Decision Support Systems 104 (2017) 113–121

Contents lists available at ScienceDirect

Decision Support Systems

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

# Integrated framework for profit-based feature selection and SVM classification in credit scoring

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

[a] Facultad de Ingeniería y Ciencias Aplicadas, Universidad de los Andes, Monseñor Álvaro del Portillo 12455, Las Condes, Santiago, Chile
[b] Department of Decision Analytics and Risk, University of Southampton, University Road, SO17 1BJ Southampton, United Kingdom
[c] Facultad de Ingeniería y Ciencias, Universidad Diego Portales, Ejército 441, Santiago, Chile

ARTICLE INFO

*Article history:*
Received 27 May 2017
Received in revised form 18 August 2017
Accepted 15 October 2017
Available online 18 October 2017

*Keywords:*
Profit measure
Group penalty
Credit scoring
Support Vector Machines
Analytics

ABSTRACT

In this paper, we propose a profit-driven approach for classifier construction and simultaneous variable selection based on linear Support Vector Machines. The main goal is to incorporate business-related information such as the variable acquisition costs, the Types I and II error costs, and the profit generated by correctly classified instances, into the modeling process. Our proposal incorporates a group penalty function in the SVM formulation in order to penalize the variables simultaneously that belong to the same group, assuming that companies often acquire groups of related variables for a given cost rather than acquiring them individually. The proposed framework was studied in a credit scoring problem for a Chilean bank, and led to superior performance with respect to business-related goals.



## 1. Introduction

Classification is a very relevant task in many profit-driven applications, with credit scoring [1] being one of the most important ones. Predicting the customers that are likely to default on loan repayment via mathematical modeling has been a very important topic in recent decades mainly because it helps companies to make profitable financial decisions while fulfilling regulatory requirements [2].

Support Vector Machine (SVM) [3] is a powerful classification approach that can be useful for decision support systems given its superior performance compared to traditional strategies, like logistic regression [4–6]. This method, however, is not able to identify the most relevant features used for classifier construction [7,8].

Despite the plethora of feature selection and classification methods available in the machine learning literature, most of the work in Analytics applies traditional, statistically grounded techniques without using business-oriented measures. Several efforts have been made in developing profit metrics for comparing the various classification methods [see e.g. Refs. 9,10] . To the best of our knowledge, however, the only work that goes one step further and adapts the idea of profit-driven metrics to the task of feature selection was presented in Maldonado et al. [11]. In that study, various profit-based measures were used for backward feature elimination with SVM for the churn prediction problem, without taking variable acquisition costs into account.

In this work, we present an integrated framework for decision support in credit assignment. This framework takes both the analysis of classification costs and benefits into account, and includes the concept of variable acquisition costs. The idea is to find the best SVM classifier by balancing the profit obtained when the model is implemented with the cost of the variables that are included in it. The problem of grouped variables is addressed by using the $l_\infty$-norm penalty [12]. This function is combined with the $l_2$ and $l_1$ regularization functions, leading to two SVM formulations for classification and embedded feature selection. Two credit scoring datasets from a Chilean bank are used. This data comes from a previously developed project that involved small loans granted to microentrepreneurs [13]. Since interpretability is of utmost importance in credit scoring due to regulatory constraints, our framework is based on linear SVM, avoiding black-box modeling such as kernel-based SVM.

This paper is structured as follows: in Section 2, the concept of profit measure is described in the context of credit scoring. Previous work on feature selection and SVM classification is discussed in Section 3. The proposed profit-based framework using SVM is described in Section 4. In Section 5, the case study is presented, and

* Corresponding author.
*E-mail address:* smaldonado@uandes.cl (S. Maldonado).

https://doi.org/10.1016/j.dss.2017.10.007
0167-9236/

experimental results are given. Finally, the main conclusions of this study are presented in Section 6.

## 2. Profit-based credit scoring

In credit scoring, the first goal is to construct a vector of characteristics $x \in \Re^n$ that describe the repayment behavior of a set of borrowers $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$, with $y_i \in \{-1, +1\}$ the objective variable describing the event of default (1) or repayment (−1) after the first year of the life of the loan. The observation window was determined studying the number of months that pass until the bad rate for a portfolio is deemed to reach stability. This window tends to be between 12 to 18 months for consumer lending [14].

From these inputs, a probability function $s(x) = p(y = 1|x)$ can be obtained, which is then used to decide whether future borrowers are creditworthy or whether the loan should be rejected. For this, a cutoff point $t$ is used, so that if $s(x_i) > t$, then the loan application will be rejected.

Given the financial nature of credit scoring, tying profits to the model analysis and evaluation is a natural step. In Bravo et al. [13], this decision was tied to determining the cutoff point, extending the widely used analysis that determined this point by using the point where the slope of the Receiver Operator Characteristic (ROC) curve intercepted with the proportion between the average cost of misclassifying a good borrower versus misclassifying a bad borrower. This idea, of obtaining the best possible cutoff given only a fraction of the data, has been key in model evaluation. The H-measure [15] has already used it, defining the cutoff points that should be maximal given the distribution of costs, and later Verbeke et al. [9] extended it by including a more thorough profit- (not cost) based framework. It is this version, specifically the credit scoring profit framework by Verbraken et al. [10], that will be the base of our analyses.

Taking a continuous model, a decision can only be made if we choose a threshold $T$. For any cutoff $s$, some cases will be accepted and some rejected, which we will describe as $F_{-1}(s)$ and $F_1(s)$, the cumulative distributions of negative and positive cases at a given cutoff $s$, respectively. Additionally, most credit scoring problems are imbalanced, since the commercial conditions of the lender, such as the current acceptance policy, its risk appetite, the market segment it targets, and the propensity of repayment in the market where the lender operates influence the bad rate. In general, most retail portfolios tend to have a significantly larger number of good loans than bad loans. We will assume that the prior probability of being good is given by $\pi_{-1}$, and the one of being bad given by $\pi_1$, such that $\pi_{-1} + \pi_1 = 1$.

The last step in defining the profit comes from the analysis of (mis)classifying the cases in the dataset. There are two (potentially stochastic) costs that are relevant for the analysis: $b_{-1}$ is the benefit of accepting a good borrower, and $c_1$ the loss of accepting a bad borrower. Under this framework, the average profit per borrower, given a threshold $t$ is given by Ref. [9]:

$$P(t; b_{-1}, c_1) = b_{-1}\pi_{-1}F_{-1}(t) - c_1\pi_1F_1(t). \tag{1}$$

When $b_{-1}$ and $c_1$ are deterministic, then the maximization of this measure leads to the Maximum Profit (MP) measure [9]. If any is stochastic, this maximization lead to the Expected Maximum Profit (EMP) measure [16], both of which permit evaluating the performance of a model in a profit driven environment, once the correct form of the cost and benefit functions has been set. In Credit Scoring, both Hand [15] and Verbraken et al. [10] have detailed the most appropriate measures for each of the functions, which are relevant to this work.

For $b_{-1}$, the benefit of accepting a good borrower, the profit obtained throughout the life of the loan has to be normalized considering multiple repayment periods, i.e. the Return On Investment (ROI) of the loan. The total interest [17] formulas give this value. Considering a principal $A$ requested at maturity (terms) $T$ at an interest rate given by $r$, the total interest $I$ follows:

$$I = \frac{AMr}{1-(1+r)^{-M}} - A = A\left(\frac{Mr}{1-(1+r)^{-M}} - 1\right). \tag{2}$$

The ROI of the loan will simply be the total interest $I$ divided by the principal $A$, i.e.

$$b_{-1} = ROI = \frac{I}{A} = \frac{Mr}{1-(1+r)^{-M}} - 1. \tag{3}$$

On the other hand, the cost of accepting a bad applicant will be given by the loss that is incurred when the borrower defaults. The Basel II Banking Regulation Accords [18] define the expected loss of a borrower as

$$L = PD \cdot LGD \cdot EAD, \tag{4}$$

where the $PD$ is equal to the Probability of Default, which is derived by the scoring function $s(x)$. The EAD is the Exposure at Default, or the amount that is outstanding when default occurs, and the LGD is the Loss Given Default, or the percentage of the EAD that cannot be recovered after all collection actions have been exhausted. Note that to estimate the cost of accepting a bad borrower, the assumption is that the borrower will default with certainty. So $PD = 1$, and only the $LGD$ and the $EAD$ must be estimated to calculate the loss for each case. With this, the cost of accepting a bad borrower will be given by:

$$c_1 = LGD \cdot EAD. \tag{5}$$

To correctly calculate $c_1$, the values of the EAD and the LGD must be available. The Exposure at Default should already be present in the test set, as it corresponds to the value that was in lieu of payment at the time of default, for defaulted loans. For non-defaulted loans, this value is zero. The LGD might not be completely available at the time of evaluation, especially if the common practice of validating using out-of-time recent samples is followed. Three options are available to the modeler:

1. An incomplete workout period can be used, using as a measure the recovery rate up to the time when the sample was created. This practice falls in line with the recommendations given by the banking regulation agreements, which mandate the use of incomplete workouts for the estimation of LGD models [19]. The LGD is then calculated as $1 - RR_{\text{inc}}$, with $RR_{\text{inc}}$ the recovery rate at the time of observation.
2. The standard regulatory parameters can be used. The Basel agreements [18] propose a set of standard parameters for the LGD that should be used by all financial institutions not implementing their own LGD models, i.e., using standard or foundational Internal Ratings-Based (IRB) models.
3. For institutions implementing their own LGD models, their internal estimates can be used.

In previous studies, an average loss has been computed over all cases. We will use the value per-loan for our estimations. Considering each sample $i$, we will refer to the benefit of accepting a good borrower as $b_{-1,i}$, while the cost of accepting a bad borrower will be equal to $c_{1,i}$. Each parameter will then take the following functional form:

$$b_{-1,i} = \frac{M_i r_i}{1-(1+r_i)^{-M_i}} - 1,$$
$$c_{1,i} = LGD_i \cdot EAD_i. \tag{6}$$

One element that has been omitted so far in previous studies is the variable acquisition cost. This cost can be relevant in credit risk management as well as in many other disciplines.

The very diverse sources of data that any modern financial company has available is also reflected in diverse variable acquisition costs. Some of these costs are, for example:

- Internal data management costs: No matter the data source, there is a cost for managing the data and store it efficiently within the systems of an organization. This cost can also increase as the data diversity increases, as for example store a mix of structured data, images, and text might require the use of a Hadoop or similar data storage system, which comes with higher human resource cost.
- Internal data creation cost: Almost all organizations create data when evaluating a customer. Most organizations require, for example, a form with the loan application. This data comes at a cost, not only due to the time the credit officer needs to support potential borrowers in their application, but also due to internal processes, such as data verification. In the UK, for example, a survey indicated that 92% of all financial institutions had procedures set in place to validate the income information of borrowers [20]. It is very likely that different data sources also come at different costs: An in-depth interview such as the ones developed for first-time applicants some segments (notably SME such as the one in this research) might be very expensive, but an evaluation for a returning customer for whom only historical structured data is used might be the exact opposite.
- External data creation cost: When external providers (consultants or surveyors) are hired to produce primary data, organizations incur external data creation costs. These costs can be assigned to each variable evenly, for example. One example for this type of data would be property valuation services.
- External data provider cost: Finally, structured or unstructured data that can be purchased from providers, such as a credit bureau. These variables need to be properly costed considering contractual and per-case costs.

In Section 4.3, we will integrate the costs and benefits introduced in this section with the variable cost measures originally developed by Maldonado et al. [21].

## 3. Theoretical background on feature selection and SVM classification

In this section, we introduce the soft-margin SVM formulation for linear classification, and several well-known feature selection strategies for SVM classification that are relevant for this study.

### 3.1. Soft-margin SVM

The traditional soft-margin SVM formulation [3] finds a hyperplane of the form $\mathbf{w}^\top \mathbf{x} + b = 0$ by solving the following quadratic programming (QP) problem:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1,\ldots,m, \\ & \xi_i \geq 0, \qquad i = 1,\ldots,m, \end{aligned} \tag{7}$$

where $C > 0$ is a parameter that controls the trade-off between margin maximization and model fit, and $\xi_i$ denotes a slack variable related to each training example. This strategy for model fit is known as hinge loss. Although non-linear classifiers can be obtained from SVM using the kernel trick, we limit ourselves to linear classifiers since interpretability is crucial in domains like credit scoring due to regulatory constraints.

### 3.2. Feature selection for SVM

Several feature selection strategies have been proposed in the literature for SVM classification, which are divided in three families: filter, wrapper, and embedded methods Guyon et al. [7]. The first approach (*filter methods*) uses statistical properties to filter out irrelevant and/or redundant variables, assessing the correlation between them and the label vector. The *Fisher Score*, for example, is a statistical measure used to rank the attributes according to their contribution before applying any classification approach [22]. This metric evaluates the absolute difference, for each feature $j$, between the means of the positive $\left(\mu_j^+\right)$ and negative class $\left(\mu_j^-\right)$, divided by a joint standard deviation $\left(\sigma_j^+\right)^2 + \left(\sigma_j^-\right)^2$.

*Wrapper methods* score various variable subsets according to their predictive power. Since the exhaustive search for the optimal subset of variables is a combinatorial problem, several heuristic approaches have been suggested, such as a greedy search or meta-heuristics Guyon et al. [7].

The *Recursive Feature Elimination SVM* (RFE-SVM) [23], is a greedy approach that first trains SVM and then eliminates those features whose removal leads to the largest margin of class separation in an iterative fashion. Formally, the absolute value of the weight vector is computed, and the variable $j$ with the smallest value of $|w_j|$ is removed.

The RFE-SVM backward elimination algorithm was modified in Maldonado and Weber [8] to include a holdout step. The training set split into a holdout-training subset and a validation subset, in which the number of misclassified instances is computed. The rationale behind this strategy, called Holdout SVM (HOSVM), is to eliminate those features whose removal has the least impact on the out-of-sample classification performance given by the accuracy computed on the validation set. This idea was further extended in Maldonado et al. [11] for profit-based feature selection for churn prediction. The accuracy measure was replaced by the profit obtained by a retention campaign, considering the respective costs and benefits (MPC and EMPC). In this work, we compare our proposal with the HOSVM method using AUC and MPC as performance metrics on the validation set.

There are important differences between the current proposal and the work by Maldonado et al. [11]. First, the applications are different: Maldonado et al. [11] focuses on churn prediction in telco, while the current proposal faces a credit scoring problem in a Chilean bank, with an alternative definition of profit. There is also a methodological difference regarding the feature selection strategy: Unlike HOSVM, our proposal does not perform an iterative strategy in which attributes are discarded based on their contribution in the profit measure. In the current paper, however, we penalize the use of (group of) features in the SVM formulation by introducing a group penalty function in the SVM formulation, a strategy that can be considered an *embedded method*.

Embedded methods find an optimal subset of features in the process of model construction. Embedded methods are able to capture dependencies between variables effectively, being computationally less demanding than wrapper methods [7]. They are, however, conceptually more complex than filter and wrapper methods, and modifications to SVM could affect it virtues, such as convexity and computational efficiency.

A well-known embedded strategy for SVM classification is to penalize the use of features by replacing the squared Euclidean norm in Formulation (7) with a regularizer that encourages sparsity. The most common approach is use of the LASSO penalty or $l_1$ norm,

which provides a good compromise between complexity reduction and feature elimination [24].

Along the same line, a *group penalty function* is a regularization strategy designed to penalize the use of a group of related variables together in such a way that sparsity is encouraged at a group level instead of by removing weights independently [25]. Such a strategy has been used in binary classification with categorical attributes with multiple levels, which are usually transformed into sets of dummy variables. In such cases, it may be desirable to remove the full set of dummy variables [25]. Feature selection can be performed simultaneously at a variable level, jointly penalizing all the weights related to one attribute in each classification function [see e.g. Ref. 26] .

The best-known group penalty is called *group-LASSO* [25]; it extends the idea of the LASSO penalty by penalizing the Euclidean norm of the weights related to a given group. This group penalty function has the following form:

$$\Gamma(\mathbf{w}) = \sum_{j=1}^{J} \sqrt{p_j} \parallel \mathbf{w}^{(j)} \parallel_2, \tag{8}$$

where $\parallel \mathbf{w}^{(j)} \parallel_2 = \sqrt{\sum_{l \in \mathcal{I}_j} w_l^2}$. The measure $\mathcal{I}_j$ represents disjoint sets of related features linked to a given attribute $j = 1, \ldots, J$, where $|\mathcal{I}_j| = p_j$ is the total number of levels considered for each nominal variable (one of the levels can be used as reference category for avoiding multicollinearity issues), and $\sum_{j=1}^{J} p_j = n$ represents the total number of estimated weights.

Next, a framework based on SVM is proposed to find an optimal solution that balances the benefits and costs of classification with the variable acquisition costs. Feature selection is performed by adding a group penalty function, given the variable acquisition costs scheme.

## 4. Proposed profit-based framework for credit scoring using SVM

The main idea of this proposal is to provide a profit-based classification framework for SVM, performing feature selection simultaneously with the classifier construction. The proposed approach is applied in the context of credit scoring, although it can be used in any application where the benefits for correct classification, the misclassification costs, and the variable acquisition costs are estimated.

The proposed method is introduced in three sections: The $l_\infty$-norm penalty function for grouped feature selection is introduced in Section 4.1. The proposed classification models are presented in Section 4.2. And finally, the use of profit metrics for feature and model selection is discussed in Section 4.3.

### 4.1. The L-infinity norm as group penalty function

Our proposal considers $n$ attributes that stem from various sources with different variable acquisition costs. In order to reduce these costs, the whole set of variables related to one source is jointly penalized, using the $l_\infty$-norm regularizer [12]. This function has the following form:

$$\Gamma(\mathbf{w}) = \sum_{j=1}^{J} \parallel \mathbf{w}^{(j)} \parallel_\infty \tag{9}$$

where $\parallel \mathbf{w}^{(j)} \parallel_\infty = \max_{l \in \mathcal{I}_j} \{|w_l|\}$, i.e. the highest weight (in magnitude) for each source of variables $j = 1, \ldots, J$ is minimized, $\mathcal{I}_j$ being the set of variables that belong to source $j$. The $l_\infty$-norm penalty was originally developed for dealing with categorical variables in binary SVM classification, under the name $F_\infty$-norm SVM. The main advantage of the $l_\infty$-norm penalty of the group LASSO is that the former can be cast easily into a smooth linear function. This strategy has not been used for selecting attributes with different acquisition costs, to the best of our knowledge.

### 4.2. The proposed models for classification and feature selection

We propose two double-regularized SVM formulations: the $l_2 l_\infty$-SVM, and $l_1 l_\infty$-SVM approaches. They differ in the regularization strategy used to control the complexity of the solution, according to the Structural Risk Minimization (SRM) principle followed by SVM [3]. The $l_2 l_\infty$-SVM method combines three objectives: Euclidean norm minimization (also known as Tikhonov regularization), $l_\infty$-norm penalization for grouped feature selection, and hinge loss minimization to guarantee an adequate model fit. Alternatively, the $l_1 l_\infty$-SVM method is equivalent to $l_2 l_\infty$-SVM, but uses the LASSO penalty instead of the Euclidean norm. The $l_2 l_\infty$-SVM model has the following form:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i + \lambda\sum_{j=1}^{J} \parallel \mathbf{w}^{(j)} \parallel_\infty \\ \text{s.t.} \quad & y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1 - \xi_i, \qquad i = 1, \ldots, m, \\ & \xi_i \geq 0, \qquad i = 1, \ldots, m, \end{aligned} \tag{10}$$

where $C, \lambda > 0$ are parameters that will be tuned via grid search with cross-validation. In order to avoid using a non-smooth function in the previous problem, we introduce a set of auxiliary variables $z_j \geq 0$, and add new constraints $|w_l| \leq z_j$ for each $l \in \mathcal{I}_j$ and $j = 1, \ldots, J$. The quadratic programming problem solved by $l_2 l_\infty$-SVM becomes:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi},\mathbf{z}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i + \lambda\sum_{j=1}^{J} z_j \\ \text{s.t.} \quad & y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1 - \xi_i, \qquad i = 1, \ldots, m, \\ & \xi_i \geq 0, \qquad i = 1, \ldots, m, \\ & -z_j \leq w_l \leq z_j, \quad l \in \mathcal{I}_j, \quad j = 1, \ldots, J. \end{aligned} \tag{11}$$

Similarly, the $l_1 l_\infty$-SVM model has the following form:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \parallel \mathbf{w} \parallel_1 + C\sum_{i=1}^{m}\xi_i + \lambda\sum_{j=1}^{J} \left\| \mathbf{w}^{(j)} \right\|_\infty \\ \text{s.t.} \quad & y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1 - \xi_i, \qquad i = 1, \ldots, m, \\ & \xi_i \geq 0, \qquad i = 1, \ldots, m, \end{aligned} \tag{12}$$

with $\parallel \mathbf{w} \parallel_1 = \sum_{i=1}^{n} |w_i|$ denoting the $l_1$-norm of $\mathbf{w}$. Again, a set of auxiliary variables is introduced in order to avoid a non-smooth optimization problem, leading to the following linear programming model:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi},\mathbf{z},\mathbf{u}} \quad & \sum_{i=1}^{n} u_i + C\sum_{i=1}^{m}\xi_i + \lambda\sum_{j=1}^{J} z_j \\ \text{s.t.} \quad & y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1 - \xi_i, \qquad i = 1, \ldots, m, \\ & \xi_i \geq 0, \qquad i = 1, \ldots, m, \\ & -z_j \leq w_l \leq z_j, \quad l \in \mathcal{I}_j, \quad j = 1, \ldots, J, \\ & -\mathbf{u} \leq \mathbf{w} \leq \mathbf{u}. \end{aligned} \tag{13}$$

### 4.3. Proposed profit metric for model and feature selection

One of the main contributions of this work is the modification of the traditional profit measure to incorporate variable acquisition costs. This metric is used for model selection for standard SVM

(tuning of parameter $C$) and our approaches (tuning of parameters $C$ and $\lambda$).

Let us consider the SVM classifier $\Lambda = \{\mathbf{w}, b\}$, and a validation subset $\mathcal{V}$ with samples $\mathbf{x}_l^v$ and labels $y_l \in \{-1, +1\}$, for $l = 1, \ldots, |\mathcal{V}|$. The profit for $\mathcal{V}$ consists of the benefits associated to the correctly classified non-defaulters (negative class), minus the losses associated with the misclassified defaulters (positive class), minus the variable acquisition costs for each source of attributes used in the classifier construction. Formally, the profit measure is redefined as follows:

$$Profit(\Lambda, \mathcal{V}) = \sum_{l \in \mathcal{V}^-} b_{-1,l} \frac{1 - sgn\left(\mathbf{w}^\top \mathbf{x}_l^v + b\right)}{2} - \sum_{l \in \mathcal{V}^+} c_{1,l} \frac{1 - sgn\left(\mathbf{w}^\top \mathbf{x}_l^v + b\right)}{2} - |\mathcal{V}| \sum_{j=1}^{J} AC_j I_j \tag{14}$$

where $b_{-1,l}$ represents the benefit for granting credit to a non-defaulter $l \in \mathcal{V}^-$, $c_{1,l}$ is the loss for granting credit to a defaulter $l \in \mathcal{V}^+$, and $AC_j$ is the variable acquisition cost for a source $j$. Additionally, $\mathcal{V}^+$ ($\mathcal{V}^-$) is the subset of positive (negative) instances in the validation set $\mathcal{V}$, $|\mathcal{V}|$ is the cardinality of this set, and $I_j$ is an indicator variable that takes the value 1 if $\max_{l \in \mathcal{I}_j}\{|w_l|\} > \epsilon$. That is to say, at least one attribute has a weight higher (in magnitude) than $\epsilon$, a sufficiently small parameter, for each source of variables $j = 1, \ldots, J$. Notice that this metric estimates the total profit of a solution on a validation set, rather than computing the expected profit of a solution based on the prior probabilities of being defaulter or non-defaulter.

Following the ideas discussed in Section 2, we assume that the benefits and losses for granting credit depend on the applicant. We compute $b_{-1,l}$ as the ROI obtained by the lender for each loan $l$, while $c_{1,l}$ is computed as the expected loss considering that the loan $l$ is already in default. Note that the interest rate and the amount granted are usually available for the customers when they apply for the loans, but the losses that a defaulter generates are not available at that moment. In this case, following Bravo et al. [13], it is possible to compute the LGD as an average, based on information of previous defaulted loans for the expected loss segment, as given by the *PD* of the borrower. We also assume that the variable acquisition costs are similar for all applicants, since they can usually be estimated as the monetary cost of purchasing certain information (e.g. from credit bureaus), or by valuating the time an analyst requires gathering the information from a given source.

## 5. Experimental results

We applied the proposed $l_2 l_\infty$-SVM and $l_1 l_\infty$-SVM approaches to two credit scoring datasets. We also studied other alternative feature selection methods described in Section 3.2 (Fisher Score, RFE-SVM, and the HOSVM method using AUC and MPC as performance metrics for the validation set) for comparison purposes.

This section is organized as follows: the credit scoring project that provided the dataset is described in Section 5.1. The experimental settings are described in Section 5.2. In Section 5.3, a summary of the performance obtained for the proposed and alternative methods is presented. Finally, the detailed feature selection performance for various metrics and subsets of selected variables is reported in Section 5.4.

### 5.1. Description of the case study

The data comes from a Chilean bank that provides loans to small and micro-companies, loans that are repaid in monthly installments. The information was collected in the period from 2004 to 2007. The target variable corresponds to the usual definition of default based on Basel II/III: one or more installments in arrears for more than 90 days during the first year of the loan [18].

The customers are divided into two datasets according to their credit history with the bank, as follows:

- New customers (NEW): A total of 1510 customers was available, of which 629 of them were defaulters. After pre-processing and filtering out irrelevant information, a total of 94 attributes was available.
- Returning customers (RET): A total of 5799 customers was available, of which 872 of them were defaulters. After pre-processing, a total of 46 attributes was available.

Besides the benefits and costs of granting loans, the variable acquisition costs were studied with the proposed framework. Specifically, the dataset includes expensive internal processes as well as data from external sources for new customers. The different sources of information can be modeled as groups of related variables with a single cost for using this source in the model. In other words, if one variable is identified as relevant and included in the final model, then all the remaining variables of its group can be included at zero cost. The following groups of related attributes were identified:

- Credit evaluation attributes: This set of attributes comes from the form that each applicant fills out. This application is subsequently analyzed by the risk department, and then registered into the company database. Since all applicants are required to fill out one of these forms, the acquisition of these variables can be seen as a sunk cost. These forms are filled out by the applicants in the company with one credit officer from the bank; a task that takes about 1 h, on average. Thus, the estimated cost per borrower is €5, assuming a monthly salary of €1000 per credit officer. After pre-processing, a total of 32 and 31 attributes from this source of variables were available for the new and returning customers, respectively.
- In-depth interview attributes: The bank conducts an in-depth interview of the applicant after the evaluation process. This interview is done during a visit to the place of work of the applicant made by an credit officer. The estimated cost for this set of attributes is €20 per application (four hours of an credit officer's time). After pre-processing, a total of 5 and 2 attributes from this source of variables was available for the new and returning customers, respectively. Few variables are relevant because this source of information is mainly used as an input for the next set of variables.
- Financial analysis attributes: Once an in-depth interview is performed, the bank estimates the cash flow of the company, which is usually not available. This is done by a specialist who has an estimated monthly salary of €2000, in about two hours per borrower. The estimated cost is then €20 per borrower. After pre-processing, a total of 34 and 13 attributes from this source of variables was available for the new and returning customers, respectively.
- System-level information: In order to enrich the information on borrowers with no credit history (new customers), the financial institution acquired information on the borrowers' standing debts in the financial system. This data source has a global fixed cost of €1000. After pre-processing, a total of 9 attributes was available.
- Financial analysis attributes based on system-level information: Some of the variables constructed during the financial analysis combined system-level information with other sources, such as evaluation or interview data. Although this source does not represent a new group *per se*, the costs of the four previous groups need to be considered if variables of this source are included in the solution. After pre-processing, a total of 14 attributes was available for the new customers.

**Table 1**
Predictive performance for all feature selection approaches. New customers. Profit as the performance metric.

| | New customers | | | | | | |
|---|---|---|---|---|---|---|---|
| | Logit | Fisher | RFE-SVM | $HOSVM_{AUC}$ | $HOSVM_{MPC}$ | $l_2l_\infty$-SVM | $l_1l_\infty$-SVM |
| AUC | 69.6 | 50.0 | 60.1 | 64.7 | 67.6 | 66.6 | 66.6 |
| Accu. | 70.4 | 58.3 | 64.2 | 66.6 | 68.5 | 68.2 | 68.2 |
| n* | 28.7 | 5 | 5 | 10 | 10 | 35.4 | 26.4 |
| J* | 4.8 | 2.9 | 2.5 | 3.4 | 3.3 | 1.1 | 1 |
| Profit | 36 | 1107 | 910 | 1742 | 1845 | 4449 | **4699** |
| Benefits | 8769 | 9965 | 8528 | 8008 | 7976 | 8342 | 8370 |
| Losses | 3235 | 5933 | 3745 | 3081 | 2696 | 3045 | 3078 |
| Acq. costs | 5498 | 2925 | 3873 | 3185 | 3435 | 849 | 592 |

Note that, unusually, more information is available in this dataset for new customers than for returning customers. The reason for this comes from evaluation process carried out for both segments: the new customers do not usually have any past credit history, a deeper (and more expensive) evaluation process was done for them, resulting in the larger number of variables for the financial analysis attribute group. For returning customers, only certain variables were captured in these segments, and past credit history was added. This results in a smaller overall number of variables for the returning customers.

### 5.2. Experimental settings

We used 10-fold cross-validation for model selection purposes, exploring the following values for $C$ and $\lambda$: $\{2^{-7}, 2^{-6}, \ldots, 2^{-1}, 2^0, 2^1, \ldots, 2^6, 2^7\}$. Our proposals perform automatic feature selection, and different combinations of $C$ and $\lambda$ lead to different solutions in terms of performance and attributes selected. In contrast, the Fisher Score, RFE-SVM, and HOSVM methods are feature ranking approaches, requiring a predefined number of attributes as an input. For such approaches, model selection was performed using all the attributes, and SVM was trained subsequently for subsets of ranked features of size $n = \{5,10, 20, 30, 40, 50, 60, 70, 80, 90\}$ and $n = \{5,10, 20, 30, 40\}$ for the new and returning customers, respectively.

Logistic regression is used as an additional benchmark approach since it is the standard model for credit scoring [2]. A backward elimination procedure is performed, removing those variables whose coefficients are not statistically significant based on the Wald test using a significance level of $\alpha = 5\%$.

A combination of undersampling and SMOTE oversampling was performed for the returning customers to deal with the class-imbalance problem [4,27]. We performed 200% oversampling for the minority class, and then undersampling to perfect balance. For the SMOTE oversampling, the nearest neighbors were set to 5, as suggested in Chawla et al. [27]. Data resampling was performed only for the training set. This resampling technique proved to be the most effective one in terms of predictive performance in our previous works on imbalanced data classification [see Refs. 11,28] .

The following pre-processing strategy was used to discard irrelevant information [see Ref. 13, for more details] :

- A first filter was applied in order to discard useless variables, eliminating those with nominal variables with more than a 99% concentration at a single level, numerical variables with zero standard deviation, or more than 30% of missing values.
- The two-sample independence tests Kolmogorov-Smirnov (KS) and $\chi^2$ were applied for numerical and nominal variables, respectively, in order to discard attributes that are statistically independent with the target variable at $\alpha = 5\%$ significance level.

### 5.3. Result summary

Next, a summary of the results is presented. Tables 1 and 2 show the performance of each method and of new and returning customers, respectively, when the profit metric presented in Section 4.3 is used for model selection. For the logistic regression with the backward elimination process, the cutoff is chosen to maximize the total profit in the validation set. This is done by evaluating the profit using the following values for the cutoff $t \in \{0,0.05, 0.1, 0.15,\ldots,0.95, 1\}$. The following performance measures are reported: AUC ( $\times 100$), overall accuracy (in percentage), number of selected variables $n*$, number of sources of variables selected $J*$, the proposed profit metric, benefits due to correct identification of non-defaulters, losses due to incorrect identification of defaulters, and variable acquisition costs. All monetary metrics are expressed in Euros for a group of approx. 150 and 580 applicants for new and returning customers, respectively (one tenth of the full sample, which the average validation sample size for the 10-fold cross-validation procedure). The best performance among all methods in terms of profit is highlighted in bold type.

In Tables 1 and 2, we observe relatively similar results for new and returning customers. The methods $l_2l_\infty$-SVM and $l_1l_\infty$-SVM tend to use only the information from the first source of attributes (credit evaluation), leading to the lowest variable acquisition costs and the best performance in terms of profit. The resulting differences in terms of profit between our proposal and the alternative methods

**Table 2**
Predictive performance for all feature selection approaches. Returning customers. Profit as the performance metric.

| | Returning customers | | | | | | |
|---|---|---|---|---|---|---|---|
| | Logit | Fisher | RFE-SVM | $HOSVM_{AUC}$ | $HOSVM_{MPC}$ | $l_2l_\infty$-SVM | $l_1l_\infty$-SVM |
| AUC | 67.7 | 62.3 | 56.9 | 66.1 | 65.5 | 63.2 | 63.2 |
| Accu. | 84.9 | 56.6 | 56.6 | 61.3 | 61.1 | 54.6 | 57.4 |
| n* | 28.8 | 20 | 5 | 20 | 20 | 31 | 31 |
| J* | 2.9 | 2.2 | 1.9 | 2 | 2.1 | 1 | 1 |
| Profit | 35,913 | 24,354 | 22,972 | 29,339 | 27,913 | 36,581 | **38,618** |
| Benefits | 67,282 | 42,881 | 39,193 | 45,401 | 45,080 | 42,419 | 44,874 |
| Losses | 10,372 | 4427 | 5077 | 3933 | 4053 | 3564 | 3982 |
| Acq. costs | 20,998 | 14,100 | 11,143 | 12,129 | 13,114 | 2274 | 2274 |

**Table 3**
Predictive performance for all feature selection approaches. New customers. AUC as the performance metric.

| | New customers | | | | | | |
|---|---|---|---|---|---|---|---|
| | Logit | Fisher | RFE-SVM | $HOSVM_{AUC}$ | $HOSVM_{MPC}$ | $l_2l_\infty$-SVM | $l_1l_\infty$-SVM |
| AUC | 69.6 | 69.6 | 70.4 | 69.0 | 69.6 | **70.7** | **70.7** |
| Accu. | 70.6 | 70.0 | 70.9 | 69.5 | 70.4 | 71.3 | 71.5 |
| n* | 28.7 | 80 | 90 | 80 | 20 | 90.5 | 58.5 |
| J* | 4.8 | 5 | 5 | 5 | 4 | 4.3 | 3.9 |
| Profit | −5 | −305 | −40 | −379 | 1271 | 1806 | 2541 |
| Benefits | 8018 | 7738 | 8003 | 7753 | 7991 | 8107 | 8114 |
| Losses | 2525 | 2287 | 2288 | 2376 | 2507 | 2342 | 2386 |
| Acq. costs | 5498 | 5756 | 5756 | 5756 | 4213 | 3960 | 3187 |

are noteworthy, mainly due to its ability to identify cheap solutions in terms of variable acquisition costs, while the alternative approaches use more than two different sources even when selecting five attributes. Results are relatively similar in terms of benefits and losses, except for the Fisher Score, whose best solution in terms of profit implies grating credit to all applicants (AUC = 0.5), leading to higher benefits but also greater losses. For this method, the benefits of improving classification performance with additional variables is not able to compensate the acquisition costs. Another exception is the logit model, which has the best predictive performance but a high variable acquisition cost since it cannot be tuned for removing expensive variables.

Tables 3 and 4 show similar information compared with Tables 1 and 2, but AUC is used instead of the profit metric to select the best model for new and returning customers, respectively. For the logistic regression with a backward elimination process, results based on the maximum likelihood cutoff of 0.5 are reported. Notice that the results for the logistic regression are equivalent to tables those shown in Tables 1 and 2 for the metrics AUC, variables selected, sources selected, and variable acquisition costs, since the only parameter tuned is the cutoff. The best performance among all methods in terms of AUC is highlighted in bold type.

In Tables 3 and 4, we observe first that higher-dimensional solutions are found, compared with the results presented in Tables 1 and 2, leading to an increase in AUC and accuracy, but to an important loss in terms of profit. This occurs because the use of the additional sources of variables leads to better classification performance, but this benefit is not able to compensate for the high cost of performing interviews and financial analyses. A comparison between the different feature selection methods shows that classification performance is relatively similar among them in terms of the highest AUC, our proposals having best performance for the new customers. Notably, even though the model is selected to maximize AUC, both $l_2l_\infty$-SVM and $l_1l_\infty$-SVM show a higher profit than the rest of the benchmarked models. This follows since our methods penalize the use of variables as groups, finding solutions that are cheap in terms of variable acquisition costs, but as accurate as the other methods studied.

At this point, it is important to highlight some characteristics of the datasets that are rather uncommon in credit risk studies. First, the AUC and accuracies for the returning customer (behavioral) model are lower than the standard. This is caused mainly by the fact that micro-entrepreneurs are more homogeneous than traditional applicants, and some variables, like income, are not relevant for this problem [see Ref. 13, for a detailed discussion around this topic]. Additionally, types I and II error costs are quite similar: interest rates are very high compared to the standard due to the higher risk, but recovery rates are also high for loans labeled as defaulted due to effective renegotiations. This segment is profitable for the bank mainly because of the high interest rate and the low LGD. Finally, default rates are unusually high for the new loans when comparing them with retail loans, which explains the high interest rates that are charged to this segment. Our framework can help elucidate which sets of variables are the most profitable given other default rate and interest rate structures, where we do not expect, for example, that using just one variable source is optimal. This follows from the combination of these factors, plus the acquisition cost structure of this particular problem.

### 5.4. Detailed feature selection performance

Next, the feature selection performance is detailed by plotting the variable acquisition costs and the profit for an increasing number of selected attributes for both datasets. Our approach is not directly comparable with feature ranking methods like Fisher Score, RFE-SVM, and HOSVM since it automatically identifies the optimal subset of features during the classifier construction, leading to single solutions. In order to make this comparison, we report the performance obtained by the alternative methods for the subsets of size $n$ discussed in Section 5.2, and compare them with various solutions of different cardinality obtained by $l_2l_\infty$-SVM and $l_1l_\infty$-SVM using different values of $C$ and λ. In order to reflect the trends that each graph follows, we graph, in addition to the points, the best polynomial that adjusts to those points. These graphs are presented in Figs. 1 and 2 for the new customers, and Figs. 3 and 4 for the returning customers. The logistic regression model is excluded from this analysis since it provides a single subset of relevant variables, in contrast to feature ranking methods, and there are no parameters to tune to obtain different solutions.

**Table 4**
Predictive performance for all feature selection approaches. Returning customers. AUC as the performance metric.

| | Returning customers | | | | | | |
|---|---|---|---|---|---|---|---|
| | Logit | Fisher | RFE-SVM | $HOSVM_{AUC}$ | $HOSVM_{MPC}$ | $l_2l_\infty$-SVM | $l_1l_\infty$-SVM |
| AUC | **67.7** | **67.8** | 65.0 | 67.0 | 67.2 | **67.7** | 67.4 |
| Accu. | 65.1 | 64.0 | 61.7 | 63.3 | 63.4 | 63.8 | 63.8 |
| n* | 28.8 | 40 | 40 | 40 | 40 | 45.6 | 44.2 |
| J* | 2.9 | 3 | 3 | 3 | 3 | 2.8 | 2.1 |
| Profit | 23,579 | 21,519 | 19,740 | 21,032 | 21,048 | 23,263 | 30,399 |
| Benefits | 48,436 | 47,218 | 45,718 | 46,938 | 46,822 | 47,008 | 47,315 |
| Losses | 3860 | 3716 | 3995 | 3922 | 3791 | 3734 | 3802 |
| Acq. costs | 20,998 | 21,983 | 21,983 | 21,983 | 21,983 | 20,012 | 13,114 |

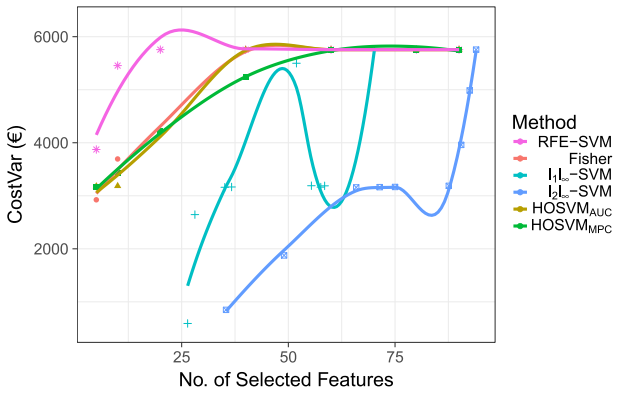


**Fig. 1.** Total variable acquisition costs for an increasing number of features. New customers.

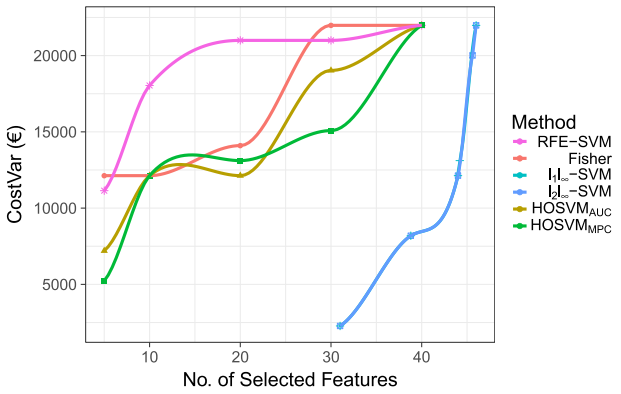


**Fig. 3.** Total variable acquisition costs for an increasing number of features. Returning customers.

For the new customers, it can be seen that the variable acquisition costs decrease significantly for $l_2l_\infty$-SVM and $l_1l_\infty$-SVM when fewer attributes are selected. This is in contrast to the alternative methods (Fig. 1), leading to important differences in terms of profit (Fig. 2).

A similar analysis can be done for the returning customers: While benefits and costs are roughly the same for all approaches, acquisition costs are much lower in our proposal compared to the alternative methods (Fig. 3), leading to a much higher profit (Fig. 4). It is important to note that no solution with less than 30 attributes is found for our method, since it tends to use all available variables from the cheapest source, given the nature of the $l_\infty$-norm regularizer.

The best overall performance in terms of profit is achieved by $l_1l_\infty$-SVM, suggesting that the $l_1$ regularization is more compatible with the $l_\infty$-norm than with the Tikhonov regularization, although the differences are very small between them in terms of performance. In terms of complexity, we recommend $l_1l_\infty$-SVM since it can be cast into a linear programming problem, while the $l_2l_\infty$-SVM requires quadratic programming solvers.

## 6. Conclusions and future developments

In this study, a profit-based framework for model and feature selection is developed. The main goal is to incorporate variable acquisition costs in the modeling decisions, and assess the performance of the solution taking this information into account together with the benefits and losses caused by correct and incorrect classification, respectively. The proposal includes two formulations that use the $l_\infty$-norm as a group penalty function, encouraging solutions that use few sources of attributes rather than a traditional feature selection scheme where all attributes have the same cost. In terms of computational complexity, $l_2l_\infty$-SVM and $l_1l_\infty$-SVM are almost equivalent to $l_2$-SVM and $l_1$-SVM, respectively. The differences are the inclusion of variables $\mathbf{z}$, one for each variable group, an extra set of constraints for the weight vector, and the additional linear term in the objective function.

The proposed framework was applied in a credit scoring project of a Chilean bank, which consists of two datasets of applicants from the micro-entrepreneur segment. A detailed cost-benefit analysis was performed for this data, including the computation of the financial losses for defaulters, the benefits of successfully granted loans, and the variable acquisition costs for each source of information. Based on this information, we compared our $l_2l_\infty$-SVM and $l_1l_\infty$-SVM formulations with well-known feature selection strategies, such as the HOSVM, the RFE-SVM, and the Fisher Score methods, showing the importance of profit-based evaluation in analytics.

From our experimental results, we can conclude that our strategy outperforms alternative methods in terms of profit thanks to its ability to identify accurate solutions using few sources of variables, in contrast to traditional feature selection methods for SVM that prioritize relevance over the source of the information. The proposal also achieves a positive performance if traditional metrics are used, leading to the highest AUC for the new customers and a similar one compared to that of the best method for the returning customers. In our case study, however, some sources of attributes are too expensive and the marginal benefits in terms of classification power gained by using these sources are lower than the acquisition costs. In consequence, selecting the best model based on AUC leads to an important loss of profit in this application.

Even though our solution comes at a slightly increased computational cost, it can be applied in multiple situations. If the data sources

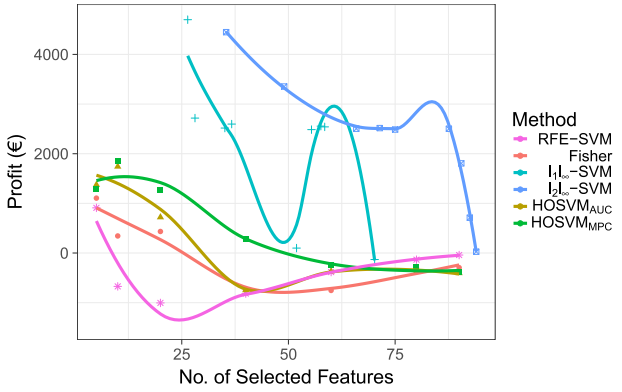


**Fig. 2.** Total profit (benefits – losses – var.acq.costs) for an increasing number of features. New customers.

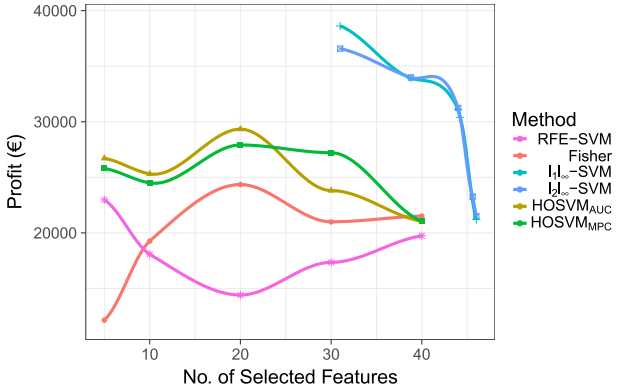


**Fig. 4.** Total profit (benefits – losses – var.acq.costs) for an increasing number of features. Returning customers.

are costly, or the modeler must decide between many providers, then a profit-optimizing model will quickly offset the computational expense. Our method can also be used as an input for a traditional model, determining the most profitable sets of variables before estimating the final predictive model.

Important future developments can be derived from this work. The following directions are viewed as future work:

- This work can be extended further to other analytics applications in which profit measures are relevant for model selection, such as churn prediction and fraud detection. For example, several sources of information can be identified in telecommunication companies, such as socio-demographic variables, call detail records, and information from external sources.
- There are interesting applications in the medical sciences for which the proposed approach could be used. Although the estimation of the classification benefits and costs can be more challenging than in analytics, our proposal can be helpful in identifying which sources of attributes are most relevant in diagnosing a disease. The Electroencephalogram (EEG) and the Electrocardiogram (ECG) are well-known sources of information that have acquisition costs and can be used jointly with the personal information of the patient.
- The $l_\infty$-norm penalization can be used in other classification methods, such as logistic regression. Consequently, our framework can be extended to other linear methods.

## Acknowledgments

The first author was supported by FONDECYT project 1160738. The third author was funded by FONDECYT project 1160894. The fourth author was supported by FONDECYT project 11160320. This research was partially funded by the Complex Engineering Systems Institute, ISCI (ICM-FIC: P05-004-F, CONICYT:FB0816).

## References

[1] B. Baesens, Analytics in a Big Data World, John Wiley and Sons. 2014.
[2] L. Thomas, J. Crook, D. Edelman, Credit Scoring and its Applications, SIAM. 2002.
[3] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[4] M. Farquad, I. Bose, Preprocessing unbalanced data using support vector machine, Decis. Support. Syst. 53 (1) (2012) 226–233.
[5] Z. Huang, H. Chen, C.-J. Hsu, W.-W. Chen, S. Wu, Credit rating analysis with support vector machines and neural networks: a market comparative study, Decis. Support. Syst. 37 (4) (2004) 543–558.
[6] J. Li, L. Wei, G. Li, W. Xu, An evolution strategy-based multiple kernels multi-criteria programming approach: the case of credit decision making, Decis. Support. Syst. 51 (2011) 292–298.
[7] I. Guyon, S. Gunn, M. Nikravesh, L.A. Zadeh, Feature Extraction, Foundations and Applications, Springer, Berlin, 2006.
[8] S. Maldonado, R. Weber, A wrapper method for feature selection using Support Vector Machines, Inf. Sci. 179 (2009) 2208–2217.
[9] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, B. Baesens, New insights into churn prediction in the telecommunication sector: a profit driven data mining approach, Eur. J. Oper. Res. 218 (1) (2012) 211–229.
[10] T. Verbraken, C. Bravo, R. Weber, B. Baesens, Development and application of consumer credit scoring models using profit-based classification measures, Eur. J. Oper. Res. 238 (2) (2014) 505–513.
[11] 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, Appl. Soft Comput. 35 (2015) 740–748.
[12] H. Zou, M. Yuan, The F-infinite norm support vector machine, Stat. Sin. 18 (2008) 379–398.
[13] C. Bravo, S. Maldonado, R. Weber, Methodologies for granting and managing loans for micro-entrepreneurs: new developments and practical experiences, Eur. J. Oper. Res. 227 (2) (2013) 358–366.
[14] N. Siddiqi, Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring, vol. 3. John Wiley & Sons. 2007.
[15] D. Hand, Measuring classifier performance: a coherent alternative to the area under the ROC curve., Mach. Learn. 77 (2009) 103–123.
[16] T. Verbraken, W. Verbeke, B. Baesens, A novel profit maximizing metric for measuring classification performance of customer churn prediction models, IEEE Trans. Knowl. Data Eng. 25 (5) (2013) 961–973.
[17] S.A. Broverman, Mathematics of Investment and Credit, Actex Publications. 2010.
[18] Basel Committee on Banking Supervision, Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework — Comprehensive Version, June 2006, http://www.bis.org/publ/bcbsca.htm.
[19] European Banking Authority, Guidelines on the Implementation, Validation and Assessment of Advanced Measurement (AMA) and Internal Ratings Based (IRB) Approaches, 2016.
[20] Financial Conduct Authority, Assessing Creditworthiness in Consumer Credit, cp17/27 ed., 2017.
[21] S. Maldonado, J. Pérez, C. Bravo, Cost-based feature selection for SVM classification — an application in credit scoring, Eur. J. Oper. Res. 261 (2) (2017) 656–665.
[22] R. Duda, P. Hard, D. Stork, Pattern Classification, Wiley-Interscience Publication. 2001.
[23] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines., Mach. Learn. 46 (1-3) (2002) 389–422.
[24] P. Bradley, O. Mangasarian, Feature selection via concave minimization and support vector machines, Machine Learning proceedings of the fifteenth International Conference (ICML'98) 82-90, San Francisco, California, Morgan Kaufmann, 1998.
[25] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J. R. Stat. Soc. Ser. B 68 (2006) 49–67.
[26] O. Chapelle, S. Keerthi, Multi-Class Feature Selection with Support Vector Machines, 2008.
[27] N.V. Chawla, L. Hall, K. Bowyer, W. Kegelmeyer, SMOTE: synthetic minority oversampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[28] S. Maldonado, Churn prediction via support vector classification: an empirical comparison, Intelligent Data Analysis 19 (S1) (2015) 135–147. special Issue in Business Analytics.