Cost-based feature selection for Support Vector Machines - An application in credit scoring

Sebastián Maldonado[a,*], Juan Pérez[a], Cristián Bravo[b]

[a] *Facultad de Ingeniería y Ciencias Aplicadas, Universidad de los Andes, Mons. Álvaro del Portillo 12455, Las Condes, Santiago, Chile.*
[b] *Southampton Business School, University of Southampton. University Road, SO17 1BJ, Southampton, United Kingdom.*

Abstract

In this work we propose two formulations based on Support Vector Machines for simultaneous classification and feature selection that explicitly incorporate attribute acquisition costs. This is a challenging task for two main reasons: the estimation of the acquisition costs is not straightforward and may depend on multivariate factors, and the inter-dependence between variables must be taken into account for the modelling process since companies usually acquire groups of related variables rather than acquiring them individually. Mixed-integer linear programming models are proposed for constructing classifiers that constrain acquisition costs while classifying adequately. Experimental results using credit scoring datasets demonstrate the effectiveness of our methods in terms of predictive performance at a low cost compared to well-known feature selection approaches.

Keywords: Analytics, Feature selection, Support Vector Machines, Mixed-integer Programming, Credit scoring.

1. Introduction

Support Vector Machine (SVM) [32] is a well-known machine learning tool for classification. Among existing methods, it provides important advantages, such as adequate generalization to new samples, absence of local minima, and a representation that depends on only a few parameters [16].

Feature selection is the process of choosing a subset of only relevant predictors for use in model estimation in order to increase stability, improve

[*] Corresponding author. Tel: +56-2-26181874. Email: smaldonado@uandes.cl

generalization power, and reduce overfitting [14]. The performance of classification methods, such as SVM, depends heavily on the proper choice of the feature set used to construct the classifier, especially in high-dimensional applications [13, 22]. In business analytics, however, two other goals of feature selection might be even more important than improving classification performance: gaining (managerial) insight into the process that generates the data, e.g. to understand the drivers that lead customers to leave a company or to default on a loan [20]; and reducing the variable acquisition costs in domains such as credit scoring, in which companies develop their models based on heterogeneous data sources.

Feature selection is an NP-hard problem that has been studied and reported extensively in the literature [13]. Most strategies propose the elimination of features independently of the classifier construction by exploiting statistical properties of each of the variables, or via greedy search [19]. All such strategies are heuristic by nature, and do not necessarily lead to models that optimize goals specified by the users or the organization they belong to.

Credit scoring corresponds to the use of statistical models to transform relevant data into numerical measures that guide credit decisions, and its main objective is to estimate the probability of default, i.e. the event of a customer not paying back the loan in a given time period [31]. These models are usually constructed based on historical data from the applicants to represent their creditworthiness, and also based on credit information from external sources, such as credit bureaus. For this work we used two credit scoring datasets from a Chilean bank. These data come from a previously developed project of small loans granted to micro-entrepreneurs [7].

In most business analytics applications, variables are grouped in such a way that if one attribute from a group is included in the model, then all the others in the same group are available at zero additional cost [8, 9]. For example, in credit scoring it is common for banks and other financial institutions to buy sets of variables from credit bureaus, and if one variable is used in the model, then the whole group of variables may be used for the modelling process.

The main contribution of this work is the incorporation of the variable acquisition costs in the feature selection procedure for a credit scoring project with a Chilean financial institution. Two novel MIP (Mixed-Integer Programming) approaches are proposed for this goal. These are based on the SVM principles of margin maximization that take into account the cost-based feature selection framework, while constructing linear classifiers. The data sets used in this work have six different groups of variables with different costs. Some attributes come from external sources, while others are combinations of the original variables from different groups in the form of financial ratios. We take advantage of this structure in our method, estimating the combination of data sources that are the most important for the model. Our approach is valid in applications where variables are acquired from different sources and at different costs. The estimation of the variable acquisition costs is also a novel contribution of this work; although it is specific for the credit assignment problem we faced and therefore cannot be extrapolated directly to other datasets.

The work is structured as follows: In Section 2 previous work on SVM classification is discussed. In Section 3 previous work on feature selection for SVM is described, including the formulations that are relevant in this work. The proposed mixed-integer linear programming methods for SVM classification and cost-based feature selection are introduced in Section 4. In Section 5 are provided experimental results using two credit scoring datasets. We present the main conclusions of this study in Section 6 and address future developments.

2. Support vector classification

In this section we briefly describe the standard l_2-SVM formulation [32], and two linear programming SVM extensions, namely the l_1-SVM formulation [6], and the LP-SVM method [35]. Both the l_1-SVM and the LP-SVM methods are used to develop a novel framework for cost-based feature selection, while standard SVM is used together with other well-known feature selection strategies as alternative approaches for benchmarking.

2.1. Standard l_2-SVM

Given training points $\mathbf{x}_i \in \Re^n$ with their respective class labels $y_i \in \{-1,+1\}$, $i = 1,\ldots,m$, SVM determines a hyperplane of the form $f(\mathbf{x}) = \mathbf{w}^\top\mathbf{x} + b$ that minimizes the classification errors, and at the same time maximizes the *margin*, which is computed as the distance between both (reduced) convex hulls. A set of slack variables $\boldsymbol{\xi}$, and a penalty parameter C that balances the trade-off between both objectives, are introduced. The methodology to calculate the parameter C is explained in Section 5.1. The soft-margin SVM formulation follows:

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

Formulation (1) is a convex quadratic programming problem that can be solved efficiently e.g. by the Sequential Minimal Optimization (SMO) algorithm [27].

2.2. l_1-Support Vector Machine

The traditional SVM formulation can be cast into a linear programming problem by replacing the use of the Euclidean norm as a regularizer with the l_1-norm or LASSO penalty [6]. The gain in doing so is twofold: on the one hand the complexity of the problem is reduced, and, on the other, the LASSO function performs embedded feature selection by reducing the number of non-zero components of the weight vector $\mathbf{w}$ [6].

The l_1-norm, however, cannot be used directly because the absolute value of $\mathbf{w}$ is nonsmooth. A set of positive variables $\bar{\mathbf{w}}$ needs to be introduced in order to cast this problem into a linear programming one. The l_1-SVM follows:

$$\begin{aligned} \min_{\mathbf{w},\bar{\mathbf{w}},b,\boldsymbol{\xi}} \quad & \sum_{j=1}^{n} \bar{w}_j + C\sum_{i=1}^{m} \xi_i \\ \text{s.t.} \quad & \forall_{i=1}^{m} : y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0; \\ & \forall_{j=1}^{n} : -\bar{w}_j \leq \ w_j \leq \bar{w}_j,\ \bar{w}_j \geq 0. \end{aligned} \tag{2}$$

2.3. Linear Programming Support Vector Machine

Another linear programming strategy based on SVM was proposed by Zhou et al. [35], in which the bound of the Vapnik-Chervonenkis (VC) dimension is relaxed properly using the l_∞-norm, resulting in an LP problem that maximizes a margin variable r. The LP-SVM method follows:

$$\begin{aligned} \min_{r,\mathbf{w},b,\boldsymbol{\xi}} \quad & -r + C\sum_{i=1}^{m} \xi_i \\ \text{s.t.} \quad & \forall_{i=1}^{m} : y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq r - \xi_i,\ \xi_i \geq 0; \\ & \forall_{j=1}^{n} : -1 \leq \ w_j \leq 1;\ r \geq 0. \end{aligned} \tag{3}$$

where $C \geq 0$ is a trade-off parameter that has a similar interpretation compared with the one presented in the standard SVM (Formulation (1)).

3. Feature selection for SVM

A plethora of feature selection methods has been proposed for SVM classification. Some strategies aim at eliminating poorly informative variables before applying the classification technique. One such approach is the Fisher Criterion Score (F), which computes each variable's contribution by comparing the means between both training patterns [13]:

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

where μ_j^+ (μ_j^-) is the mean of the j-th feature for class $+1$ (-1), and σ_j^+ (σ_j^-) is the respective standard deviation. The Fisher Score is simple to implement and performs quickly, but does not take the interactions between the variables and the classifier into account. The Fisher Score is one of the most intuitive best-known approaches in the literature to assess feature relevance. Alternative two-sample independence metrics, such as the KS and chi-squared tests, usually achieve similar results compared with the Fisher Score.

Another family of methods consists of search strategies that explore various subsets of variables, assessing them in terms of their performance. Since exhaustive search is usually intractable [13], greedy algorithms such as Sequential Forward Selection (SFS) and Sequential Backward Elimination (SBE) have been proposed in the literature [19]. Recursive Feature Elimination (RFE-SVM) [15] is a popular SBE method that tries to find a subset of relevant variables by eliminating those whose removal leads to the largest margin of class separation. Compared to the Fisher Score and other filter methods, RFE-SVM is computationally more demanding, but it does takes the interaction between the variables and the classifier into account, leading to a potentially better predictive performance [13].

Feature selection can also be a part of the optimization process used to construct the classifiers [34]. Such methods have the advantage of being computationally less intensive than search strategies [13]. Some MIP models have been proposed for this purpose. For example, an MIP for feature selection based on the assumption of feature independence was introduced in Iannarilli and Rubin [18]. Recently, Bertsimas et al. [5] proposed a MIP-based two-step algorithm for a linear regression problem. Their proposal finds good feasible solutions in the first stage, which are used as warm starts to a MIP problem in the second step.

Some methods have incorporated binary variables in the SVM classification problem. In Mangasarian and Wild [24], a kernel-based SVM classifier is proposed, in which a linear model is used to construct the hyperplane while the feature selection is performed by successive updates of the binary variables related to each attribute. In Carrizosa et al. [9], another MIP model was proposed for multi-class classification, in which bi-objective optimization was used to balance the trade-off between fit and feature elimination. The authors also propose a framework for addressing the variable acquisition cost minimization problem, which is also discussed below in our proposal. Finally, we extended the ideas of Carrizosa et al. [9] in Maldonado et al. [21], where two MIP models that included an additional budget constraint were proposed.

In this work we propose two novel mixed-integer linear programming models based on SVM, in which cost-based feature selection is performed by taking the variable acquisition costs into account. To the best of our knowledge, there are no studies devoted to the estimation of the variable acquisition costs in the context of SVM, and classification in general. In Carrizosa et al. [9], the authors used well-known data sets from the UCI Repository [3] for the experimental section, and simulated costs were studied. In Maldonado et al. [21], microarray datasets were studied, and all variables (genes) were treated independently, and with similar acquisition costs.

4. Proposed cost-based framework for feature selection and SVM classification

In this section, we extend two SVM formulations based on linear programming to MIP, namely, the l_1-SVM and the LP-SVM methods (Formulations (2) and (3)). We use these formulations instead of the classical l_2-SVM (Formulation (1)) because the latter has a higher complexity since it is a quadratic formulation instead of a linear one, assuming that the inclusion of binary variables causes a significant increase in running times due to the higher complexity. The proposed mixed-integer linear problems can be solved efficiently via state-of-the-art optimization tools, while a mixed-integer quadratic model would be much more expensive computationally.

Section 4.1 presents the description of the credit scoring project used in this paper, with a description of the datasets and the estimation of the variable acquisition costs for each group of variables. The datasets and their characteristics are presented before our proposal to enhance readability, and for a better understanding of the framework. Next, in Section 4.2, we present the framework used to incorporate variable acquisition costs into the SVM models. Subsequently, the two proposed MILP formulations for SVM are described in Section 4.3. Finally, in Section 4.4 there is a discussion of an additional issue, the class-imbalance problem, and the strategy we used to deal with skewed class distributions.

4.1. Dataset description and cost analysis

The dataset used in this paper consists of 7309 loans granted to small and micro companies, repaid in monthly instalments, granted by a local bank during the period 2004-2007. The dataset includes a total of 676 variables characterizing the loans, the borrowers, and the financial history of the borrower which is available for all returning customers. The dataset was split according to the credit history with the bank, resulting in two datasets: new customers (NC) with 1510 loans, and returning customers (RC) with 5799 loans. The numbers of defaulters are 629 and 872 for the new and returning customers, respectively, leading to imbalance ratios (IRs) of 1.4 and 5.65 for the new and returning customers, respectively. The IR is computed as the number of samples from the majority class divided by the number of samples from the minority class.

The objective variable for this problem follows the usual Basel II/III definition of default: one or more instalments in arrears for more than 90 days during the first year of the loan [4]. Following the methodology presented in Bravo et al. [7], a preprocessing step was applied to eliminate noisy and irrelevant attributes from the datasets: first, variables whose values have more than a 99% concentration in a single value, or more than 30% of missing values were discarded. Secondly, we used two-sample tests to discard poorly informative variables quickly by comparing whether the two groups (defaulters and non-defaulters) were independent or not. We used the Kolmogorov-Smirnov and χ^2 tests for numerical and nominal variables, respectively, where a p-value higher than 0.05 was used to remove irrelevant features. After preprocessing, the final datasets consisted of 94 and 46 variables for new customers and returning customers, respectively. The RC dataset has fewer variables than the NC because some variables were not captured in the evaluation process of the former dataset since it has a greater amount of historical information available.

Although datasets with fewer than one hundred variables may seem low-dimensional from a machine-learning perspective, it was of prime interest for the company to construct risk models with no more than 10 variables for two main reasons: first, they wanted models that could be understood easily in terms of the variables that conform to them in order to gain managerial insight into the customers and make better decisions; and second, to reduce variable acquisition costs, the main motivation for this study. In particular, they wanted to assess whether or not expensive internal processes for collecting and transforming information and data from external sources were truly explicative.

The following groups of variables were identified:

- Credit evaluation variables: These variables come from the application form that each potential borrower fills out. The application is then processed by the risk area of the bank, and entered into the company database. Since every borrower has to fill out one of these applications, acquiring this set of variables can be considered as a sunk cost. Each application has to be filled out in an office,where an executive is assigned to assist the potential customer. On average, each application takes one hour, so the estimated cost per processed application corresponds to €5 given the monthly salary of the executives (approximately €1000 per month, using the exchange rate on 4 April, 2016).

- In-depth interview: When there is little past credit information available for a given customer, the bank may choose to conduct an in-depth interview of the potential borrower. This requires a visit to the place of work of the potential borrower, which takes approximately four hours of the time of an executive. We estimate the acquisition cost for this group of variables as €20 per application.

- Financial analysis: If an in-depth interview is conducted, the bank may also choose to attempt to reconstructing the cash flow of the company. This is particularly relevant in micro-companies that are not required to keep detailed logs of their transactions, and as such, may not have cash flows readily available. This requires two hours of a specialist's time, for a total of €20 per loan, in addition to the €20 per interview. Also, some variables are calculated only if there is system-level information available.

- System-level information: The bank may also choose to acquire a database of the standing debts of the customers in the financial system. This information is provided on a monthly basis for all borrowers in the country, and requires a fixed cost of €1000. According to the policy of the bank, these variables are obtained only if there is no credit history available for the application.

In addition to all these previously mentioned costs, we consider a per-variable cost of €0.001 per application, which accounts for the application

Table 1: Groups of variables available per dataset.

Variable Group	NC	RC
Credit evaluation (g_1)	32	31
In-depth interview (g_2)	5	2
Financial analysis (g_3)	34	13
System-level information (g_4)	9	N/A
System-level + Financial analysis (g_5)	14	N/A

processing costs incurred by the company. This cost is not estimated like the other costs; it is sufficiently small value chosen to avoid any trade-off with the variable acquisition costs. Its goal is to allow the model to prefer solutions with few attributes when including the same groups of variables. Table 1 summarizes the available variables for each dataset and for each group.

In Table 1 we observe that most variables are obtained from the credit evaluation (group 1 or g_1) and the financial analysis (group 3 or g_3). The in-depth interview (group 2 or g_2) contributes few variables for both datasets, while system level information (group 4 or g_4) is available only for the new customers. A new group results from the construction of ratios generated as a combination of variables collected during the in-depth interview and system-level information (group 5 or g_5), which are only available for the new customers.

4.2. Framework for variable acquisition cost modelling

Following the literature in variable cost modelling [see e.g. 9, 26], given a set of selected variables $\mathcal{S}$, and a cost per variable of $c_j \; \forall \; j \in \mathcal{S}$, the total variable acquisition cost follows:

$$\pi = \sum_{j \in \mathcal{S}} c_j \tag{5}$$

The main issue with the previous definition is that linearity is assumed, which is not realistic in our case. The concept of groups of variables leads to *precedence relations* between them, that is, if one variable of the group is selected, then all the other variables from the group can be included in the model with zero cost (or at the application processing cost). Formally, if the use of an attribute b requires the inclusion of an attribute a, then a partial order relation $a \preceq b$ is given [9]. Equation (5) can be redefined by introducing an auxiliary variable $z_j \in \{0, 1\}, \; j = 1, \ldots, n$, where $z_j = 1$ if a

Table 2: Precedence relations for each group of variables.

Group	Evaluation	Interview	Fin. anal.	System info.
g_1	✓			
g_2	✓	✓		
g_3	✓	✓	✓	
g_4	✓			✓
g_5	✓	✓	✓	✓

payment for variable j is required, i.e. if $k \in \mathcal{S}$ for any k such as $k \preceq j$, and $z_j = 0$ otherwise. The redefined variable acquisition cost equation follows:

$$\pi = \sum_{j=1}^{n} c_j z_j \tag{6}$$

The previous approach is suitable for both types of precedence relations we face in our credit scoring project. First, there are precedence relations between groups, since, for example, an in-depth interview can be conducted only after a credit evaluation, and therefore at least one variable from the latter group should be included. Alternatively, ratios between system-level information and credit evaluation variables can be performed via the financial analysis (and its respective cost), but also require the payment for both sources of information: system-level and evaluation variables. Secondly, if one variable from a group is included, all the others from the same group can be included by paying only the application processing cost. This is equivalent to defining the cost of one of the variables as the acquisition cost for the whole group, defining precedence relations of the form $a \prec b$ if and only if attributes a and b belong to the same group and a has no additional cost.

Next, we formalize the precedence relations for the five available groups defined in Table 1. These relations are presented in Table 2.

We propose the following framework for modelling the cost analysis described above: two sets of binary decision variables are introduced: a vector $\mathbf{v}$ of size n that indicates the selection of attributes, and a binary vector $\mathbf{g}$ of size $K = 5$ that represents the groups of attributes. We define as parameters the per-variable cost VC_j with $j = 1, \ldots, n$; the per-group costs GC_k with $k = 1, \ldots, K$; and a monetary budget B designed to constrain the variable acquisition costs. This constraint can be formulated as follows:

$$\sum_{j=1}^{n} VC_j v_j + \sum_{k=1}^{K} GC_k g_k \leq B \tag{7}$$

Precedence relations can be defined as constraints under this framework. For example, group 2 also requires the payment for group 1, i.e. $g_1 \prec g_2$; which translates into $g_1 \geq g_2$. All precedence relations can be written compactly as

$$\forall_{k,k' \in \{1,\ldots,K\} | k > k'} : \; g_k \geq g_{k'}. \tag{8}$$

Additionally, the relationships between both sets of variables need to be defined in the model. For each group of variables, the attributes within the group can be selected only if the group is activated, or

$$\forall_{k=1}^{K} : \; M_1 g_k \geq \sum_{j \in \mathcal{G}_k} v_j, \tag{9}$$

where $M_1 >> 0$ should be at least the cardinality of $\mathcal{G}_k$, the set of variables that belongs to the group k, with $k = 1, \ldots, K$. Next, the two proposed MILP models using this framework are proposed.

4.3. Proposed cost-based MILP formulations

In this section we incorporate the acquisition cost framework that can be summarized as equations (7), (8), and (9), into two novel MILP formulations based on the structural risk minimization principle used in SVM classification.

The first formulation is an extension of l_1-SVM (Formulation (2)). It is basically the same formulation but includes the three constraints mentioned above, plus two extra sets of constraints that relate $\bar{\mathbf{w}}$, the positive variable related to the absolute values of the weights, and $\mathbf{v}$, our indicator variable for feature selection. If a given attribute j is activated, i.e. $v_j = 1$, then $\bar{w} > 0$. In contrast, if j is not activated, i.e. $v_j = 0$, then $\bar{w} = 0$. Thus we include the constraints $M_2 \bar{w}_j \geq v_j$ and $M_3 v_j \geq \bar{w}_j$ for all $j = 1, \ldots, n.$, where $M_2, \; M_3 >> 0$. We refer to this formulation as l_1 Mixed-Integer SVM, or simply l_1-MISVM.

$$
\begin{aligned}
\min_{\mathbf{w},\mathbf{g},\bar{\mathbf{w}},\mathbf{v},b,\boldsymbol{\xi}} \quad & \sum_{j=1}^{n} \bar{w}_j + C\sum_{i=1}^{m} \xi_i \\
\text{s.t.} \quad & \forall_{i=1}^{m} : y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0; \\
& \forall_{k=1}^{K} :\ M_1 g_k \geq \sum_{j \in \mathcal{G}_k} v_j,\ g_k \in \{0,1\}; \\
& \forall_{k,k' \in \{1,\ldots,K\} | k>k'} :\ g_k \geq g_{k'}; \\
& \forall_{j=1}^{n} : -\bar{w}_j \leq\ w_j \leq \bar{w}_j,\ M_2\bar{w}_j \geq v_j,\ M_3 v_j \geq \bar{w}_j; \\
& \forall_{j=1}^{n} : \bar{w}_j \geq 0,\ v_j \in \{0,1\}; \\
& \sum_{j=1}^{n} VC_j v_j + \sum_{k=1}^{K} GC_k g_k \leq B.
\end{aligned}
\tag{10}
$$

The second formulation extends the LP-SVM model (Formulation (3)), and combines this LP model with the cost framework by including variables **v** and **g**, and the three previously mentioned constraints (Eqs.(7), (8), and (9)). The constraints in LP-SVM that bound the weights between -1 and 1 are redefined as $-v_j \leq\ w_j \leq v_j$ in order to incorporate the relation between **v** and **w**. We refer to this formulation as LP Mixed-Integer SVM, or simply LP-MISVM.

$$
\begin{aligned}
\min_{\mathbf{w},\mathbf{g},\mathbf{v},b,\boldsymbol{\xi}} \quad & -r + C\sum_{i=1}^{m} \xi_i \\
\text{s.t.} \quad & \forall_{i=1}^{m} : y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq r - \xi_i,\ \xi_i \geq 0; \\
& \forall_{k=1}^{K} :\ M_1 g_k \geq \sum_{j \in \mathcal{G}_k} v_j,\ g_k \in \{0,1\}; \\
& \forall_{k,k' \in \{1,\ldots,K\} | k>k'} :\ g_k \geq g_{k'}; \\
& \forall_{j=1}^{n} : -v_j \leq\ w_j \leq v_j,\ v_j \in \{0,1\}; \\
& \sum_{j=1}^{n} VC_j v_j + \sum_{k=1}^{K} GC_k g_k \leq B.
\end{aligned}
\tag{11}
$$

Formulations (10) and (11) were solved by using a Branch and-Cut strategy for the instances of a Chilean financial institution analyzed in this paper.

4.4. The class-imbalance problem

One additional challenge is the class-imbalance problem. This issue arises when the class distribution is significantly skewed [17]. Imbalance ratios of 5:1 or higher are common in credit scoring since most risk models are created based on rules used to reject those applicants who are most likely to default, being the applicants accepted who do not repay their loans, strongly underrepresented in the dataset [31].

The class-imbalance problem affects SVM negatively since it usually constructs a classifier that assigns all data points to the majority class, i.e. predicts that all applicants will not default and no risk model is needed. Since accepting a defaulter usually has a higher cost for the financial institutions than rejecting a good borrower [1], this outcome is far from ideal.

There is strong evidence that suggests the need for balancing credit scoring samples [11], so we used two well-known resampling approaches to deal with this issue. Resampling consists of adjusting the imbalance ratio of the training set artificially by either discarding samples from the majority class, downsizing it, or by generating new samples from the minority class. The first resampling approach is known as *undersampling*, and is usually performed randomly, while the second strategy is known as *oversampling*. Arguably the most popular oversampling method is called the Synthetic Minority Over-sampling Technique (SMOTE) [10] in which new data points are created artificially by interpolating the pre-existing pairs of samples from the minority class. The techniques we used have been proven to give good results in credit scoring by Marques et al. [25].

Both undersampling and oversampling have advantages and disadvantages. On the one hand, undersampling may lead to a loss of relevant data if too many points are discarded for the sake of a balanced training set. On the other hand, oversampling is prone to overfitting, and it may increase the size of the training set significantly, causing longer running times [17].

In our work we explore the following two approaches in the dataset that presents class-imbalance (Returning Customers, IR=5.65): random undersampling until perfect class balance; and 200% SMOTE oversampling, i.e. the minority class is doubled via the generation of new, artificially generated points, and then random undersampling is used until perfect class balance is achieved. Resampling until perfect balance is a well-known strategy for dealing with the class-imbalance problem in business applications [see e.g. 33].

Besides data resampling, there are various techniques designed to be trained from imbalanced data sets without data resampling. Cost-sensitive approaches, for example, consider different misclassification costs to represent the fact that the cost of misclassifying a minority class sample is usually higher than the one of misclassifying a majority class instance. For example, to the decision threshold can be adjusted in order to favour the minority class [17]. For SVM, the expression that controls the model fit $C\sum_{i=1}^{m}\xi_i$ can be split in two terms, $C_+\sum_{i\in I^+}\xi_i$ and $C_-\sum_{i\in I^-}\xi_i$, where I^+ (I^-) is the set of positive(negative) samples, and $C_+ > C_-$ are two trade-off parameters [2].

Alternatively, one-class techniques can be used for dealing with the class-imbalance problem. Originally developed for outlier detection, one-class classification aims at constructing classifiers by learning from a training set containing only the objects of one of the classes. One technique is Support Vector Data Description (SVDD), which finds the smallest sphere of radius R that contains most of the data points [30].

5. Experimental results

In this section we provide the classification results using the two proposed methods (l_1-MISVM and LP-MISVM), and the two alternative approaches for feature selection and SVM classification that are described in Section 3: the Fisher Score as a filter strategy for feature ranking using standard SVM as baseline classifier, and RFE-SVM.

5.1. Experimental settings and summary of results

For these approaches, the first step of the experimental setting is model selection for linear SVM without feature selection. Parameter C was tuned using 10-fold cross-validation, and we explored the following set of parameters applying line search, based on previous research [23]:
$C \in \{2^{-7}, 2^{-6}, ..., 2^{-1}, 2^{0}, 2^{1}, ..., 2^{6}, 2^{7}\}$.

Feature selection was performed on the training set for each of the folds, using the best parameter C found in the previous step, where the Area Under the Curve (AUC) was used as the performance metric. The Receiver Operating Characteristic (ROC) curve is a graphical representation of the true positive rate against the false negative rate at various discrimination threshold settings, and the AUC is simply the area under this curve. This measure provides an adequate balance between the true positive and the true negative rates, being more suitable than the overall accuracy for assessing the performance of binary classification problems when facing class-imbalanced datasets [29]. For the Fisher Score and RFE-SVM, we explored the classification performance for an increasing number of ranked features:

$n = \{5, 10, 20, 30, 40, 50, 60, 70, 80, 90\}$ and $n = \{5, 10, 20, 30, 40\}$ for the new and returning customers, respectively.

Regarding our proposals, we performed a similar line search for parameter C using 10-fold cross-validation. Since the financial institution did not have a precise value for the budget, we varied this parameter in order to obtain different solutions, and to assess the performance of the proposed methods as a function of the number of variables, and the total cost.

Tables 3 and 4 summarize the results for the new and returning customers, respectively. These tables present the best performance in terms of AUC among all subsets of features, the selected number of variables, and the total cost.

Table 3: Performance summary for different feature selection approaches. New customers.

	AUC	n	Cost [€]
Fisher+SVM	70.6	50	57,320
RFE-SVM	69.4	90	57,320
l_1-MISVM	70.4	13	57,320
LP-MISVM	69.5	13	6,189

Table 4: Performance summary for different feature selection approaches. Returning customers.

	Undersampling			Under & Oversampling		
	AUC	n	Cost [€]	AUC	n	Cost [€]
Fisher+SVM	67.3	30	219,831	67.5	40	219,831
RFE-SVM	63.6	40	219,831	64.1	40	219,831
l_1-MISVM	67.0	31	29,443	67.6	31	29,443
LP-MISVM	66.2	32	122,678	67.8	32	122,678

In Tables 3 and 4 we first notice that no method seems to outperform the others for these datasets. The Fisher Score performs slightly better than our approaches, but uses 50 attributes instead of the 13 selected by our proposals, and includes all variable groups. If we are willing to sacrifice one percentage point in AUC, we can reduce the acquisition costs significantly (from €57,320 to €6,189). We provide further details regarding this trade-off between cost and performance at the end of this section. On the other hand, the best average performance is achieved with our proposals for the returning customers and at a significantly lower variable acquisition cost (one tenth of the total costs for l_1-MISVM). We also observe that the best performance for each method is reached when using the combination of under- and oversampling (SMOTE) instead of random undersampling as the sole resampling strategy.

To support our previous analysis, the Holm's test is used to identify if any of the methods outperform others statistically, as is recommended by Demšar [12]. This test computes a Z statistic based on the average ranks for each technique. The best approach (the one with with the lowest mean rank) is set as the baseline, and then pairwise comparisons are performed between this method and the other techniques. The results for this test are reported in Table 5. From this analysis we conclude that no approach outperforms the others for these two datasets, being LP-MISVM the approach with the best average performance.

Method	Mean Rank	Mean AUC	p value	$\alpha/(k-i)$	Action
LP-MISVM	1.50	68.65	-	-	-
Fisher	2.00	69.05	0.70	0.05	not reject
l_1-MISVM	2.75	68.50	0.33	0.025	not reject
RFE-SVM	3.75	66.75	0.08	0.016	not reject

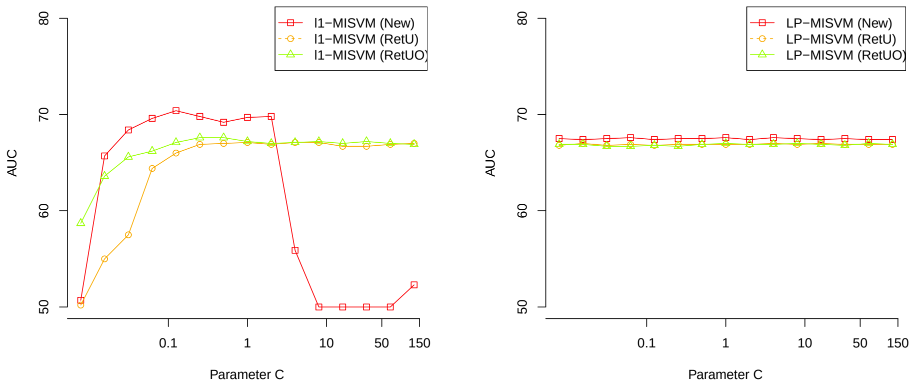
Table 5: Holm's test for pairwise comparisons between methods.

The proposed approaches are MILP implementations, which are known to be more time-consuming than linear and quadratic programming. According to our results, our proposals achieved tractable running times. For the dataset with new customers, the mean running times were 0.72 and 1.01 seconds for l_1-MISVM and LP-MISVM, respectively; while the mean running times for the dataset with returning customers were 5.07 and 2.49 seconds for l_1-MISVM and LP-MISVM, respectively. These values were obtained by averaging all running times for different folds on a laptop with 16 GB RAM, i7-6650U processor with 2.20 GHz, and using Microsoft Windows 10 Operating System (64-bits).

In terms of computational complexity, the MI problems solved by l_1-MISVM consisted in 1798 decision variables (99 binary variables) and 1912 constraints for the dataset with new customers, and 5943 decision variables (51 binary variables) and 6009 constraints for the dataset with returning customers. For LP-MISVM, the MI problems consisted in 1705 decision variables (99 binary variables) and 1724 constraints for the dataset with new customers, and 5898 decision variables (51 binary variables) and 5917 constraints for the dataset with returning customers.

5.2. *Sensitivity analysis for parameter C*

Next, we analyze the influence of the parameter C on the performance of the proposed methods. On the one hand, stable results would show robustness in terms of performance, which is a desirable feature in machine learning. And on the other hand, a strong influence of this parameter in the final solution would suggest that the model calibration procedure used in this work is highly recommended in order to achieve adequate performance. In Figures 1(a) and 1(b) we report the AUC as a function of C for l_1-MISVM and LP-MISVM, respectively. The three curves presented in each plot represent the three datasets: new customers, returning customers with random undersampling as the resampling technique (RetU), and returning customers with the combination of undersampling and SMOTE oversampling as the resampling technique (RetUO).



(a) l_1-MISVM

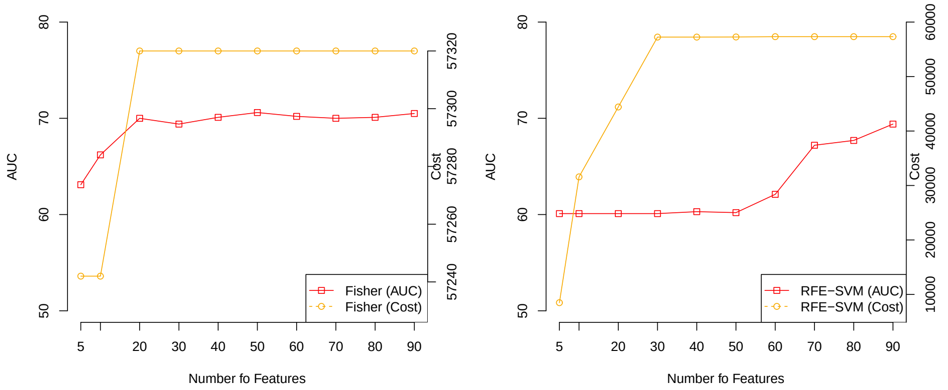
(b) LP-MISVM

Figure 1: Sensitivity analiysis for parameter C. Proposed methods.

In Figure 1(a) we observe that l_1-MISVM shows relatively stable results when $0.1 \leq C \leq 1$ (the highest AUC), but performance decreases significantly when $C \leq 0.1$ or $C \geq 1$. In contrast, the LP-MISVM method achieves very stable performance for all C values (See Figure 1(b)). We conclude from these experiments that, although the results are relatively stable in terms of performance for the different values of C, an adequate validation step based on line search is strongly recommended, as is suggested in the SVM literature [16].

5.3. Trade-off between performance and acquisition costs

A deeper analysis of the trade-off between predictive performance and variable acquisition costs is reported here. First, we report the AUC and the total variable acquisition cost for an increasing number of ranked attributes for the new customers: Figure 2(a) for the Fisher Score and Figure 2(b) for the RFE-SVM method. Then, this is shown for the returning customers: Figure 3(a) for the Fisher Score and Figure 3(b) for RFE-SVM. The results reported for the latter dataset consider only the use of the combination of undersampling and SMOTE oversampling as the resampling technique; the results obtained when using random undersampling as the sole resampling technique were omitted in this analysis since they are always outperformed by the combined strategy.

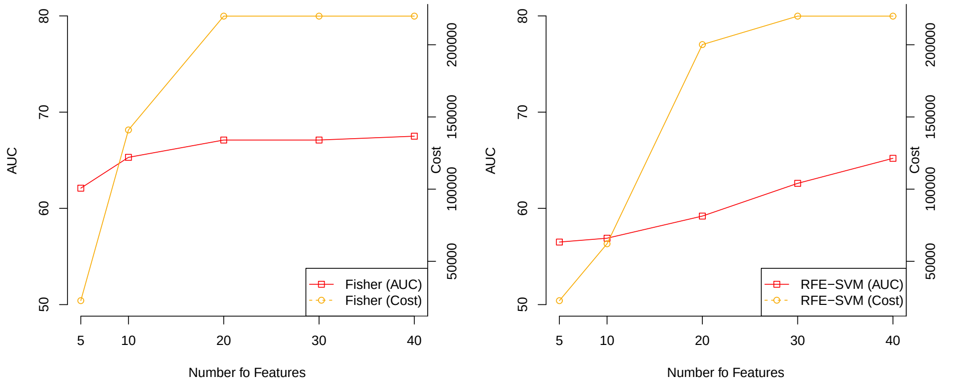


(a) Fisher+SVM

(b) RFE-SVM

Figure 2: Performance and cost for an increasing number of attributes. New Customers. Alternative feature selection methods.

In Figures 2 and 3 we observe similar patterns for both datasets and methods: in all cases both the AUC and the total cost remain stable and high when using around 20% of the attributes or more, but performance decreases significantly when using only 5 or 10 variables. This drop in terms of AUC also implies lower costs, but the performance gap is too high to accept these low-dimensional classifiers as valid candidates for implementation. We also observe that the Fisher Score performs better than the RFE-SVM on these datasets. We conclude from these experiments that, on the one hand,



(a) Fisher+SVM

(b) RFE-SVM

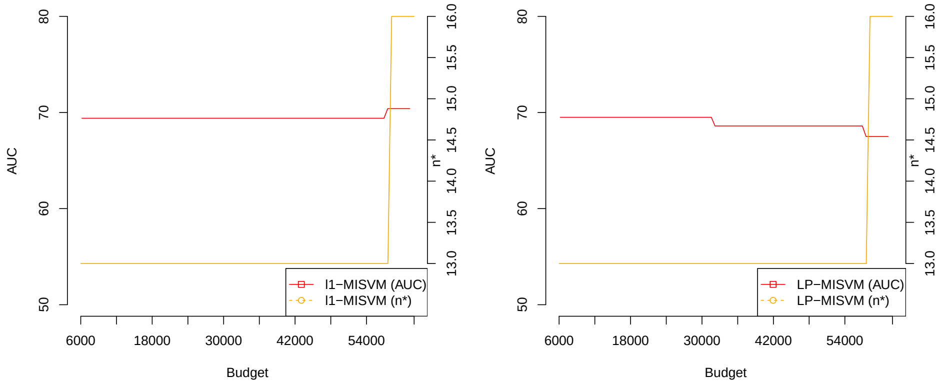
Figure 3: Performance and cost for an increasing number of attributes. Returning Customers, Under- and Oversampling. Alternative feature selection methods.

feature selection methods like the Fisher Score and RFE-SVM are successful at identifying irrelevant attributes correctly, leading to good predictive performances even with only 20% of the original variables, but they fail at constructing low-dimensional classifiers with both good predictive performance and low variable acquisition costs.

In contrast to feature ranking approaches, our proposals find the optimal number of selected attributes automatically for a given budget. This is a desirable attribute in machine learning, since it avoids additional validation steps to determine it [22]. In order to make both approaches comparable, we report the AUC and the number of selected variables for an increasing value of the budget parameter B for the new customers (Figure 4(a) for l_1-MISVM and Figure 4(b) for LP-MISVM), and for the returning customers (Figure 5(a) for l_1-MISVM and Figure 5(b) for LP-MISVM).

In Figures 4 and 5 we also observe similar patterns for both datasets and methods: classification performance (AUC) remains very steady while varying the budget parameter, and very good solutions can be achieved with few attributes, and, most importantly, at a very low cost. Only a few different solutions are obtained for each model, showing robustness. For each figure we observe the following results:

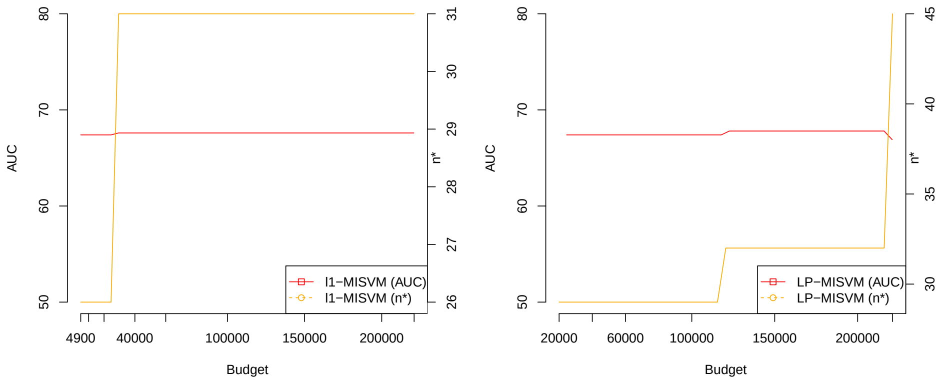
- For l_1-MISVM, new customers (Figure 4(a)), we observe two valid so-



(a) l_1-MISVM

(b) LP-MISVM

Figure 4: Influence of the budget parameter. New Customers.



(a) l_1-MISVM

(b) LP-MISVM

Figure 5: Influence of the budget parameter. Returning Customers, Under- and Oversampling.

lutions, using 13 attributes with an AUC of 69.4 and a cost of €6,189, or using 16 attributes with an AUC of 70.4 and a cost of €57,320. Both solutions are sparse (only around one tenth of the attributes were selected), while a decision-maker that ignores the cost would probably prefer the solution with 16 variables. The inclusion of the cost information in the modelling leads to a solution that is 1% worse in terms of AUC, but reduces the variable acquisition costs up to one tenth by removing three expensive features. A decision-maker that takes this information into account would probably prefer this second alternative.

- For LP-MISVM, new customers (Figure 4(b)), we observe that the cheapest solution (13 selected attributes, AUC of 69.5, and a cost of €6,189) is also the best one in terms of performance. This result is very similar to the one obtained by l_1-MISVM for this dataset.
- For l_1-MISVM, returning customers (Figure 5(a)), we again observe two valid solutions: using 26 attributes with an AUC of 67.4 and a cost of €4,907, or using 31 attributes with an AUC of 67.6 and a cost of €29,443. For this case, if a loss of 0.2 in performance leads to a reduction of the variable acquisition cost to one sixth, a decision-maker would most likely prefer the cheapest alternative. Compared to the alternative approaches, the proposed method achieves similar performance compared to the best strategy, while reducing the costs up to one tenth.
- For LP-MISVM, returning customers (Figure 5(b)), we also observe two valid solutions: using 29 attributes with an AUC of 67.4 and a cost of €24,536, or using 32 attributes with an AUC of 67.8 and a cost of €122,678. For this case, the decision-maker has to trade-off a sacrifice in performance of 0.4 in order to reduce the variable acquisition cost to one fifth.

From these experiments we conclude that our proposal achieves similar predictive performance compared with the best alternative feature selection model (the Fisher Score) while selecting fewer attributes and, most importantly, reducing the variable acquisition costs by about one tenth. Alternative methods failed at finding cheap solutions with adequate predictive performance. A comparison between l_1-MISVM and LP-MISVM shows that both methods present similar performance in terms of AUC and cost.

6. Conclusions

In this work we have presented two SVM-based strategies for simultaneous classification and embedded feature selection. The identification of relevant attributes was achieved through the introduction of binary variables, while a budget constraint was introduced in order to find accurate solutions at a low variable acquisition cost. These costs were estimated for a credit scoring problem: two datasets of loans granted to small and micro companies by a Chilean bank.

A comparison between our proposals and other feature selection methods for SVM in these real-world datasets showed the advantages of the proposed approaches. First, they achieved similar or better performance compared to the best alternative approach, using fewer attributes and with a total variable cost of about one tenth. Secondly, they identified the optimal subset of attributes automatically for a given budget, avoiding additional calibration steps required for feature ranking methods. Additionally, solving the feature selection problem while constructing the classifier takes all variable interactions and the relationship between those and the classifiers into account, leading to best predictive performance and a robust feature selection scheme. Finally, the linearity of our models allows the use of our proposal in practice for credit risk modelling, and it is easily extrapolated to other applications within business analytics.

From the results of our experiments we can conclude that the proposals are stable in terms of performance for different values of the parameters C and B, the budget for variable acquisition costs. A sensitivity analysis for the latter suggests that different solutions can be obtained in terms of cost and predictive performance, allowing the decision-maker to choose between several alternatives according to this trade-off. Finally, both the l_1-MISVM and LP-MISVM methods achieved relatively similar results, both being excellent alternatives for SVM classification, even without the need of estimating the variable acquisition costs.

Interestingly, predictive results are very similar between new and returning applicants. Although the returning customers are better at repaying (85% good applicants compared with 58% for the new customers), their behaviour is not easier to predict. This can be due to the fact that more variables for external sources are collected for the new applicants, compensating the lack of behavioural variables. Predictive performance in general is also rather low compared to other credit scoring studies. A reason for this could be that the applicants are micro-entrepreneurs, which is a riskier group that receives loans with high interest rates. Traditional credit scoring variables like income are no longer relevant for micro-entrepreneurs, resulting in less accurate predictive models [7].

Future work can be carried out in several directions. First, it would be interesting to apply these methods to other application domains, such as medicine and biotechnology. For example, the detection of respiratory diseases like the Obstructive Sleep Apnea Syndrome requires data in the form of signals from different sources, such as an electrocardiogram, or a nasal airflow sensor, and each of these tests has different costs [28]. Additionally, other areas within business analytics besides credit risk could be studied using these approaches, such as churn prediction, or fraud detection. Another possible future development is the extension of our proposal to include profit-based measures. The trade-off between a less accurate solution but cheaper in terms of variable acquisition costs can be better assessed under a cost-benefit setting by computing the total profit of the solution instead of the AUC. Finally, the proposed methods can be extended to kernel methods, which may lead to better predictive performance in domains that are not regulated in terms of the type of models that can be used, such as credit assignment.

Acknowledgments

The first author was supported by FONDECYT project 1160738. The second author was funded by FONDECYT project 11160320. This research was partially funded by the Complex Engineering Systems Institute, ISCI (ICM-FIC: P05-004-F, CONICYT: FB0816). The authors are grateful to the anonymous reviewers, who contributed to improving the quality of the paper.

References

[1] R. Anderson, The Credit Scoring Toolkit, Oxford University Press, 2007.

[2] F. Bach, D. Heckerman, E. Horvitz, Considering cost asymmetry in learning classifiers, Journal of Machine Learning Research 7 (2006) 1713–1741.

[3] K. Bache, M. Lichman, UCI machine learning repository, 2013. URL: http://archive.ics.uci.edu/ml.

[4] Basel Committee on Banking Supervision, Basel II: International convergence of capital measurement and capital standards: A revised framework - comprehensive version, 2006. URL: http://www.bis.org/publ/bcbsca.htm.

[5] D. Bertsimas, A. King, R. Mazumder, Best subset selection via a modern optimization lens, The Annals of Statistics 44 (2016) 813–852.

[6] P. Bradley, O. Mangasarian, Feature selection vía concave minimization and support vector machines, in: Machine Learning proceedings of the fifteenth International Conference (ICML'98) 82-90, San Francisco, California, Morgan Kaufmann.

[7] C. Bravo, S. Maldonado, R. Weber, Methodologies for granting and managing loans for micro-entrepreneurs: New developments and practical experiences, European Journal of Operational Research 227 (2013) 358–366.

[8] E. Carrizosa, B. Martín-Barragán, R.M. D., Detecting relevant variables and interactions in supervised classification, European Journal of Operational Research 213 (2011) 260–269.

[9] E. Carrizosa, B. Martín-Barragán, D. Romero-Morales, Multi-group support vector machines with measurement costs: A biobjective approach, Discrete Applied Mathemathics 156 (2008) 950–966.

[10] N.V. Chawla, L. Hall, K. Bowyer, W. Kegelmeyer, Smote: Synthetic minority oversampling technique, Journal of Artificial Intelligence Research 16 (2002) 321–357.

[11] S.F. Crone, S. Finlay, Instance sampling in credit scoring: An empirical study of sample size and balancing, International Journal of Forecasting 28 (2012) 224 – 238.

[12] J. Demšar, Statistical comparisons of classifiers over multiple data set, Journal of Machine Learning Research (2006) 1–30.

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

[14] I. Guyon, A. Saffari, G. Dror, G. Cawley, Model selection: Beyond the bayesian frequentist divide, Journal of Machine Learning Research 11 (2009) 61–87.

[15] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Machine Learning 46 (2002) 389–422.

[16] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, second ed., Springer-Verlag, 2009.

[17] H. He, E. García, Learning from imbalanced data, IEEE Transactions on Knowledge and Data Engineering 21 (2009) 1263–1284.

[18] F.J. Iannarilli, P.A. Rubin, Feature selection for multiclass discrimination via mixed-integer linear programming, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (2003) 779–783.

[19] J. Kittler, Pattern recognition and signal processing, Pattern Recognition and Signal Processing, Sijthoff and Noordhoff, Netherlands, 1978, pp. 41–60.

[20] S. Maldonado, A. Flores, T. Verbraken, B. Baesens, R. Weber, Profit-based feature selection using support vector machines - general framework and an application for customer churn prediction, Applied Soft Computing 35 (2015) 740–748.

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

[22] S. Maldonado, R. Weber, A wrapper method for feature selection using support vector machines, Information Sciences 179 (2009) 2208–2217.

[23] S. Maldonado, R. Weber, J. Basak, Kernel-penalized SVM for feature selection, Information Sciences 181 (2011) 115–128.

[24] O.L. Mangasarian, E.W. Wild, Feature selection for nonlinear kernel support vector machines, in: Seventh IEEE International Conference on Data Mining, IEEE, Omaha, NE, 2007, pp. 231–236.

[25] A.I. Marques, V. Garcia, J.S. Sanchez, On the suitability of resampling techniques, J Oper Res Soc 64 (2013) 1060–1070.

[26] P. Paclik, R. Duin, G. van Kempen, R. Kohlus, Structural, syntactic, and statistical pattern recognition. lecture notes in computer science, Structural, Syntactic, and Statistical Pattern Recognition. Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2002, pp. 461–469.

[27] J. Platt, Advances in kernel methods-support vector learning, Advances in Kernel Methods-Support Vector Learning, MIT Press, Cambridge, MA, 1999, pp. 185–208.

[28] S. Ríos, L. Erazo, An automatic apnea screening algorithm for children, Expert Systems with Applications 48 (2016) 42–54.

[29] M. Sokolova, N. Japkowicz, S. Szpakowicz, Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation, in: Advances in Artificial Intelligence, Springer, Berlin Heidelberg, 1015-1021, 2006.

[30] D.M.J. Tax, R. Duin, Support vector data description, Machine Learning 54 (2004) 45–66.

[31] L. Thomas, J. Crook, D. Edelman, Credit Scoring and its Applications, SIAM, 2002.

[32] V. Vapnik, Statistical Learning Theory, John Wiley and Sons, 1998.

[33] T. Verbraken, C. Bravo, R. Weber, B. Baesens, Development and application of consumer credit scoring models using profit-based classification measures, European Journal of Operational Research 238 (2014) 505–513.

[34] J. Weston, S. Mukherjee, O. Chapelle, M. Ponntil, T. Poggio, V. Vapnik, Feature selection for SVMs, Advances in Neural Information Processing Systems, in: Advances in Neural Information Processing Systems 13, volume 13.

[35] W. Zhou, L. Zhang, L. Jiao, Linear programming support vector machines, Pattern Recognition 35 (2002) 2927–2936.