Computers and Operations Research 87 (2017) 137–145

Contents lists available at ScienceDirect

Computers and Operations Research

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

Lagrangian relaxation for SVM feature selection

M. Gaudioso [a,*], E. Gorgone [d,e], M. Labbé [b], A.M. Rodríguez-Chía [c]

[a] Dipartimento di Ingegneria Informatica, Modellistica, Elettronica e Sistemistica, Università della Calabria, Rende (CS) 87036, Italia
[b] Département d'Informatique, Université Libre de Bruxelles, CP 212 Boulevard du Triomphe, Brussels B-1050, Belgium
[c] Faculty of Sciences, Universidad de Cádiz Avenida República Saharaui s/n 11510. Puerto Real Cádiz, Spain
[d] Indian Institute of Management Bangalore (IIMB), Bannerghatta Road, Bangalore 560076, India
[e] Dipartimento di Matematica e Informatica, Università di Cagliari, Cagliari 09124, Italy

ARTICLE INFO

Article history:
Received 20 July 2016
Revised 1 June 2017
Accepted 1 June 2017
Available online 15 June 2017

Keywords:
SVM classification
Feature selection
Lagrangian relaxation
Nonsmooth optimization

ABSTRACT

We discuss a Lagrangian-relaxation-based heuristics for dealing with feature selection in the Support Vector Machine (SVM) framework for binary classification. In particular we embed into our objective function a weighted combination of the L_1 and L_0 norm of the normal to the separating hyperplane. We come out with a Mixed Binary Linear Programming problem which 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.

1. Introduction

The focus of pattern classification is to recognize similarities in the data, categorizing them in different subsets (Cristianini and Shawe-Taylor, 2000; Carrizosa and Morales, 2013; Vapnik, 1995). In many fields, such as the financial and the medical ones (Guyon et al., 2002; Lee and Tsung-Lin, 2009), classification of data (samples in Machine Learning language) is useful for analysis or diagnosis purposes.

Quite often datasets are formed by a small number of samples, which in turn are characterized by a huge number of attributes (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 relevant for classification purposes.

Hence it is necessary to screen off the relevant features from those which are irrelevant (Bi et al., 2003).

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. Guyon and Elisseeff (2003); Guyon et al. (2006); Kittler (1986); Meyer et al. (2008)). As far as mathematical programming-based approaches are concerned, we cite here Weston et al. (2000) and, more specifically, Bradley et al. (1998), where (FS) is pursued by formulating a mathematical program with a parametric objective function embedding a concave approximation of the zero-norm of the feature vector. Other approaches based on zero-norm minimization are in Weston et al. (2003) and Rinaldi and Sciandrone (2010). In a recent paper (Aytug, 2015) a model based on penalization of the number of features entering into the classification process is treated by means of generalized Benders decomposition.

The objective of this paper is to treat explicitly the (FS) problem as a Mixed Binary Programming (MBP) one (Bertolazzi et al., 2016; Maldonado et al., 2014), in the framework of the SVM (Support Vector Machine) (Chen and Lin, 2006; Cristianini and Shawe-Taylor, 2000; Nguyen and de la Torre, 2010; Vapnik, 1995) approach. Consequently the objective of our model is threefold: to minimize the classification error, to maximize the separation margin and to minimize the number features playing a role in the classification process. To this aim we embed into the objective function, in addition to a measure of the classification error, the weighted combination of the L_1 and L_0 norms of the feature vector. We note that use of such norms (in particular the L_1 is also known as LASSO penalty) has been adopted in several papers in the SVM framework.

* Corresponding author.
E-mail addresses: gaudioso@dimes.unical.it (M. Gaudioso), enrico.gorgone@iimb.ernet.in, egorgone@unica.it (E. Gorgone), mlabbe@ulb.ac.be (M. Labbé), antonio.rodriguezchia@uca.es (A.M. Rodríguez-Chía).

http://dx.doi.org/10.1016/j.cor.2017.06.001
0305-0548/

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 Maldonado et al. (2014), and in the use of the method described in Frangioni (2002) which belongs to the class of the bundle ones (Astorino et al., 2013, 2011; Hiriart-Urruty and Lemaréchal, 1993). It implements an ascent procedure for solving the related Lagrangian Dual problem, which is, of course, a nonsmooth one (Fortz et al., 2017; Frangioni and Gorgone, 2014; Frangioni et al., 2017; Gaudioso et al., 2009). 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 datasets from cancer classification (Meyer et al., 2008) as well as on some datasets widely used in classification literature. Some conclusions are finally drawn in Section 5.

2. The model

Given two point-sets $\mathcal{A} \triangleq \{a_1, \ldots, a_{m_1}\}$ and $\mathcal{B} \triangleq \{b_1, \ldots, b_{m_2}\}$ in $\mathbb{R}^n$, we seek a hyperplane (w, γ) that separates $\mathcal{A}$ and $\mathcal{B}$. We define the *classification error variables* ξ_i and ζ_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 Maldonado et al. (2014), 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 Maldonado et al. (2014) budget-type constraints are in action. In particular we come out with the following MBP formulation of our SVM-feature-selection problem:

$$z^* = \min_{w,\gamma,\xi,\zeta,y} \|w\|_1 + C\left(\sum_{i=1}^{m_1}\xi_i + \sum_{l=1}^{m_2}\zeta_l\right) + D\sum_{k=1}^{n} y_k \tag{2.1}$$

subject to

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

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

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

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

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

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

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

where $u_k > 0$ is an upper bound on the modulus of the kth component of w, whose setting is discussed later in Section 4. 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, throughout the paper, the L_1 norm), (ii) the classification error, (iii) the number of active features (the L_0 norm of w). Note that the rationale behind minimizing both the L_1 and L_0 norm of w is to achieve, respectively, large separation margin and selection of the relevant features (see also Liu and Wu (2007) for an alternative computational approach). The positive parameters C and D weight the different objectives. We remark that in the SVM approach minimization of $\|w\|_1$ corresponds to maximize the separation margin and that nonlinearity in the model can be easily eliminated by letting:

$$w_k = w_k^+ - w_k^-, \quad w_k^+ \ge 0, \quad w_k^- \ge 0, \quad k = 1, \ldots, n \text{ and } \|w\|_1 = \sum_{k=1}^{n}(w_k^+ + w_k^-) \tag{2.9}$$

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:

$$LR(\lambda, \mu) = \left| \begin{array}{l} z(\lambda, \mu) = \min_{w,\gamma,\xi,\zeta,y} \|w\|_1 + C\left(\sum_{i=1}^{m_1}\xi_i + \sum_{l=1}^{m_2}\zeta_l\right) + D\sum_{k=1}^{n} y_k \\ \quad + \sum_{k=1}^{n}\lambda_k(w_k - u_k y_k) - \sum_{k=1}^{n}\mu_k(w_k + u_k y_k) \\ \text{subject to } (2.2-2.3), (2.5-2.8)\,. \end{array} \right.$$

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, γ, ξ and ζ while the binary variables y_k are confined to the latter one. In details, one comes out with:

$$LR_1(\lambda, \mu) = \left| \begin{array}{l} z_1(\lambda, \mu) = \min_{w,\gamma,\xi,\zeta} \|w\|_1 + C\left(\sum_{i=1}^{m_1}\xi_i + \sum_{l=1}^{m_2}\zeta_l\right) \\ \quad + \sum_{k=1}^{n}(\lambda_k - \mu_k)w_k \\ \text{subject to } (2.2-2.3), (2.5-2.7)\,. \end{array} \right.$$

and

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

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, whose proof is in the Appendix.

Proposition 1. *There exists an optimal solution to the Lagrangian dual satisfying the condition*

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

Remark 1. At points (λ, μ) satisfying (2.10) the dual function $z(\lambda, \mu)$ exhibits a kink. Moreover at such points it is $z_2(\lambda, \mu) = 0$.

Proposition 1 and the above remark allows us to eliminate the variables μ_k, thus substantially reducing the number of variables of the Lagrangian dual which we rewrite in the form:

$$\begin{aligned} z_{LD} = \quad & \max z(\lambda) \\ & 0 \leq \lambda_k \leq D/u_k, \quad k = 1,\ldots,n \end{aligned} \tag{2.11}$$

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

$$\begin{aligned} z(\lambda) = \min_{w,\gamma,\xi,\zeta} & \sum_{k=1}^{n}(w_k^+ + w_k^-) + C\left(\sum_{i=1}^{m_1}\xi_i + \sum_{l=1}^{m_2}\zeta_l\right) \\ & + \sum_{\substack{k=1 \\ u_k\neq 0}}^{n}(2\lambda_k - D/u_k)(w_k^+ - w_k^-) \end{aligned} \tag{2.12}$$

subject to

$$\sum_{k=1}^{n} a_i^k(w_k^+ - w_k^-) + \gamma \leq \xi_i - 1, \qquad i = 1,\ldots,m_1 \tag{2.13}$$

$$-\sum_{k=1}^{n} b_l^k(w_k^+ - w_k^-) - \gamma \leq \zeta_l - 1 \qquad l = 1,\ldots,m_2 \tag{2.14}$$

$$0 \leq w_k^+ \leq u_k, \qquad k = 1,\ldots,n \tag{2.15}$$

$$0 \leq w_k^- \leq u_k, \qquad k = 1,\ldots,n \tag{2.16}$$

$$\xi_i \geq 0, \qquad i = 1,\ldots,m_1 \tag{2.17}$$

$$\zeta_l \geq 0, \qquad l = 1,\ldots,m_2 \tag{2.18}$$

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

Remark 2. An intuitive explanation of the third term of the objective function (2.12) is obtained by considering an iterative dual ascent algorithm for solving the Lagrangian dual. We assume D sufficiently large to guarantee $D/u_k > 1$.

Observe first that from the definition of problem 2.11, the cost coefficients $(2\lambda_k - D/u_k)$ in (2.12) is in the range $[-D/u_k, +D/u_k]$, $k = 1,\ldots,n$.

Now let $\bar{\lambda}_k$, $k = 1,\ldots,n$, be any current configuration of the multipliers and $\bar{w}^+$ $\bar{w}^-$ the correspondent optimal vectors of problem (2.12)–(2.18). Suppose, without loss of generality, $\bar{w}_{\bar{k}}^+ > 0$ and "small" for some index $\bar{k}$ (it is of course $\bar{w}_{\bar{k}}^- = 0$). In this case the effect of setting at next iteration the $\bar{k}$th multiplier to zero (that is letting $2\lambda_{\bar{k}} - D/u_{\bar{k}} = -D/u_{\bar{k}}$) and leaving the remaining multipliers unchanged, is both the dual ascent and the suppression of feature $\bar{k}$ (provided that variables ξ_is and ζ_ls do not change in the new multiplier configuration by effect of the (small) variation in $w_{\bar{k}}^+$).

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 (2.11). In particular we adopt the (generalized) bundle method (GBM) introduced in Frangioni (2002), which is an iterative ascent method. In fact during the ascent process, every time a new estimate of the optimal multiplier vector λ^* 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 λ. 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

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

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

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

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 (2.11) 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 (2.11), 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 MBP 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 results

We have implemented the proposed approach within a general-purpose C++ bundle code developed by Frangioni (2002) which, in turn, embeds a specialized quadratic solver described in Frangioni (1996). The software has been already used with success in other applications such as Frangioni and Gorgone (2014) and Fortz et al. (2017).

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_k, $k = 1, \ldots, n$ in all our experiments has been made for each dataset through a pre-processing phase based on the following steps.

- The LP relaxation of problem (2.1)–(2.8) is solved by dropping the variables y_ks together with constraints (2.4), (2.5) and (2.8). Letting $\bar{w}$ the optimal value of w, the set $K_1 \triangleq \{k \mid |\bar{w}_k| > 0\}$ is calculated;
- The problem (2.1)–(2.8) is solved with variables y_k fixed to zero for $k \notin K_1$ and the u_ks arbitrarily large. The optimal objective function value of such problem, say $\hat{z}$, is calculated;
- The following problem is solved:

$$\max_{w,\gamma,\xi,\zeta,y} \|w\|_1 \tag{4.1}$$

subject to

$$a_i^\top w + \gamma \le \xi_i - 1, \qquad i = 1, \ldots, m_1 \tag{4.2}$$

$$-b_l^\top w - \gamma \le \zeta_l - 1 \qquad l = 1, \ldots, m_2 \tag{4.3}$$

$$\|w\|_1 + C\left(\sum_{i=1}^{m_1} \xi_i + \sum_{l=1}^{m_2} \zeta_l\right) + D\sum_{k=1}^{n} y_k \le \hat{z} \tag{4.4}$$

$$-u_k y_k \le w_k \le u_k y_k, \qquad k = 1, \ldots, n \tag{4.5}$$

$$-u_k \le w_k \le u_k, \qquad k = 1, \ldots, n \tag{4.6}$$

$$\xi_i \ge 0, \qquad i = 1, \ldots, m_1 \tag{4.7}$$

$$\zeta_l \ge 0, \qquad l = 1, \ldots, m_2 \tag{4.8}$$

$$0 \le y_k \le 1, \qquad k = 1, \ldots, n, \tag{4.9}$$

and, in particular, the optimal value w^* of w is considered.

- The values u_ks to be used in the numerical experiments are finally obtained by setting:

$$u_k = \min\{u_k, w_k^*\}.$$

We have performed our experiments on two groups of five datasets each. Datasets 1–5 (Group 1), available at http://www.tech.plym.ac.uk/spmc/ , are characterized by relatively few samples with respect to the number of features. For Datasets 6–10 (Group 2), available at https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/ , the number of samples, instead, is dominant. For the latter ones a certain class overlap is then expected. Detailed descriptions are in Meyer et al. (2008) and Chang and Lin (2011).

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

Before testing our approach, we have performed the so called *model selection* phase, see Bi et al. (2003), whose objective, in the SVM framework, is the tuning of the parameter C. To this aim we have solved the following problem:

$$\bar{z} = \min_{w,\gamma,\xi,\zeta} \|w\|_1 + C\left(\sum_{i=1}^{m_1} \xi_i + \sum_{l=1}^{m_2} \zeta_l\right) \tag{4.10}$$

subject to

$$a_i^\top w + \gamma \le \xi_i - 1, \qquad i = 1, \ldots, m_1 \tag{4.11}$$

$$-b_l^\top w - \gamma \le \zeta_l - 1 \qquad l = 1, \ldots, m_2 \tag{4.12}$$

$$\xi_i \ge 0, \qquad i = 1, \ldots, m_1 \tag{4.13}$$

$$\zeta_l \ge 0, \qquad l = 1, \ldots, m_2, \tag{4.14}$$

which is exactly the optimization problem described in Bradley and Mangasarian (1998), where the SVM approach embedding into the objective function the L_1 norm of w is adopted.

A motivation for selecting the above model is that it has been shown in Bradley and Mangasarian (1998) that it is also a valid tool for feature selection purposes. Consequently, the results will be used in the sequel also for comparison purposes with our approach. We will refer to the above model as to L_1-SVM. We remark (see Mangasarian (1997) for a detailed discussion on the use of different norms) that minimization of the L_1 norm of w corresponds to maximize the separation margin, measured by means of the L_∞ norm.

We have considered a grid of 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 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 *Test*) and in the training one (column *Train*) is reported in Table 4.2, together with the average L_1 norm of w in column $|w|_1$.

Finally, on the basis of the results we have obtained, we have decided to proceed to test our feature selection approach setting $C = 1$ and $C = 10$ in the experiments concerning, respectively, the datasets of Group 1 and Group 2.

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.

In particular, in Tables 4.3 and 4.4 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 again reserved 10% of

Table 4.1
Description of the instances.

#	Datasets of Group 1	m	n	#	Datasets of Group 2	m	n
1	Carcinoma (CARC)	36	7457	6	Breast Cancer (BC)	683	10
2	DLBCL	77	7129	7	PIMA Indians Diabetes (PIMA)	768	8
3	Leukemia (LEK)	72	5327	8	HEART	270	13
4	Tumor1 (TUM1)	60	7129	9	Ionosphere (IONO)	351	34
5	Tumor2 (TUM2)	50	12625	10	Liver Disorders (LIVER)	145	5

Table 4.2
SVM - TenFold Cross Validation.

	$C = 0.1$			$C = 1$			$C = 10$			$C = 100$		
Dataset	Test	Train	$\|w\|_1$	Test	Train	$\|w\|_1$	Test	Train	$\|w\|_1$	Test	Train	$\|w\|_1$
CARC	75.83	89.51	1.17	94.17	100.00	2.61	94.17	100.00	2.61	94.17	100.00	2.61
DLBCL	75.71	75.37	0.00	97.14	100.00	4.88	94.05	100.00	5.07	94.05	100.00	5.07
LEUK	93.81	96.52	2.54	95.00	100.00	4.43	95.00	100.00	4.43	95.00	100.00	4.43
TUM1	66.67	66.05	0.00	69.67	100.00	7.04	69.67	100.00	7.06	69.67	100.00	7.06
TUM2	62.00	83.27	1.42	81.50	100.00	4.97	79.50	100.00	4.99	79.50	100.00	4.99
BC	96.09	96.62	2.74	96.74	97.25	5.16	96.74	97.24	6.01	96.74	97.22	6.37
PIMA	75.43	75.82	4.31	76.44	77.47	7.65	76.30	77.62	8.30	76.30	77.58	8.33
HEART	86.00	86.69	2.68	84.33	86.51	6.00	84.37	86.83	7.27	83.95	86.92	7.38
IONO	89.24	89.77	3.65	88.57	94.20	16.13	88.91	95.46	43.43	87.63	95.46	51.74
LIVER	62.03	61.83	0.00	72.58	73.20	4.23	72.58	74.38	5.14	74.12	74.47	5.31

Table 4.3
MBP- Cplex implementation-Group 1-Tenfold cross validation.

	$C = 1$, $D = 0.01$							$C = 1$, $D = 0.1$						
Dataset	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9
CARC	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
DLBCL	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
LEUK	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
TUM1	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
TUM2	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
	$C = 1$, $D = 1$							$C = 1$, $D = 10$						
Datasets	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9
CARC	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
DLBCL	94.25	100	5.98	0.04	0.09	0.09	0.09	71.43	77	0.57	0.00	0.00	0.00	0.00
LEUK	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
TUM1	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
TUM2	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 4.4
MBP- Cplex implementation-Group 2-Tenfold cross validation.

	$C = 10$, $D = 0.01$							$C = 10$, $D = 0.1$						
Datasets	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9
BC	96.74	97.22	6.00	9.00	92.00	92.00	92.00	96.74	97.22	5.97	9.00	90.00	90.00	90.00
PIMA	76.30	77.62	8.30	26.25	100.00	100.00	100.00	76.30	77.62	8.30	26.25	100.00	100.00	100.00
HEART	84.37	86.83	7.27	12.31	100.00	100.00	100.00	84.37	86.83	7.27	12.31	100.00	100.00	100.00
IONO	88.91	95.46	43.35	50.29	93.82	93.82	93.82	88.90	95.43	43.28	50.00	89.71	89.71	89.71
LIVER	72.58	74.39	5.14	48.00	100.00	100.00	100.00	72.58	74.39	5.14	48.00	100.00	100.00	100.00
	$C = 10$, $D = 1$							$C = 10$, $D = 10$						
Datasets	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9
BC	96.58	97.24	5.80	9.00	80.00	80.00	80.00	96.42	97.14	5.73	17.00	68.00	68.00	68.00
PIMA	76.30	77.60	8.27	26.25	96.25	96.25	96.25	75.86	77.26	7.74	26.25	78.75	78.75	78.75
HEART	83.95	86.88	7.33	13.07	98.46	98.46	98.46	83.95	86.97	5.16	9.23	58.46	58.46	58.46
IONO	88.59	95.46	42.13	49.71	80.30	80.30	80.30	88.28	94.61	27.06	36.18	45.30	45.30	45.30
LIVER	72.58	74.39	5.14	48.00	98.00	98.00	98.00	71.10	74.05	4.45	52.00	68.00	68.00	68.00

the samples for the testing and then we have partitioned the remaining 90% in ten subsets of identical size. We have trained ten classifiers, using every time nine out of the ten subsets and, finally, every classifier has been tested against the independent test set.

We report the average testing and training correctness in columns Test and Train, respectively, together with the average norm of w. We add the columns $ft0$, $ft2$, $ft4$, $ft9$ reporting the average percentage of features for which the corresponding component of w has turned out to be greater than $1, 10^{-2}, 10^{-4}, 10^{-9}$, respectively. Such values help in catching how strong is the feature suppression effect of any model. For example, relatively small values of $ft9$ indicate that many features are characterized by corresponding components of w smaller than or equal to 10^{-9} and thus can be considered not relevant.

Table 4.5
MBP-Lagrangian relaxation- Group 1-Tenfold cross validation.

	$C = 1$, $D = 0.01$							$C = 1$, $D = 0.1$						
Datasets	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9
CARC	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
DLBCL	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
LEUK	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
TUM1	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
TUM2	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
	$C = 1$, $D = 1$							$C = 1$, $D = 10$						
Datasets	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9
CARC	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
DLBCL	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
LEUK	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
TUM1	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
TUM2	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 4.6
MBP-Lagrangian relaxation- Group 2-Tenfold cross validation.

	$C = 10$, $D = 0.01$							$C = 10$, $D = 0.1$						
Datasets	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9
BC	96.74	97.22	6.00	9.00	92.00	92.00	92.00	96.74	97.22	5.99	9.00	91.00	91.00	91.00
PIMA	76.30	77.62	8.30	26.25	100.00	100.00	100.00	76.30	77.62	8.30	26.25	100.00	100.00	100.00
HEART	84.37	86.83	7.27	12.31	100.00	100.00	100.00	84.37	86.83	7.27	12.31	100.00	100.00	100.00
IONO	88.91	95.46	43.38	58.59	94.41	94.41	94.41	88.90	95.39	43.26	50.00	90.00	90.00	90.00
LIVER	72.58	74.39	5.14	48.00	100.00	100.00	100.00	72.58	74.39	5.14	48.00	100.00	100.00	100.00
	$C = 10$, $D = 1$							$C = 10$, $D = 10$						
Datasets	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9
BC	96.58	97.23	5.80	9.00	80.00	80.00	80.00	96.41	97.18	5.78	17.00	71.00	71.00	71.00
PIMA	76.30	77.60	8.27	26.25	96.25	96.25	96.25	76.01	77.42	8.02	26.25	87.50	87.50	87.50
HEART	83.95	86.88	7.32	13.08	98.46	98.46	98.46	83.95	86.79	6.71	12.31	82.31	82.31	82.31
IONO	88.28	95.46	42.47	49.41	87.94	87.94	87.94	87.93	95.18	34.31	40.00	67.65	67.65	67.65
LIVER	72.58	74.39	5.14	48.00	98.00	98.00	98.00	70.33	74.13	4.69	54.00	76.00	76.00	76.00

We have run the Cplex code with a maximum time bound of 1000 sec (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.

We observe different effects of increasing values of parameter D for the datasets of Group 1 and of Group 2. As for Group 1, we observe a drastic reduction in the number of relevant features, accompanied by a notable reduction of the testing correctness for four out of five datasets. On the other hand, for data sets of Group 2, the significant reduction in the number of relevant features does not impair testing correctness.

For comparison purposes we focus on the papers Bradley et al. (1998) and Maldonado et al. (2014). We consider in particular the two datasets *Ionosphere* and *PIMA Indian Diabetes*. In our results we assume not relevant any feature k such that $|w_k| \leq 10^{-9}$.

Considering first the Ionosphere dataset, where the total number of feature is 34, we have (see Table 4.4) an average percentage of non relevant features of approximately 6%, 10%, 20% and 55% for $D = 0.01;\ 0.1;\ 1;\ 10$, respectively. The average testing correctness is in the range (88.28, 88.91). In Bradley et al. (1998) the results of four different feature selection algorithms are presented. The number of non selected features corresponding to the maximum testing correctness achieved by each algorithm ranges in the interval (35%–58%). The testing correctness is slightly smaller than in our approach.

In Maldonado et al. (2014) the maximum testing correctness of 88.9 is achieved with 18 non selected features, corresponding to the 53% of the total.

As fo PIMA Indian Diabetes dataset, where the total number of features is 8, the results of our testing indicate an average percentage of non relevant features ranging in the interval 0%–11%, with average testing correctness in the range (75.86, 76.30). In Maldonado et al. (2014) the best result obtained in terms of testing correctness is 78.0 with all features selected.

To address the second issue, about the performance of our Lagrangian heuristics, similar experiments have been performed by replacing the Cplex solver by our Lagrangian relaxation approach. The corresponding results are summarized in Tables 4.5 and 4.6.

In analysing such results in comparison with those of the exact approach reported in Tables 4.3 and 4.4, we observe that the number of the suppressed features is uniformly higher in the exact solution, while no uniform behaviour can be detected as far as correctness is concerned.

To have a closer view in comparing the Lagrangian relaxation approach with the exact one, we have made some additional experiment. We have set the parameters (C, D) to the values (1, 0.01) and (10, 10), respectively, in the experiments for Group 1 and Group 2 datasets. 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 sec. The corresponding results are in the Tables 4.7–4.9. The results of the Lagrangian relaxation approach are in Table 4.10.

Note that the problems related to the datasets of Group 2 have been all solved at the optimum by CPLEX within 5 sec, thus it has not been necessary to adopt for them the time limits of 10 and 1000 sec. Note, in addition, that the Lagrangian relaxation approach provides for all of them the same upper bounds and the same results in terms of all the evaluation parameters, but time, for which CPLEX is faster.

Table 4.7
MBP- Cplex implementation-Testing (5 s).

	$C=1$, $D=0.01$									
Diabetes	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9	time	LB	UB
CARC	100.00	100	3.00	0.00	0.27	0.27	0.27	5.23	3.03	3.20
DLBCL	100.00	100	5.24	0.00	0.43	0.46	0.46	5.01	5.30	5.57
LEUK	100.00	100	5.02	0.00	0.64	0.69	0.69	4.99	5.06	5.39
TUM1	66.67	100	7.95	0.00	0.63	0.63	0.63	5.01	7.98	8.40
TUM2	80.00	100	6.02	0.00	0.29	0.29	0.29	5.00	6.07	6.39
	$C=10$, $D=10$									
Diabetes	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9	time	LB	UB
BC	98.53	97.07	5.99	30.00	70.00	70.00	70.00	0.08	490.97	490.97
PIMA	80.26	77.02	7.90	25.00	75.00	75.00	75.00	0.09	3759.37	3759.37
HEART	77.78	86.42	4.95	7.69	61.54	61.54	61.54	0.10	879.70	879.70
IONO	88.57	94.63	32.47	38.24	55.88	55.88	55.88	3.42	694.90	694.90
LIVER	92.86	75.57	5.37	60.00	80.00	80.00	80.00	0.03	815.94	815.94

Table 4.8
MBP- Cplex implementation-Testing (10 s).

	$C=1$, $D=0.01$									
Datasets	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9	time	LB	UB
CARC	100.00	100.00	3.00	0.00	0.23	0.23	0.23	9.99	3.04	3.17
DLBCL	100.00	100.00	5.24	0.00	0.43	0.46	0.46	9.97	5.30	5.57
LEUK	100.00	100.00	5.04	0.00	0.51	0.51	0.51	10.03	5.08	5.31
TUM1	66.67	100.00	7.95	0.00	0.63	0.63	0.63	9.98	8.00	8.40
TUM2	80.00	100.00	6.02	0.00	0.29	0.29	0.29	10.03	6.07	6.39

Table 4.9
MBP- Cplex implementation-Testing (1000 s).

	$C=1$, $D=0.01$									
Datasets	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9	time	LB	UB
CARC	100.00	100.00	3.00	0.00	0.20	0.20	0.20	996.80	3.13	3.15
DLBCL	100.00	100.00	5.29	0.00	0.32	0.32	0.32	997.05	5.44	5.52
LEUK	100.00	100.00	5.07	0.00	0.41	0.41	0.41	996.87	5.20	5.29
TUM1	66.67	100.00	8.03	0.00	0.43	0.43	0.43	996.88	8.14	8.34
TUM2	80.00	100.00	6.06	0.00	0.20	0.20	0.20	996.77	6.20	6.31

Table 4.10
MBP- Lagrangian relaxation-Testing.

	$C=1$, $D=0.01$									
Datasets	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9	time	LB	UB
CARC	100.00	100.00	3.00	0.00	0.20	0.20	0.20	2.50	3.03	3.15
DLBCL	100.00	100.00	5.29	0.00	0.32	0.32	0.32	9.49	5.30	5.52
LEUK	100.00	100.00	5.05	0.00	0.45	0.45	0.45	4.62	5.06	5.29
TUM1	83.33	100.00	8.00	0.00	0.49	0.49	0.49	4.14	7.98	8.35
TUM2	80.00	100.00	6.06	0.00	0.21	0.21	0.21	3.88	6.07	6.33
	$C=10$, $D=10$									
Datasets	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9	time	LB	UB
BC	98.53	97.07	5.99	30.00	70.00	70.00	70.00	0.30	464.77	490.97
PIMA	80.26	77.02	7.90	25.00	75.00	75.00	75.00	0.25	3738.23	3759.37
HEART	77.78	86.42	4.95	7.69	61.54	61.54	61.54	0.26	837.95	879.70
IONO	88.57	94.63	32.47	38.24	55.88	55.88	55.88	15.89	591.25	694.90
LIVER	92.86	75.57	5.37	60.00	80.00	80.00	80.00	0.05	799.69	815.94

As far as datasets of Group 1 are concerned, the Lagrangian relaxation approach provides in less than 10 sec upper bounds which are better than those obtained by CPLEX with the same time limit and very close to those related to the time limit of 1000 sec.

To finally assess how worth is the numerical effort needed to tackle our model in comparison with other methods for feature selection, we summarize in next Table 4.11 some results extracted from Tables 4.2, 4.5 and 4.6. In fact, as previously mentioned, the L_1-SVM method can be considered a tool for feature selection, which just requires solution of a Linear Program. For comparison purposes we add to the results extracted from Table 4.2 the corresponding columns $ft\,0$, $ft\,2$, $ft\,4$, $ft\,9$.

The comparison indicates that, while in general our method provides results less accurate in terms of classification correctness, it exhibits a definitely stronger ability to suppress non relevant features.

Table 4.11
L_1-SVM vs. Lagrangian Relaxation.

	L_1-SVM $C = 1$							Lagrangian relaxation $C = 1$, $D = 10$						
Datasets	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9
CARC	94.17	100	2.61	0,00	0.23	0.25	0.25	93.33	95.00	3.57	0.02	0.03	0.03	0.03
DLBCL	97.14	100	4.88	0.00	0.42	0.44	0.44	85.71	89.00	3.97	0.03	0.05	0.05	0.05
LEUK	95.00	100	4.43	0.00	0.54	0.58	0.58	97.14	99.00	6.32	0.03	0.19	0.19	0.19
TUM1	69.67	100	7.04	0.00	0.53	0.56	0.56	62.00	94.00	6.80	0.01	0.36	0.38	0.38
TUM2	81.50	100	4.97	0.00	0.23	0.25	0.25	76.50	93	5.96	0.02	0.05	0.05	0.05
	L_1-SVM $C = 10$							Lagrangian relaxation $C = 10$, $D = 10$						
Datasets	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9	Test	Train	$\|w\|_1$	ft 0	ft 2	ft 4	ft 9
BC	96.74	97.24	6.01	9.00	94.00	94.00	94,00	96.41	97.18	5.78	17.00	71.00	71.00	71.00
PIMA	76.30	77.62	8.30	26.25	100.00	100.00	100.00	76.01	77.42	8.02	26.25	87.50	87.50	87.50
HEART	84.37	86.83	7.27	12.31	100.00	100.00	100.00	83.95	86.79	6.71	12.31	82.31	82.31	82.31
IONO	88.91	95.46	42.47	50.59	95.00	95.59	95.59	87.93	95.18	34.31	40.00	67.65	67.65	67.65
LIVER	72.58	74.38	5.14	48.00	100.00	100.00	100.00	70.33	74.13	4.69	54.00	76.00	76.00	76.00

5. Conclusions

In this paper we have adopted a mixed binary formulation for feature selection in a SVM framework based on the use of the L_1 norm. We have mainly focussed on the performance of our Lagrangian relaxation approach.

The numerical examples we have worked indicate that the MBP model is able to provide an acceptable tradeoff between classification quality and number of relevant features. On the other hand the Lagrangian relaxation approach we have introduced appears to produce good quality solution at a quite affordable computational cost, and this is encouraging in view of possible application to very large datasets.

Finally we have discussed the performance of our method w.r.t. a well established method for feature selection.

Acknowledgments

We wish to cordially thank the Area Editor and three Reviewers whose constructive comments have been very useful to improve the paper.

The research of second author has been supported by the Interuniversity Attraction Poles Programme P7/36 “COMEX: combinatorial optimization metaheuristics & exact methods” of the Belgian Science Policy Office.

The research of fourth author has been partially supported by FEDER/Ministerio de Economía y Competitividad under grants MTM2013-46962-C02-02 and MTM2016-74983-C2-2-R.

Appendix A

Proof of Proposition (1).

Proof. Let (λ^*, μ^*) 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

$$u_{\bar{k}}(\hat{\lambda}_{\bar{k}} + \hat{\mu}_{\bar{k}}) - D = 0. \tag{A.1}$$

The setting:

$$\begin{cases} \hat{\lambda}_{\bar{k}} = \min\{\lambda^*_{\bar{k}}, D/u_{\bar{k}}\} \\ \hat{\mu}_{\bar{k}} = D/u_{\bar{k}} - \min\{\lambda^*_{\bar{k}}, D/u_{\bar{k}}\}, \end{cases} \tag{A.2}$$

satisfies condition (A.1), moreover, letting

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

we prove it is:

$$|(\hat{\lambda}_{\bar{k}} - \hat{\mu}_{\bar{k}}) - (\lambda^*_{\bar{k}} - \mu^*_{\bar{k}})| \leq \Delta_{\bar{k}}. \tag{A.3}$$

Consider in fact the two cases in (A.2):

- $\lambda^*_{\bar{k}} = \min\{\lambda^*_{\bar{k}}, D/u_{\bar{k}}\}$
  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
  $$D/u_{\bar{k}} - \lambda^*_{\bar{k}} \leq 0 \tag{A.4}$$
  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 (A.4) we obtain:
  $$\begin{aligned} -\Delta_{\bar{k}} &= D/u_{\bar{k}} - \lambda^*_{\bar{k}} - \mu^*_{\bar{k}} \leq D/u_{\bar{k}} - \lambda^*_{\bar{k}} + \mu^*_{\bar{k}} \\ &= D/u_{\bar{k}} - \lambda^*_{\bar{k}} + \Delta_{\bar{k}} - \lambda^*_{\bar{k}} + D/u_{\bar{k}} \\ &= \Delta_{\bar{k}} + 2(D/u_{\bar{k}} - \lambda^*_{\bar{k}}) \leq \Delta_{\bar{k}}, \end{aligned}$$
  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 A.3, the following holds:

$$\begin{aligned} z_1(\hat{\lambda}, \hat{\mu}) &= \|\hat{w}\| + C\left(\sum_{i=1}^{m_1} \hat{\xi}_i + \sum_{l=1}^{m_2} \hat{\zeta}_l\right) + \sum_{k=1}^{n} (\hat{\lambda_k} - \hat{\mu_k})\hat{w}_k \\ &= \|\hat{w}\| + C\left(\sum_{i=1}^{m_1} \hat{\xi}_i + \sum_{l=1}^{m_2} \hat{\zeta}_l\right) + \sum_{k \neq \bar{k}} (\lambda^*_k - \mu^*_k)\hat{w}_k \\ &\quad + (\hat{\lambda}_{\bar{k}} - \hat{\mu}_{\bar{k}})\hat{w}_{\bar{k}} + (\lambda^*_{\bar{k}} - \mu^*_{\bar{k}})\hat{w}_{\bar{k}} - (\lambda^*_{\bar{k}} - \mu^*_{\bar{k}})\hat{w}_{\bar{k}} \end{aligned}$$

$$= \|\hat{w}\| + C\left(\sum_{i=1}^{m_1}\hat{\xi}_i + \sum_{l=1}^{m_2}\hat{\zeta}_l\right) + \sum_{k=1}^{n}(\lambda_k^* - \mu_k^*)\hat{w}_k$$
$$+[(\hat{\lambda}_{\bar{k}} - \hat{\mu}_{\bar{k}}) - (\lambda_{\bar{k}}^* - \mu_{\bar{k}}^*)])\hat{w}_{\bar{k}} \geq z_1(\lambda^*, \mu^*)$$
$$+[(\hat{\lambda}_{\bar{k}} - \hat{\mu}_{\bar{k}}) - (\lambda_{\bar{k}}^* - \mu_{\bar{k}}^*)])\hat{w}_{\bar{k}} \geq z_1(\lambda^*, \mu^*)$$
$$-u_{\bar{k}}\Delta_{\bar{k}}, \tag{A.5}$$

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

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

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

$$\begin{cases} \hat{\lambda}_{\bar{k}} = \lambda_{\bar{k}}^* + \dfrac{\Gamma_{\bar{k}}}{2} \\ \hat{\mu}_{\bar{k}} = \mu_{\bar{k}}^* + \dfrac{\Gamma_{\bar{k}}}{2}, \end{cases} \tag{A.6}$$

Note that it is now

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

and thus, see A.5, 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 (λ, μ). □

References

Astorino, A., Frangioni, A., Fuduli, A., Gorgone, E., 2013. A nonmonotone proximal bundle method with (potentially) continuous step decisions. SIAM J. Optim. 23 (3), 1784–1809.

Astorino, A., Frangioni, A., Gaudioso, M., Gorgone, E., 2011. Piecewise quadratic approximations in convex numerical optimization. SIAM J. Optim. 21 (4), 1418–1438.

Aytug, H., 2015. Feature selection for support vctor machines using generalized benders decomposition. Eur. J. Oper. Res. 244, 210–218.

Bertolazzi, P., Felici, G., Festa, P., Fiscon, G., Weitschek, E., 2016. Integer programming models for feature selection: new extensions and a randomized solution algorithm. Eur. J. Oper. Res. 250 (2), 389–399.

Bi, J., Bennett, K., Embrechts, M., Breneman, K., Song, M., 2003. Dimensionality reduction via sparse support vector machines. J. Mach. Learn. Res. 3, 1229–1243.

Bradley, P., Mangasarian, O., 1998. Feature selection via concave minimization and support vector machines. In: Shavlik, J. (Ed.), Machine Learning proceedings of the Fifteenth International Conference (ICML98). Morgan Kaufmann, pp. 82–90.

Bradley, P., Mangasarian, O., Street, W., 1998. Feature selection via mathematical programming. INFORMS J. Comput. 10, 209–217.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM : a library for support vector machines. ACM Trans. Intell. Syst.Technol. 2 (27), 1:27.

Chen, Y., Lin, C.-J., 2006. Combining SVMS with various feature selection strategies. In: Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L. (Eds.), Feature Extraction, Foundations and Applications. Springer.

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

Carrizosa, E., Morales, D.R., 2013. Supervised classification and mathematical optimization. Comput. Oper. Res. 40 (1), 150–165.

Fortz, B., Gorgone, E., Papadimitriou, 2017. A Lagrangian heuristic algorithm for the time-dependent combined network design and routing problem. Networks 69 (1), 110–123.

Frangioni, A., 1996. Solving semidefinite quadratic problems within nonsmooth optimization algorithms. Comput. Oper. Res. 21, 1099–1118.

Frangioni, A., 2002. Generalized bundle methods. SIAM J. Optim. 13 (1), 117–156.

Frangioni, A., Gorgone, E., 2014. Generalized bundle methods for sum-functions with "easy" components: applications to multicommodity network design. Math. Program. 145 (1), 133–161.

Frangioni, A., Gorgone, E., Gendron, B., 2017. On the computational efficiency of subgradient methods: a case study with lagrangian bounds. Math. Programm. Comput.. To appear.

Gaudioso, M., Giallombardo, G., Miglionico, G., 2009. On solving the Lagrangian dual of integer programs via an incremental approach. Comput. Optim. Appl. 44, 117–138.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.

Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L., 2006. Feature extraction, foundations and applications. Springer, Berlin.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Mach. Learn. 46 (1–3), 389–422.

Hiriart-Urruty, J.-B., Lemaréchal, C., 1993. Convex analysis and minimization algorithms II—advanced theory and bundle methods. Grundlehren Math. Wiss., 306. Springer-Verlag, New York.

Kittler, J., 1986. Feature selection and extraction. In: Young, A. (Ed.), Handbook of Pattern Recognition and Image Processing. Academic Press, New York.

Lee, E., Tsung-Lin, W., 2009. Classification and disease prediction via mathematical programming. In: Pardalos, P., Romeijn, H. (Eds.), Handbook of Optimizationin Medicine. Springer Optimization and Its Applications 26.

Liu, Y., Wu, Y., 2007. Variable selection via a combination of the L_0 and L_1 penalties. J. Comput. Graphical Stat. 16 (4), 782–798.

Maldonado, S., Pérez, J., Weber, R., Labbé, M., 2014. Feature selection for support vector machines via mixed integer linear programming. Inf. Sci. 279, 163–175.

Mangasarian, O., 1997. Arbitrary-norm separating plane. Oper. Res. Lett. 24, 15–23.

Meyer, P.E., Schretter, C., Bontempi, G., 2008. Information-theoretic feature selection in microarray data using variable complementarity. IEEE Signal Process. Soc. 20, 261–274.

Nguyen, M.H., de la Torre, F., 2010. Optimal feature selection for support vector machines. Pattern Recognit. 43, 584–591.

Rinaldi, F., Sciandrone, M., 2010. Feature selection combining linear support vector machines and concave optimization. Optim. Methods Software 10 (1), 117–128.

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

Weston, J., Elisseeff, A., Schölkopf, B., Tipping, M., 2003. Use of the zero-norm with linear models and kernel methods. J. Mach. Learn. Res. 3, 1439–1461.

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., Vapnik, V., 2000. Feature selection for SVMs. Adv. Neural Inf. Process. Syst. 12, 668–674.