamma}, \hat{\xi}, \hat{\zeta}, \hat{y})$ be any optimal solution for $LR(\hat{\lambda}, \hat{\mu})$

We discuss, separately, the objective function values associated to such solution for both $LR_1(\hat{\lambda}, \hat{\mu})$ and $LR_2(\hat{\lambda}, \hat{\mu})$.

As for $LR_1(\hat{\lambda}, \hat{\mu})$, taking into account 2.13, the following holds:

$$
\begin{array}{l}
z_1(\hat{\lambda}, \hat{\mu}) = \\
\|\hat{w}\| + C\left(\sum_{i=1}^{m_1} \hat{\xi}_i + \sum_{l=1}^{m_2} \hat{\zeta}_l\right) + \sum_{k=1}^{n} (\hat{\lambda_k} - \hat{\mu_k})\hat{w}_k = \\
\|\hat{w}\| + C\left(\sum_{i=1}^{m_1} \hat{\xi}_i + \sum_{l=1}^{m_2} \hat{\zeta}_l\right) + \sum_{k \neq \bar{k}} (\lambda_k^* - \mu_k^*)\hat{w}_k + (\hat{\lambda}_{\bar{k}} - \hat{\mu}_{\bar{k}})\hat{w}_{\bar{k}} + \\
+(\lambda_{\bar{k}}^* - \mu_{\bar{k}}^*)\hat{w}_{\bar{k}} - (\lambda_{\bar{k}}^* - \mu_{\bar{k}}^*)\hat{w}_{\bar{k}} = \\
\|\hat{w}\| + C\left(\sum_{i=1}^{m_1} \hat{\xi}_i + \sum_{l=1}^{m_2} \hat{\zeta}_l\right) + \sum_{k=1}^{n} (\lambda_k^* - \mu_k^*)\hat{w}_k + [(\hat{\lambda}_{\bar{k}} - \hat{\mu}_{\bar{k}}) - (\lambda_{\bar{k}}^* - \mu_{\bar{k}}^*)])\hat{w}_{\bar{k}} \geq \\
z_1(\lambda^*, \mu^*) + [(\hat{\lambda}_{\bar{k}} - \hat{\mu}_{\bar{k}}) - (\lambda_{\bar{k}}^* - \mu_{\bar{k}}^*)])\hat{w}_{\bar{k}} \geq z_1(\lambda^*, \mu^*) - u_{\bar{k}}\Delta_{\bar{k}},
\end{array}
\tag{2.15}
$$

which implies that in the new multiplier setting the decrease in the optimal value of problem $LR_1$ is bounded by $u_{\bar{k}}\Delta_{\bar{k}}$.

On the other hand, considering problem $LR_2(\hat{\lambda}, \hat{\mu})$, we have

$$z_2(\hat{\lambda}, \hat{\mu}) = z_2(\lambda^*, \mu^*) + u_{\bar{k}}\Delta_{\bar{k}}.$$

Summing up, we conclude that $z(\hat{\lambda}, \hat{\mu}) \geq z(\lambda^*, \mu^*)$.

Now consider the second possibility that for some index $\bar{k}$ it is

$$u_{\bar{k}}(\lambda_{\bar{k}}^* + \mu_{\bar{k}}^*) - D < 0,$$

with the corresponding $y_{\bar{k}}^* = 0$ and let

$$\Gamma_{\bar{k}} \stackrel{\triangle}{=} D/u_{\bar{k}} - (\lambda_{\bar{k}}^* + \mu_{\bar{k}}^*) > 0.$$

We consider now the following feasible solution $(\hat{\lambda}, \hat{\mu})$ for the Lagrangian dual:

$$\hat{\lambda}_k = \lambda_k^*, \quad \hat{\mu}_k = \mu_k^* \quad \text{for} \quad k \neq \bar{k},$$

and

$$
\begin{cases}
\hat{\lambda}_{\bar{k}} = \lambda_{\bar{k}}^* + \dfrac{\Gamma_{\bar{k}}}{2} \\
\hat{\mu}_{\bar{k}} = \mu_{\bar{k}}^* + \dfrac{\Gamma_{\bar{k}}}{2},
\end{cases}
\tag{2.16}
$$

Note that it is now

$$(\hat{\lambda}_{\bar{k}} - \hat{\mu}_{\bar{k}}) - (\lambda^*_{\bar{k}} - \mu^*_{\bar{k}}) = 0$$

and thus, see 2.15, it is $z_1(\hat{\lambda}, \hat{\mu}) \geq z_1(\lambda^*, \mu^*)$.

The thesis follows noting, finally, that the optimal value of $LR_2$ does not change in consequence of the modification of the variables $(\lambda, \mu)$.
□

REMARK 2.2. *At points $(\lambda, \mu)$ satisfying (2.10) the dual function $z(\lambda, \mu)$ exhibits a kink. Moreover at such points it is $z_2(\lambda, \mu) = 0$.*

Proposition 2.1 and the above remark allows us to eliminate the variables $\mu_k$, thus substantially reducing the number of variables of the Lagrangian dual which we rewrite in the form:

$$\begin{aligned} z_{LD} = \quad & \max z(\lambda) \\ & 0 \leq \lambda_k \leq D/u_k, \quad k = 1, \ldots, n \end{aligned} \tag{2.17}$$

where $z(\lambda)$ is the Lagrangian function defined as:

$$\begin{aligned} z(\lambda) = \quad & \min_{w,\gamma,\xi,\zeta,y} \sum_{k=1}^{n} (w_k^+ + w_k^-) + C \left( \sum_{i=1}^{m_1} \xi_i + \sum_{l=1}^{m_2} \zeta_l \right) \\ & + \sum_{\substack{k=1 \\ u_k \neq 0}}^{n} (2\lambda_k - D/u_k)(w_k^+ - w_k^-) \end{aligned} \tag{2.18}$$

subject to

$$\sum_{k=1}^{n} a_i^k (w_k^+ - w_k^-) + \gamma \leq \xi_i - 1, \qquad i = 1, \ldots, m_1 \tag{2.19}$$

$$-\sum_{k=1}^{n} b_l^k (w_k^+ - w_k^-) - \gamma \leq \zeta_l - 1 \qquad l = 1, \ldots, m_2 \tag{2.20}$$

$$0 \leq w_k^+ \leq u_k, \qquad k = 1, \ldots, n \tag{2.21}$$

$$0 \leq w_k^- \leq u_k, \qquad k = 1, \ldots, n \tag{2.22}$$

$$\xi_i \geq 0, \qquad i = 1, \ldots, m_1 \tag{2.23}$$

$$\zeta_l \geq 0, \qquad l = 1, \ldots, m_2 \tag{2.24}$$

where we have introduced the transformation of variables (2.9).

**3. The algorithm.** The approach we propose to tackle problem (2.1)-(2.8) is based on the use of a nonsmooth optimization method to solve the Lagrangian dual

(2.17). In particular we adopt the (generalized) bundle method (GBM) introduced in [5], which is an iterative ascent method. In fact during the ascent process, every time a new estimate of the optimal multiplier vector $\lambda^*$ is achieved, we come out with an improved lower bound for the original problem. In addition a feasible solution (and consequently an upper bound) for the same problem can be easily obtained starting from $w(\lambda)$, the $w$ component of the optimal solution vector corresponding to the current setting of the multiplier vector $\lambda$. It is in fact sufficient setting $y_k = 1$ whenever it is $|w_k(\lambda)| > 0$ and $y_k = 0$ otherwise.

However in many practical cases, typically for datasets where the number of features is large while the number of samples is small, the percentage of features $k$ such that $w_k(\lambda) \neq 0$ is very small. If this occurs, we prefer to act in a slightly different way to obtain a feasible solution of the original problem from an optimal solution of the relaxed one. In fact we define the set

$$\Omega_\epsilon(\lambda) \stackrel{\triangle}{=} \{k = 1, \ldots, n \,:\, |w_k(\lambda)| > \epsilon\, u_k\},$$

for some $0 < \epsilon < 1$, and state the restricted MBP problem:

$$P(\Omega_\epsilon(\lambda)) = \left| \begin{array}{ll} z_R(\Omega_\epsilon(\lambda)) = \min\limits_{w,\gamma,\xi,\zeta,y} \sum\limits_{k\in\Omega_\epsilon(\lambda)} (w_k^+ + w_k^-) + & \\ \qquad C\left(\sum\limits_{i=1}^{m_1} \xi_i + \sum\limits_{l=1}^{m_2} \zeta_l\right) + D \sum\limits_{k\in\Omega_\epsilon(\lambda)} y_k & \\ \quad \text{subject to} & \\ \sum_{k\in\Omega_\epsilon(\lambda)} a_i^k (w_k^+ - w_k^-) + \gamma \leq \xi_i - 1, & i = 1, \ldots, m_1 \\ -\sum_{k\in\Omega_\epsilon(\lambda)} b_l^k (w_k^+ - w_k^-) - \gamma \leq \zeta_l - 1 & l = 1, \ldots, m_2 \\ w_k^+ - w_k^- \leq u_k y_k, & k \in \Omega_\epsilon(\lambda) \\ w_k^- - w_k^+ \leq u_k y_k, & k \in \Omega_\epsilon(\lambda) \\ \xi_i \geq 0, & i = 1, \ldots, m_1 \\ \zeta_l \geq 0, & l = 1, \ldots, m_2 \\ y_k \in \{0, 1\}, & k \in \Omega_\epsilon(\lambda) \end{array} \right.$$

Of course the optimal solution of $P(\Omega_\epsilon(\lambda))$ is feasible for (2.1)–(2.8) and it is quite easy to obtain in all cases where $|\Omega_\epsilon(\lambda)| \ll n$.

The algorithm is summarized in the following.

0. Choose any initial estimate $\lambda^{(0)} \geq 0$ of the optimal solution to (2.17) and set the bundle ascent iteration counter $t = 0$. Calculate $z(\lambda^{(0)})$ and initialize the upper bound $UB = z_R(\Omega_\epsilon(\lambda^{(0)}))$.
1. Run the GBS until either stopping at the current $\lambda^{(t)}$ upon satisfaction of the GBM stopping criterion occurs or a new estimate $\lambda^{(t+1)}$ of the optimal solution to (2.17), with $z(\lambda^{t+1}) > z(\lambda^{(t)})$, is obtained.
2. Calculate a feasible solution to problem (2.1)-(2.8) by solving the restricted problem $P(\Omega_\epsilon(\lambda^{(t+1)}))$. Update the upper bound by setting $UB = \min\{UB, z_R(\Omega_\epsilon(\lambda^{(t+1)}))\}$
3. Set $t = t + 1$ and return to step 1.

We remark that throughout the algorithm, every time function $z(\lambda)$ is to be evaluated, either at step 0 or inside GBM, a call to any LP solver is needed. A call to a MBP solver is needed as well whenever the restricted MBP problem is to be solved at step 0 or at step 2.

**4. Numerical approach.** We have implemented the proposed approach within a general-purpose `C++` bundle code developed by A. Frangioni [5] and already used with success in several other applications [6].

Whenever it has been necessary to solve a Linear Program or a Mixed Binary Program the solver provided by the Cplex package has been adopted.

The setting of the $u_j$, $j = 1, \ldots, n$ in all our experiments has been made after a preprocessing phase, where for each dataset the LP relaxation of problem (2.1)-(2.8) have been solved for different values of $C$ and $D$ and for very large values of the $u_j$'s. After that, the $u_j$'s as been set to the maximal value of the corresponding $w_j$'s ever attained during pre-processing.

The following datasets available at *http://www.tech.plym.ac.uk/spmc/* and described in [13] have been adopted to test our code (here $m = m_1 + m_2$ is the total number of samples):

| Dataset name | $m$ | $n$ |
|---|---|---|
| Carcinoma | 36 | 7457 |
| DLBCL | 77 | 7129 |
| Leukemia | 72 | 5327 |
| Tumor1 | 60 | 7129 |
| Tumor2 | 50 | 12625 |

TABLE 4.1
*Description of the instances*

Before testing our appraoch, we have performed the so called *model selection* phase, whose objective is tuning of the parameter $C$ in a standard SVM setting. To this aim we have solved the continuous relaxation of problem (2.1)-(2.8) with $D = 0$ for a grid of values of $C$. In particular we have considered four values of $C$ equally spaced in the interval $[10^{-1}, 10^2]$. We have kept completely separated the *training*

and the *testing* phase. Thus 10% of the samples of each dataset has been reserved for the testing and then we have operated according to the *tenfold-cross-validation* paradigm. In fact we have partitioned the remaining 90% in ten subsets of identical size and we have trained ten classifiers, using every time nine out of the ten subsets. Every classifier has been finally tested against the independent test set. As usual, classification correctness has been defined as the percentage of well classified points over the cardinality of the considered sample set.

For each dataset and for each value of $C$, the average classification correctness, both in the testing phase (column $c_1$) and in the training one (column $c_2$), is reported in table 4.2, together with the average $L_1$ norm of $w$ in column $|w|_1$.

| | $C = 0.1$ | | | $C = 1$ | | | $C = 10$ | | | $C = 100$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # | $c_1$ | $c_2$ | $\|w\|_1$ | $c_1$ | $c_2$ | $\|w\|_1$ | $c_1$ | $c_2$ | $\|w\|_1$ | $c_1$ | $c_2$ | $\|w\|_1$ |
| 1 | 75.83 | 89.51 | 1.17 | 94.17 | 100 | 2.61 | 94.17 | 100 | 2.61 | 94.17 | 100 | 2.61 |
| 2 | 75.71 | 75.37 | 0.00 | 97.14 | 100 | 4.88 | 94.05 | 100 | 5.07 | 94.05 | 100 | 5.07 |
| 3 | 93.81 | 96.52 | 2.54 | 95.00 | 100 | 4.43 | 95.00 | 100 | 4.43 | 95.00 | 100 | 4.43 |
| 4 | 66.67 | 66.05 | 0.00 | 69.67 | 100 | 7.04 | 69.67 | 100 | 7.06 | 69.67 | 100 | 7.06 |
| 5 | 62.00 | 83.27 | 1.42 | 81.50 | 100 | 4.97 | 79.50 | 100 | 4.99 | 79.50 | 100 | 4.99 |

TABLE 4.2
*SVM - TenFold Cross Validation*

For sake of completeness, we have operated according to the above described guidelines with the tenfold cross validation replaced by the *leave-one-out* one. The correspondent results are in table 4.3.

Finally, on the basis of the results we have obtained, we have decided to proceed setting $C = 1$ in the experiments aimed at testing effectiveness of the proposed method in terms of feature selection.

In particular our tests have been designed to address two different issues:

- The effect of the weighting parameter $D$ on the number of relevant features and on classification correctness;
- The performance of the proposed Lagrangian heuristic w.r.t. the exact solution of the MBP formulation.

As for the first issue, the grid of values $(0.01, 0.1, 1, 10)$ has been chosen for parameter $D$.

| | $C = 0.1$ | | | $C = 1$ | | | $C = 10$ | | | $C = 100$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # | $c_1$ | $c_2$ | $\|w\|_1$ | $c_1$ | $c_2$ | $\|w\|_1$ | $c_1$ | $c_2$ | $\|w\|_1$ | $c_1$ | $c_2$ | $\|w\|_1$ |
| 1 | 90.63 | 96.27 | 1.50 | 96.88 | 100 | 2.71 | 96.88 | 100 | 2.71 | 96.88 | 100 | 2.71 |
| 2 | 75.36 | 75.36 | 0.00 | 97.10 | 100 | 5.23 | 97.10 | 100 | 5.23 | 97.10 | 100 | 5.23 |
| 3 | 93.81 | 96.52 | 2.54 | 93.75 | 100 | 4.58 | 95.31 | 100 | 4.59 | 95.31 | 100 | 4.59 |
| 4 | 66.04 | 66.04 | 0.00 | 67.92 | 100 | 7.43 | 67.92 | 100 | 7.43 | 67.92 | 100 | 7.43 |
| 5 | 54.55 | 84.99 | 1.60 | 81.82 | 100 | 5.22 | 81.82 | 100 | 5.22 | 81.82 | 100 | 5.22 |

TABLE 4.3
*SVM - Leave one out*

In particular, in tables 4.4 and 4.5 we show the results obtained by Cplex for the MBP formulation (2.1)-(2.8). For each value of $D$ in the grid and for each dataset we have considered here the 90% of the available samples and on such subset we have adopted a classic tenfold-cross-validation approach. We report the average training and testing correctness $c_1$ and $c_2$, respectively and the average norm of $w$. In addition, we report the average percentage of "significant" features, where a feature $k$ is defined significant whenever it is $|w_k| > 10^{-x}$ for values of $x = 0, 2, 4, 9$ (columns $ft\,0$, $ft\,2$, $ft\,4$, $ft\,9$, respectively). We have run the Cplex code with a maximum time bound of 1000 seconds (reporting the best solution found at the stop). We remark that the 10% part of each dataset has not been used at all during such experiment.

| | $C = 1\,,\ D = 0.01$ | | | | | | | $C = 1\,,\ D = 0.1$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # | $c_1$ | $c_2$ | $\|w\|_1$ | ft 0 | ft 2 | ft 4 | ft 9 | $c_1$ | $c_2$ | $\|w\|_1$ | ft 0 | ft 2 | ft 4 | ft 9 |
| 1 | 90.83 | 100 | 2.84 | 0.00 | 0.19 | 0.19 | 0.19 | 87.50 | 100 | 3.08 | 0.00 | 0.10 | 0.10 | 0.10 |
| 2 | 95.71 | 100 | 5.05 | 0.00 | 0.29 | 0.29 | 0.29 | 91.43 | 100 | 5.45 | 0.00 | 0.15 | 0.15 | 0.15 |
| 3 | 92.62 | 100 | 4.83 | 0.00 | 0.41 | 0.41 | 0.41 | 89.05 | 100 | 5.30 | 0.01 | 0.21 | 0.21 | 0.21 |
| 4 | 70.33 | 100 | 7.46 | 0.00 | 0.40 | 0.40 | 0.40 | 55.67 | 100 | 8.13 | 0.01 | 0.22 | 0.22 | 0.22 |
| 5 | 70.50 | 100 | 5.66 | 0.00 | 0.18 | 0.18 | 0.18 | 66.50 | 100 | 6.12 | 0.00 | 0.11 | 0.11 | 0.11 |

TABLE 4.4
*FS-SVM - TenFold Cross Validation - MBP (1)*

| | $C = 1\,,\ D = 1$ | | | | | | | $C = 1\,,\ D = 10$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # | $c_1$ | $c_2$ | $\|w\|_1$ | ft 0 | ft 2 | ft 4 | ft 9 | $c_1$ | $c_2$ | $\|w\|_1$ | ft 0 | ft 2 | ft 4 | ft 9 |
| 1 | 90.00 | 99 | 3.45 | 0.01 | 0.05 | 0.05 | 0.05 | 90.83 | 94 | 3.59 | 0.01 | 0.01 | 0.01 | 0.01 |
| 2 | 90.00 | 100 | 5.98 | 0.04 | 0.09 | 0.09 | 0.09 | 71.43 | 77 | 0.57 | 0.00 | 0.00 | 0.00 | 0.00 |
| 3 | 86.19 | 100 | 6.05 | 0.05 | 0.12 | 0.12 | 0.12 | 89.76 | 92 | 3.26 | 0.03 | 0.03 | 0.03 | 0.03 |
| 4 | 59.67 | 100 | 9.53 | 0.04 | 0.22 | 0.22 | 0.22 | 65.33 | 65 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 5 | 78.00 | 100 | 7.13 | 0.02 | 0.08 | 0.08 | 0.08 | 41.50 | 57 | 1.86 | 0.01 | 0.01 | 0.01 | 0.01 |

TABLE 4.5
*FS-SVM - TenFold Cross Validation - MBP (2)*

As expected, increasing values of $D$ result in deterioration of both classification correctness and margin, while we are better off with the number of significant features which reduces.

To address the second issue, similar experiments have been performed by replacing the Cplex solver by our Lagrangian relaxation approach. The corresponding results are summarized in tables 4.6 and 4.7

To have a closer view to comparison of our method with the exact approach we have made some additional experiment. We have set the parameters $C$ and $D$ to the values 1 and 0.01, respectively. Then we have trained the classifier, for each data set, on the previously mentioned 90% of the total number of samples and tested it on the remaining 10%. As for the Cplex solver we have fixed the time bound to 5, 10 and 1000 seconds. The corresponding results are in the tables 4.8, 4.9 and 4.10.

| | $C = 1$ , $D = 0.01$ | | | | | | | $C = 1$ , $D = 0.1$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # | $c_1$ | $c_2$ | $\|w\|_1$ | ft 0 | ft 2 | ft 4 | ft 9 | $c_1$ | $c_2$ | $\|w\|_1$ | ft 0 | ft 2 | ft 4 | ft 9 |
| 1 | 90.83 | 100 | 2.82 | 0.00 | 0.26 | 0.26 | 0.26 | 87.50 | 100 | 3.08 | 0.00 | 0.13 | 0.14 | 0.14 |
| 2 | 95.71 | 100 | 5.03 | 0.00 | 0.36 | 0.36 | 0.36 | 94.29 | 100 | 5.09 | 0.00 | 0.33 | 0.33 | 0.33 |
| 3 | 92.62 | 100 | 4.80 | 0.00 | 0.52 | 0.53 | 0.53 | 87.86 | 100 | 4.92 | 0.00 | 0.45 | 0.46 | 0.46 |
| 4 | 70.33 | 100 | 7.44 | 0.00 | 0.47 | 0.47 | 0.47 | 66.67 | 100 | 7.67 | 0.00 | 0.36 | 0.36 | 0.36 |
| 5 | 65.50 | 100 | 5.65 | 0.00 | 0.21 | 0.22 | 0.22 | 68.00 | 100 | 5.87 | 0.00 | 0.16 | 0.16 | 0.16 |

TABLE 4.6
*FS-SVM - TenFold Cross Validation - Lag. relaxation (1)*

| | $C = 1$ , $D = 1$ | | | | | | | $C = 1$ , $D = 10$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # | $c_1$ | $c_2$ | $\|w\|_1$ | ft 0 | ft 2 | ft 4 | ft 9 | $c_1$ | $c_2$ | $\|w\|_1$ | ft 0 | ft 2 | ft 4 | ft 9 |
| 1 | 90.00 | 100 | 3.73 | 0.01 | 0.09 | 0.09 | 0.09 | 93.33 | 95 | 3.57 | 0.02 | 0.03 | 0.03 | 0.03 |
| 2 | 94.29 | 100 | 6.12 | 0.01 | 0.16 | 0.16 | 0.16 | 85.71 | 89 | 3.97 | 0.03 | 0.05 | 0.05 | 0.05 |
| 3 | 91.19 | 100 | 6.11 | 0.03 | 0.24 | 0.24 | 0.24 | 97.14 | 99 | 6.32 | 0.03 | 0.19 | 0.19 | 0.19 |
| 4 | 58.00 | 99 | 9.21 | 0.05 | 0.19 | 0.19 | 0.19 | 62.00 | 94 | 6.80 | 0.01 | 0.36 | 0.38 | 0.38 |
| 5 | 70.00 | 100 | 6.94 | 0.02 | 0.09 | 0.09 | 0.09 | 76.50 | 93 | 5.96 | 0.02 | 0.05 | 0.05 | 0.05 |

TABLE 4.7
*FS-SVM - TenFold Cross Validation - Lag. relaxation (2)*

| | $C = 1$ , $D = 0.01$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| # | $c_1$ | $c_2$ | $\|w\|_1$ | ft 0 | ft 2 | ft 4 | ft 9 | time | LB | UB |
| 1 | 100.00 | 100 | 3.00 | 0.00 | 0.27 | 0.27 | 0.27 | 5.23 | 3.03 | 3.20 |
| 2 | 100.00 | 100 | 5.24 | 0.00 | 0.43 | 0.46 | 0.46 | 5.01 | 5.30 | 5.57 |
| 3 | 100.00 | 100 | 5.02 | 0.00 | 0.64 | 0.69 | 0.69 | 4.99 | 5.06 | 5.39 |
| 4 | 66.67 | 100 | 7.95 | 0.00 | 0.63 | 0.63 | 0.63 | 5.01 | 7.98 | 8.40 |
| 5 | 80.00 | 100 | 6.02 | 0.00 | 0.29 | 0.29 | 0.29 | 5.00 | 6.07 | 6.39 |

TABLE 4.8
*FS-SVM - Testing - MBP (time 5 sec)*

| | $C = 1$ , $D = 0.01$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| # | $c_1$ | $c_2$ | $\|w\|_1$ | ft 0 | ft 2 | ft 4 | ft 9 | time | LB | UB |
| 1 | 100.00 | 100 | 3.00 | 0.00 | 0.23 | 0.23 | 0.23 | 9.99 | 3.04 | 3.17 |
| 2 | 100.00 | 100 | 5.24 | 0.00 | 0.43 | 0.46 | 0.46 | 9.97 | 5.30 | 5.57 |
| 3 | 100.00 | 100 | 5.04 | 0.00 | 0.51 | 0.51 | 0.51 | 10.03 | 5.08 | 5.31 |
| 4 | 66.67 | 100 | 7.95 | 0.00 | 0.63 | 0.63 | 0.63 | 9.98 | 8.00 | 8.40 |
| 5 | 80.00 | 100 | 6.02 | 0.00 | 0.29 | 0.29 | 0.29 | 10.03 | 6.07 | 6.39 |

TABLE 4.9
*FS-SVM - Testing - MBP (time 10 sec)*

The results obtained by our method, together with the corresponding computing time, are reported in table 4.11.

Comparison of the results in tables 4.8, 4.9 and 4.10 with those of table 4.11 indicates that the results obtained by using our Lagrangian relaxation approach are of rather good quality in terms of lower and upper bounds and are obtained at a reasonable computational effort (see, in particular, table 4.9).

| | $C = 1\,,\ D = 0.01$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| # | $c_1$ | $c_2$ | $\|w\|_1$ | ft 0 | ft 2 | ft 4 | ft 9 | time | LB | UB |
| 1 | 100.00 | 100 | 3.00 | 0.00 | 0.20 | 0.20 | 0.20 | 996.80 | 3.13 | 3.15 |
| 2 | 100.00 | 100 | 5.29 | 0.00 | 0.32 | 0.32 | 0.32 | 997.05 | 5.44 | 5.52 |
| 3 | 100.00 | 100 | 5.07 | 0.00 | 0.41 | 0.41 | 0.41 | 996.87 | 5.20 | 5.29 |
| 4 | 66.67 | 100 | 8.03 | 0.00 | 0.43 | 0.43 | 0.43 | 996.88 | 8.14 | 8.34 |
| 5 | 80.00 | 100 | 6.06 | 0.00 | 0.20 | 0.20 | 0.20 | 996.77 | 6.20 | 6.31 |

TABLE 4.10
*FS-SVM - Testing - MBP (time 1000 sec)*

| | $C = 1\,,\ D = 0.01$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| # | $c_1$ | $c_2$ | $\|w\|_1$ | ft 0 | ft 2 | ft 4 | ft 9 | time | LB | UB |
| 1 | 100.00 | 100 | 3.00 | 0.00 | 0.20 | 0.20 | 0.20 | 2.50 | 3.03 | 3.15 |
| 2 | 100.00 | 100 | 5.29 | 0.00 | 0.32 | 0.32 | 0.32 | 9.49 | 5.30 | 5.52 |
| 3 | 100.00 | 100 | 5.05 | 0.00 | 0.45 | 0.45 | 0.45 | 4.62 | 5.06 | 5.29 |
| 4 | 83.33 | 100 | 8.00 | 0.00 | 0.49 | 0.49 | 0.49 | 4.14 | 7.98 | 8.35 |
| 5 | 80.00 | 100 | 6.06 | 0.00 | 0.21 | 0.21 | 0.21 | 3.88 | 6.07 | 6.33 |

TABLE 4.11
*FS-SVM - Testing - Lag. relaxation*

**5. Conclusions.** The numerical examples we have worked indicate that the MBP model is able to provide an acceptable tradeoff between classification quality and number of significant features. On the other hand the Lagrangian relaxation approach we have introduced appears to produce good quality solution at a quite affordable computational cost.

REFERENCES


[1] H. Aytug, *Feature selection for support vctor machines using generalized benders decomposition*, ejor, 244 (2015), pp. 210–218.

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

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

[4] E.Carrizosa and D. Romero Morales, *Supervised classification and mathematical optimization*, Computers and Operations Research, 40 (2013).

[5] A. Frangioni, *Generalized Bundle Methods*, SIAM Journal on Optimization, 13 (2002), pp. 117–156.

[6] A. Frangioni and E. Gorgone, *Generalized bundle methods for sum-functions with "easy" components: Applications to multicommodity network design*, Mathematical Programming, 145 (2014), pp. 133–161.

[7] M. Gaudioso, G. Giallombardo, and G. Miglionico, *On solving the lagrangian dual of integer programs via an incremental approach*, Computational Optimization and Applications, 44 (2009), pp. 117–138.

[8] I. Guyon and A. Elisseef, *An introduction to variable and feature selection*, The Journal of Machine Learning Research, 3 (2003), pp. 1157–1182.

[9] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* II—*Advanced Theory and Bundle Methods*, vol. 306 of Grundlehren Math. Wiss., Springer-Verlag, New York, 1993.

[10] J. Kittler, *Feature selection and extraction*, in Handbook of Pattern Recognition and Image Processing, A .Young, ed., Academic Press, New York, 1986.

[11] E. Lee and W. Tsung-Lin, *Classification and disease prediction via mathematical programming*, in Handbook of Optimizationin Medicine, P.M. Pardalos and H.E. Romeijn, eds., Springer Optimization and Its Applications 26, 2009.

[12] S. Maldonado, J. Pérez, R. Weber, and M. Labbé, *Feature selection for support vector machines via mixed integer linear programming*, Information Sciences, 279 (2014), pp. 163–175.

[13] P. E. Meyer, C. Schretter, and G. Bontempi, *Information-theoretic feature selection in microarray data using variable complementarity*, IEEE Signal Processing Society, 20 (2008), pp. 261–274.

[14] F. Rinaldi and M. Sciandrone, *Feature selection combining linear support vector machines and concave optimization*, Optimization Methods and Software, 10 (2010), pp. 117–128.

[15] V. Vapnik, *The nature of the statistical learning theory*, Springer Verlag, New York, 1995.

[16] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, *Feature selection for SVMs*, Advances in Neural Information Processing Systems, 12 (2000), pp. 668–674.