Applied Soft Computing 60 (2017) 831–843

Contents lists available at ScienceDirect

Applied Soft Computing

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

Probabilistic modeling and visualization for bankruptcy prediction

Francisco Antunes [a,*], Bernardete Ribeiro [a], Francisco Pereira [b]

[a] CISUC, University of Coimbra, Department of Informatics Engineering, Pólo 2, Pinhal de Marrocos, 3030–290 Coimbra, Portugal
[b] DTU, Technical University of Denmark, Bygningstorvet, 2800 Kongens-Lyngby, Denmark

ARTICLE INFO

Article history:
Received 12 April 2016
Received in revised form 7 June 2017
Accepted 23 June 2017
Available online 1 July 2017

Keywords:
Bankruptcy prediction
Machine learning
Gaussian processes
Graphical visualization

ABSTRACT

In accounting and finance domains, bankruptcy prediction is of great utility for all of the economic stakeholders. The challenge of accurate assessment of business failure prediction, specially under scenarios of financial crisis, is known to be complicated. Although there have been many successful studies on bankruptcy detection, seldom probabilistic approaches were carried out. In this paper we assume a probabilistic point-of-view by applying Gaussian processes (GP) in the context of bankruptcy prediction, comparing it against the support vector machines (SVM) and the logistic regression (LR). Using real-world bankruptcy data, an in-depth analysis is conducted showing that, in addition to a probabilistic interpretation, the GP can effectively improve the bankruptcy prediction performance with high accuracy when compared to the other approaches. We additionally generate a complete graphical visualization to improve our understanding of the different attained performances, effectively compiling all the conducted experiments in a meaningful way. We complete our study with an entropy-based analysis that highlights the uncertainty handling properties provided by the GP, crucial for prediction tasks under extremely competitive and volatile business environments.

1. Introduction

Finance and accounting are complex domains in which multiple components often interact, making it an active area of research for uncovering the challenges posed by the domain complexity. In addition, the challenge of accurate assessment of business failure prediction under scenarios of financial crisis is known to be complicated. A particularly disruptive case is when the markets' volatility and unpredictability introduce great uncertainty. In such scenarios, credit risk is a primary concern for banks and investors that screen firms and monitor their efforts. As a consequence, there is a great need for accurate analytical tools that are able to predict corporate bankruptcy among the companies in which investors are willing to place their investments, along with the design of early warning systems.

In the past years, weakened by the fall down of economic growth, many small and medium enterprises (SMEs) announced zero profits or losses. A large number of firms of all type of industries have become insolvent across the financial system, nationally and globally, spilling into the real economy. Most governments were forced to implement rescue plans for the banking systems with more effective credit risk assessment [1].

Two methodologies are commonly used to estimate financial distress. The first one is based on accounting information while the second regards the market information [2]. The former takes a technique that analyses financial ratios where each component is recorded as a financial statement. In this paper we take this view and look at the problem of bankrupcty prediction in terms of several financial ratios which are intrinsically linked to the financial health of the companies in study. We address this problem by taking a Bayesian perspective, applying the Gaussian process (GP) classification framework [3,4] for data inference aimed to estimate bankruptcy probabilities in a real-world dataset of thousands of French companies. Although there have been many successful studies on bankruptcy detection, seldom probabilistic approaches were carried out. Moreover, the application of GPs to the bankruptcy prediction problem is still rarely used. The GP provides a very flexible framework which, encoded with a kernel function, improves the potential of modeling complex non-linear relationships between accounting ratios and bankruptcy risk simultaneously [5].

In this work we compare the GP classifier against standard discriminative classification approaches, namely, the support vector machine (SVM) and the econometric logistic regression (LR) model. First we show that a richer and more flexible class of models, such as the GP, can improve the classification performance when com-

* Corresponding author.
E-mail addresses: fnibau@dei.uc.pt (F. Antunes), bribeiro@dei.uc.pt (B. Ribeiro), camara@dtu.dk (F. Pereira).

http://dx.doi.org/10.1016/j.asoc.2017.06.043
1568-4946/

pared with the SVM and LR. For our study, we mainly focus on a set of financial ratios from a private bankruptcy database of French companies and present a research design that includes parameter and model selection, while statistical tests assess the significance of the attained results. Additionally, we add several publicly available credit risk datasets to our experiments in order to further test our experimental design and to expand our conclusions. Second, we generate a meaningful graphical visualization that allows to understand the trade-off between metrics and the attained performances via several research designs. Finally, a sensitivity analysis is conducted by computing the probabilistic confusion entropy matrix of the GP classifier, which reinforce that the GP is successfully able to deal with uncertainty present in the studied bankruptcy datasets.

The remainder of this paper is organized as follows. In the next section we provide the problem formulation and review the relevant literature for this work, describing linked research conducted in this area. In Section 3, we briefly introduce the Gaussian Process classification framework. The description of the our experimental design is resumed in Section 4. In Section 5 the results are discussed and then some practical implications are summarized in Section 6. We finalize this paper by drawing several conclusions and by addressing future lines of work.

2. Literature review

Among many possible definitions, bankruptcy (or failure) occurs when an entity is not able to fulfill its financial obligations [6]. On the other hand, financial credit risk indicates the amount of loss that arises from a borrower not being able to pay the lenders, or declaring bankruptcy. Therefore, and focusing on the corporate case, bankruptcy risk assessment attempts to solve the following problem: given a number of companies labeled as bad/good credit or bankrupt/healthy, and a set of financial variables that describe the status of a company over a given period of time, predict the probability of these companies belonging to a high risk group or becoming bankrupt in the future. The former problem is called credit rating or scoring, whereas the latter one is called bankruptcy prediction or corporate financial distress forecast. Both are solved in a similar way as a binary classification task. Nowadays, there are many credit rating agencies that exclusively dedicate their activity to the creditworthiness and bankruptcy assessment of financial entities, such as other companies, local governments, countries, sovereign nations or even economic zones.

The earliest research on financial credit risk assessment can be traced back to FitzPatrick [7] and the well-known Altman [8] models. To date, a large variety of approaches have been employed to evaluate the creditworthiness of applicants using the traditional statistical methods or advanced machine learning techniques, each of which with particularities depending on the size, structure and nature of the dataset, model selection, type of problem (credit rating or bankruptcy prediction), among other factors. The focus of this paper is on the bankruptcy prediction problem. Although distinct, the concepts of credit risk and bankruptcy are much intertwined which explains why our review of the literature bounces between both topics. In fact, several authors address both issues within the same experimental setting [9–11].

In [12] a survey is conducted regarding the application of Machine Learning techniques within the financial crisis prediction problems in the period between 1995 and 2010. With respect to the application of general intelligent techniques to the bankruptcy prediction problem, Kirkos [13] presents a systematic literature review on the topic. Particularly focused on credit risk prediction, Galindo and Tamayo [14] conducted a comparative study of different statistical and machine learning techniques on mortgage loan data, eventually proposing a methodology based on error curves analysis. In [15], the authors present an extensive review on statistical and intelligent techniques used to predict bankruptcy in the context of banks and firms. Also in [1], a survey on the various methods to approach the financial risk assessment problem is provided. Yet another review, this time focusing on the application of evolutionary computing for credit-worthiness assessment, is conducted by Marqués et al. [16]. Additionally, a thorough number of approaches can be found, among which we highlight the following: Statistical methods [17,18], artificial neural network [19–21], decision trees [22,23], case-base reasoning [24,25], support vector machines (SVM) [26,27] and hybrid learning models [28,29].

In 2011, Chen [30] proposed hybrid adaptive-network-based fuzzy inference system (ANFIS) model to predict business failures within the Taiwanese electronics companies market. In [31], using a similar financial dataset, the author conducts an experimental comparison of several statistical, intelligent and evolutionary models, namely, linear discriminant analysis (LDA), logistic regression (LR), decision trees (C5.0 and CART), self-organizing map (SOM), learning vector quantization (LVQ), SVM, particle swarm optimization integrated with SVM (PSO-SVM) and finally genetic algorithm with SVM (GA-SVM). The results show that the traditional statistical approaches are more suited for large datasets, whereas for smaller datasets the intelligent techniques seem to be more adequate. The same author proposes in [32] a multi-phased model integrating SOM and SVM techniques for both visualization and dynamic evaluation of corporate financial structure.

In 2012, Chaudhuri and De [33] apply a Fuzzy SVM to solve bankruptcy classification problems. This enhanced approach combines the advantages of both machine learning and fuzzy sets, which compares favorably to the probabilistic neural network (PNN) in terms of clustering power. In [34], a novel Fuzzy SVM is proposed to evaluate credit risk. Recently, Cleofas-Sánchez et al. [35] explored a hybrid associative classifier with translation (HACT) neural network over nine real-world financial datasets. The authors show that associative memories can be a good approach for financial distress assessment, generally outperforming, for example, the SVM and the standard logistic regression.

Techniques based on combination of classifiers have also been suggested by several authors, such as, for example, Marqués et al. [36], Sun and Li [37] and very recently du Jardin [38]. In [39], the authors show that the AdaBoost algorithm can properly evaluate the financial risk within a universe of Korean construction companies. Tsai et al. [40] conduct a comparative study of three classifier ensembles based on several data mining and machine learning techniques. An interesting survey on hybrid and ensemble-based soft computing techniques, and their application to bankruptcy prediction, is conducted in [41].

In more recent works, a learning model with privileged information was introduced in a financial setting [27]. This transfer learning technique has attracted attention since it allows the integration of additional knowledge into the training process of a classifier, even when this comes in the form of a data modality that is not available at test time [42]. The authors in [27] showed that the advanced SVM+ approach not only significantly improves the prediction performance over the baseline SVM but also that it constitutes a better model when compared to a similar approach of multi-tasking learning (MTL).

In another line of research, Yu et al. [43] proposed an enhanced extreme learning machine (ELM) approach based on shallow neural networks which are used for feature selection. Each best neuron is selected using the PREdiction Sum of Squares (PRESS) statistic and the Leave-One-Out (LOO) method. The studied data regards 500 and 520 French retail companies operating in years 2002 and 2003, respectively. Both datasets are balanced and featuring 41 financial indicators. In [44], the author alerts for the importance of variable selection and its impact on the classification performance, showing

that a neural network achieves better prediction accuracy when the explanatory variables are selected based on a criterion specially adapted to the network, rather than those suggested in the financial literature. Deep neural networks models have also been applied to bankruptcy forecasting. In [45], a deep belief network (DBN) is proposed and tested. The authors successfully showed that the underlying financial model, enhanced with deep learning, has a great potential when compared to standard approaches such as SVM or single restricted Boltzmann machine (RBM).

Bayesian settings for bankruptcy prediction problems can be traced back to Ribeiro et al. [46] and have been recently gaining interest. This work demonstrated the competitive performance of relevance vector machine (RVM) when compared to SVMs. The approach therein used Automatic Relevance Determination (ARD) and the closely related sparse Bayesian learning (SBL) framework which are effective tools for pruning large numbers of irrelevant features leading to a sparse explanatory subset. Closely related with the SBL framework, the Gaussian Processes (GP) approach was applied to bankruptcy in [2]. The authors compared several machine learning techniques, including the fisher linear discriminant analysis (LDA) and logistic regression model, in bankruptcy prediction using US Federal Deposit Insurance Corporation data. They concluded that the GP showed a competitive classification performance with respect to the well-known approaches such as the Altman's *Z*-score or the logistic regression model.

More recently, in [5] a GP model was applied in the setting of Swedish companies and compared to the linear logistic model. The author has empirically shown the model is able to improve the highly non-linear mapping between the financial ratios and bankruptcy risk as compared to linear logistic model. A multi-class classifier GP (GPC) was used by Huang [47] for credit rating forecasting within Taiwan's security market. The results show that the GPC outperformed other traditional multi-class approaches. The data used in this study consist of 36 features for 88 Taiwanese technology companies, along the years 2000 and 2004, with five possible output ratings. In [48], the GP framework was explored in the peer-to-peer (P2P) lending market. In this approach, textual information, such as loan actions, for semantic analysis was also introduced. The authors showed that their model outperforms traditional baseline models.

In spite of the enormous volume of the related literature, our understanding is that the existing models that estimate bankruptcy probabilities need further development and more applications. In particular, the application of the GP classifier in such field is still scarce.

3. Gaussian Process classification model

The Gaussian Process (GP) framework has been recently rediscovered mostly due to the emerging of the kernel-based machine learning and artificial intelligence techniques. These models provide a full non-linear Bayesian framework and a wide range of applications, gaining much attention within many research fields. Closely following Rasmussen and Williams [3], we briefly present the GP classification framework.

A GP is a stochastic process, $f = (f_t, t \in \mathcal{T})$, with $\mathcal{T}$ a set of indexes, in which any finite set of variables forms a joint Gaussian distribution [3]. The simplicity of the GP model is that it is completely characterized by its mean and covariance (or kernel) functions, respectively denoted as $m_f(\mathbf{x})$ and $k_f(\mathbf{x}, \mathbf{x}')$, where $\mathbf{x}'$ and $\mathbf{x}'$ are two input vectors with dimension D. Thus, a Gaussian Process is simply denoted as $\mathcal{GP}(m_f(\mathbf{x}), k_f(\mathbf{x}, \mathbf{x}'))$ where

$$m_f(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$$
$$k_f(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m_f(\mathbf{x}))(f(\mathbf{x}') - m_f(\mathbf{x}'))].$$

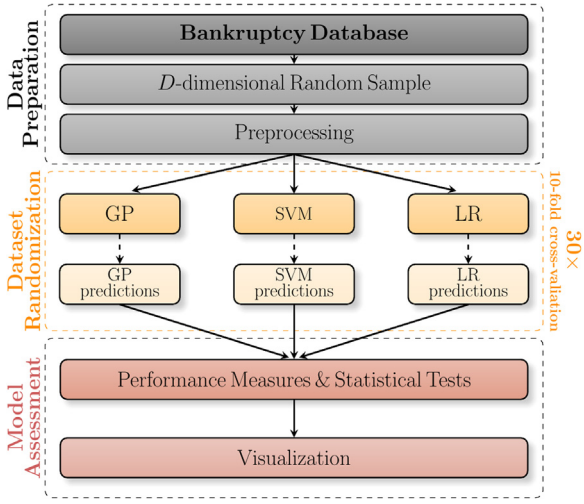

Fig. 1. Bankruptcy prediction and visualization framework.

The GP framework assumes a GP prior over functions, i.e., $y = f(\mathbf{x}) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$, $f(\mathbf{x}) \sim \mathcal{GP}(m_f(\mathbf{x}), k_f(\mathbf{x}, \mathbf{x}'))$ and y is the target variable. Fixing $m_f(\mathbf{x}) = 0$, the prior over the latent function is then given by

$$p(\mathbf{f}|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \mathcal{N}(0, K_f),$$

where $\mathbf{f} = [f_1, f_2, \ldots, f_n]^\top$, $f_i \triangleq f(\mathbf{x}_i)$ and K_f is the covariance matrix, with its elements given by $[K_f]_{ij} = k_f(\mathbf{x}_i, \mathbf{x}_j)$. Typically, the covariance function has a number of free parameters, which can be obtained directly from the training data by marginal likelihood maximization.

For prediction purposes regarding a new test point $\mathbf{x}_*$, first we compute the distribution of the latent variable corresponding to the new test case,

$$p(f_*|X, \mathbf{y}, \mathbf{x}_*) = \int p(f_*|X, \mathbf{x}_*, \mathbf{f}) p(\mathbf{f}|X, \mathbf{y}) d\mathbf{f},$$

where, $p(\mathbf{f}|X, \mathbf{y})$ is the posterior distribution over the latent variables, and then use it over f_*,

$$\bar{\pi}_* \triangleq p(y_* = +1|X, \mathbf{y}, \mathbf{x}_*) = \int \delta(f_*) p(f_*|X, \mathbf{y}, \mathbf{x}_*) df_*,$$

where $\delta(.)$ is any sigmoid function, commonly the logistic or probit functions. From $\bar{\pi}_*$ we obtain the probability of the new target variable y_* belonging to the positive class. The main difference between the GP and other discriminative classification approaches is that each class prediction comes in the form of a probability. This probabilistic classification allows us to sense how certain the model is about the state of bankruptcy of a given company. The higher the value of $\bar{\pi}_*$ is, the more likely is that the company will default, i.e., the higher probability is that the same company belongs to the positive class. The same reasoning is equally valid for credit risk assessment, where more informed decisions based on probabilities can be taken to either grant or reject a credit application. This characteristic contrasts with other type of models that only provide a single discriminative prediction with no associated uncertainty.

4. Experimental design

Our experimental design follows three main axes and can be gleaned from Fig. 1. Using a real-word and private bankruptcy data of French companies, we conduct an in-depth comparison between the GP, SVM and the LR. Moreover, in order to achieve more reliable and general conclusions, we expand our study with multiple credit

Table 1
Datasets in study.

Dataset	Size	Features	Class balance
DIANE	1334	30	50:50
	2000	30	30:70
	2000	30	20:80
Australian	690	14	50:50
German	1000	24	30:70
Japanese	653	15	50:50

risk datasets, publicity available from UC Irvine machine learning repository (UCI) [49], to further validate our comparison setting. A summary of all the used datasets is presented in Table 1. During the preprocessing stage we treat the data, remove missing values that might be present and finally normalize it. In some cases we had to additionally convert nominal attributes to discrete integer features. Then, a thoroughly and consistent design for model and parameter selection is performed, followed by analysis and discussion of the results. The model training process is conducted in a stratified 10-fold cross-validation scheme, where each model is trained/tested in parallel on the same training/testing blocks, so that the performance results are directly comparable. This process is repeated 30 times for statistical significance.

For performance assessment several well-known metrics were used, namely, accuracy, Type I and Type II errors, F1-score, Precision, recall and receiver operating characteristic (ROC) curves. Then, hypothesis statistical test were performed to assess the significance of the results. We additionally generated a complete graphical visualization of the different design experiments, datasets and results. The proposed cobweb graph effectively compiles the performance results in a meaningful and appealing way.

Finally, a GP sensitivity probability analysis was performed and an entropy-based confusion matrix is proposed unveiling the uncertainty properties nicely handled by the GP framework. The probabilistic classification provided by the GP can prove to be extremely useful for prediction tasks under extremely competitive and volatile business environments, where uncertainly can acquire its uppermost forms.

4.1. DIANE database

The main dataset used in this study was extracted from the DIANE database, containing ca. 540,000 records with several financial indicators from bankrupt and healthy companies in France, from year 2002 to 2006. A summary of all the used financial indicators is shown in Table 2. DIANE is a well-known, extensive and non-public database provided and maintained by the French credit risk provider COFACE[1] (*Compagnie Française d'Assurance pour le Commerce Extérieur*) Credit Insurance Company.

4.2. Data sampling and preprocessing

To handle the earlier mentioned large scale financial database we firstly needed to sample and preprocess the data that served as input for the models. Apart from being a very complete and comprehensive database, it also contains a sizable number of missing values. In order to avoid artificial replacement of values, we queried the database in such a way that the selected samples would not contain missing attribute values. After this, we noticed that, within the positive class, there were only 667 records without missing values thus we decided to build our samples around this number. We carried out a simple random sampling to extract

[1] http://www.coface.com/.

Table 2
DIANE database features.

Number of employees last year	Capital employed/fixed assets
Financial Debt/capital employed	Depreciation of tangible assets
Working capital/current assets	Current ratio
Liquidity ratio	Stock turnover days
Collection period days	Credit period days
Turnover per employee EUR	Interest/turnover
Debt period days	Financial Debt/equity
Financial Debt/cashflow	Cashflow/turnover
Working capital/turnover days	Net current assets/turnover days
Working capital needs/Turnover	Export
Added value per employee EUR	Total assets turnover
Operating profit margin	Net profit margin
Added valued margin	Part of employees
Return on capital employed	Return on total assets
EBIT margin	EBITDA margin

three samples with different default-healthy proportions, 667–667, 667–1333 and 400–1600, forwardly designated as data samplings 50:50, 30:70 and 20:80, respectively. Thus, two of our samples are purposely imbalanced, favoring the negative class. We provide this analysis to evaluate the models' behavior under different bankrupted/healthy proportions. By doing so, we are able to represent different financial scenarios and to mimic the possible lack of data which can occur in the real-world. Our samples consists of companies that were declared bankrupted in the beginning of 2006 and 2007, with business activity along the years 2005 and 2006, respectively. The healthy companies were randomly sampled among the ones active during the year of 2006.

Following Ribeiro et al. [27], the features were logarithmized and then linearly normalized in order not only to decrease the variability of the data distribution, but also to even the feature weights for the classification task. After these transformations the features' values range in the [0, 1] domain. The total number of observations, n, is 1334 and 2000, for the balanced and unbalanced settings, respectively. Finally, the number of features corresponding to the problem dimensionality is $D = 30$.

4.3. Classification models

As already mentioned, the classification models used in this work are the GP, SVM and LR. For the GP we used the implementation available from [3], with the Logistic likelihood function and the Laplace approximation method, whereas for the SVM and LR we used the built-in tools of Matlab [50].

For both kernel-based models (GP and SVM), we choose the Squared Exponential (SE) function, generically defined by $k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||x-x'||^2}{2\sigma^2}\right)$, where σ is the characteristic length-scale. In the particular case of the GP, it can take a more general form as follows

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{x}')^\top M(\mathbf{x}-\mathbf{x}')\right),$$

where σ_f^2 corresponds to the variance of the underlying signal function f, $M = diag(\boldsymbol{\sigma})^{-2}$ and $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \ldots, \sigma_D]^\top$ is a positive real-valued vector. Hence, it uses different length-scales on each of the D dimensions of the input $\mathbf{x}$. This kernel is often called squared exponential with automatic relevance determination (SE-ARD), as it can effectively attenuate the effect of irrelevant input features on the inference process, while giving more importance to the most relevant ones [3]. Note that the standard SE is a particular case of the SE-ARD where $\sigma_1 = \sigma_2 = \ldots = \sigma_D$.

One of the advantages of the probabilistic GP framework is, as mentioned in Section 3, its ability to obtain the kernel parameters directly from the training data via marginal likelihood optimization. Even with the SE-ARD this process is rather simple and reason-

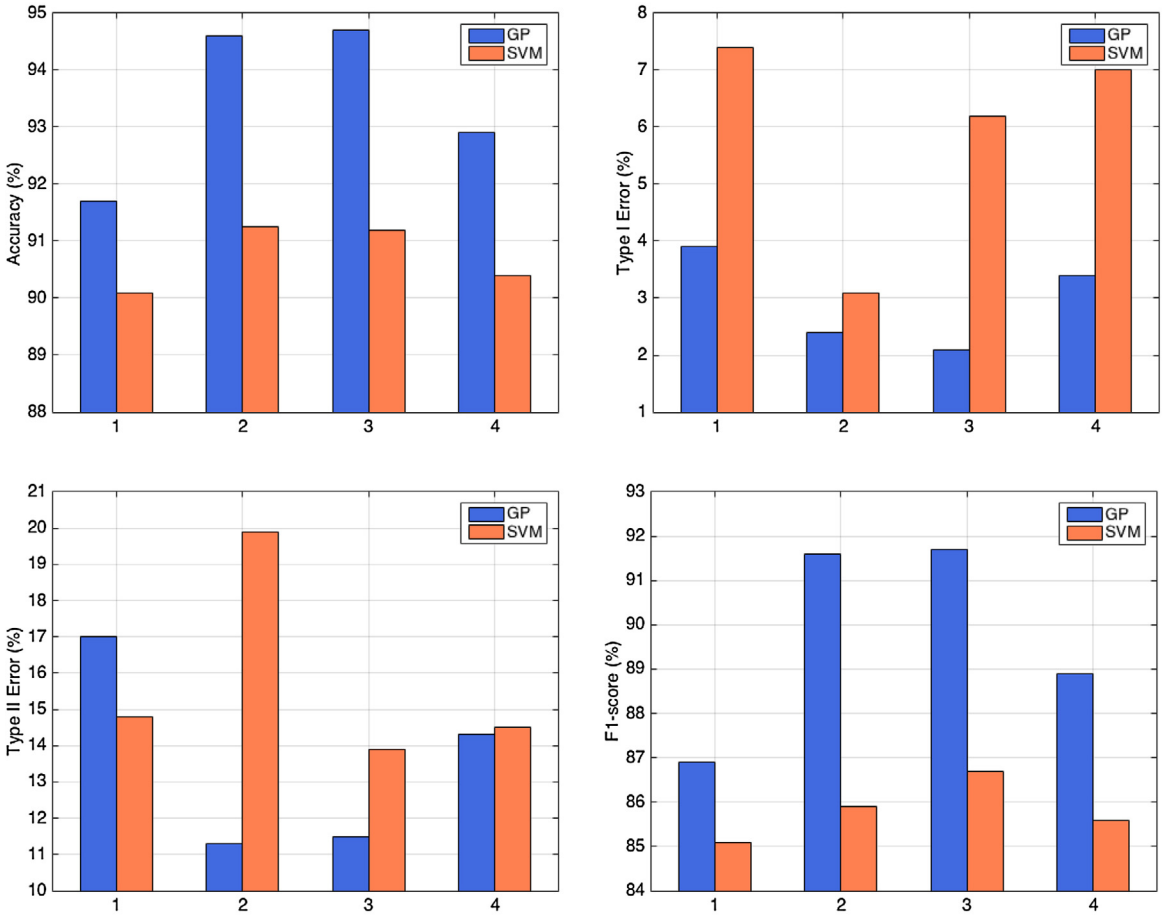

Fig. 2. Assessing the performance of prediction models using four parameter configuration scenarios for the DIANE 30–70 sampling: (1) $\sigma_f = \sigma_1 = \sigma_2 = 1$, $C = Inf$; (2) $\sigma_f = 20$, $\sigma_1 = \sigma_2 = 2$, $C = 1$; (3) $\sigma_f = 10$, $\sigma_1 = 2$, $\sigma_2 = 1$, $C = 10$ and (4) $\sigma_f = 10$, $\sigma_1 = \sigma_2 = 0.5$, $C = 10$.

ably fast. Note that while for the standard SE we have only two parameters, in the SE-ARD case we have $D+1$ parameters for which we want to optimize the likelihood function with respect to. Non-probabilistic approaches, such as the SVM, often require cross-validation schemes combined with mesh-grid searches to select the optimal set of parameters in a $(D+1)$-dimensional space. When D is large, this search can prove to be computationally very demanding, not to say virtually intractable for some extreme cases. This explains why SVMs usually implement kernel functions with a reduced number of parameters.

Overall, these kernel functions are widely known for encoding the notion of closeness, or similarity, into their structure. An expected property is that two distinct input vectors, $\mathbf{x}$ and $\mathbf{x}'$, are more likely to have similar target values if they are closer than if they are too far apart. The importance of this distance is then weighted by σ, thus having a clear role of length scaling. In our study, whenever we use the SE in its standard form, we use σ_1 and σ_2 to distinguish between the GP and the SVM kernel characteristic length-scales, respectively.

4.4. Evaluation metrics

The ultimate goal of this work is to assess the classification performances of the GP when applied to the bankruptcy data previously described and compare it with the remaining approaches. One important task when approaching binary classification problems is to clearly define which label corresponds to the positive class and, conversely, which label corresponds to the negative class. This correspondence differs from problem to problem, depending on its nature and on the consequences of a misclassification. From the investors point-of-view, the error of classifying a potential bankrupted company as healthy has a completely different consequence than the occurrence of false alarm, i.e., a healthy company misclassified as defaulted. While the former corresponds to a "credit risk", the latter can be regarded as a "commercial risk" [51]. These concepts are crucial within the financial risk assessment applications as they are able, up to a certain extent, to account for the different kind of losses incurred by the stakeholders under the two dichotomous scenarios. Note that when a healthy company is classified as bankrupted, there is no direct loss for the investor. Nevertheless, this type of misclassification will eventually keep potential investor from applying their financial resources in a good business opportunity. On the other hand, a misclassified healthy company can suffer from lack of investment. From now on we define the state of bankruptcy to be our positive class, i.e., it is labeled by "+1" or simply "1".

With the definition of the positive class, the interpretation of the classification metrics gain a real practical meaning. The Accuracy refers to the overall correct matches, which simply cannot be used for classification quality assessment. The Type I error indicates a misclassification of a healthy company, whereas the Type II error indicates, conversely, the misclassification of a default company. The F1-score quantifies the trade-off between Precision and Recall. All these metrics range from 0 to 1. Ideally, both accuracy and F1-score should be as close to 1 as possible, whereas the Type I and II errors should lie close to 0. Finally, the ROC curve, which depicts the trade-off between true positive and false positive rates, provides

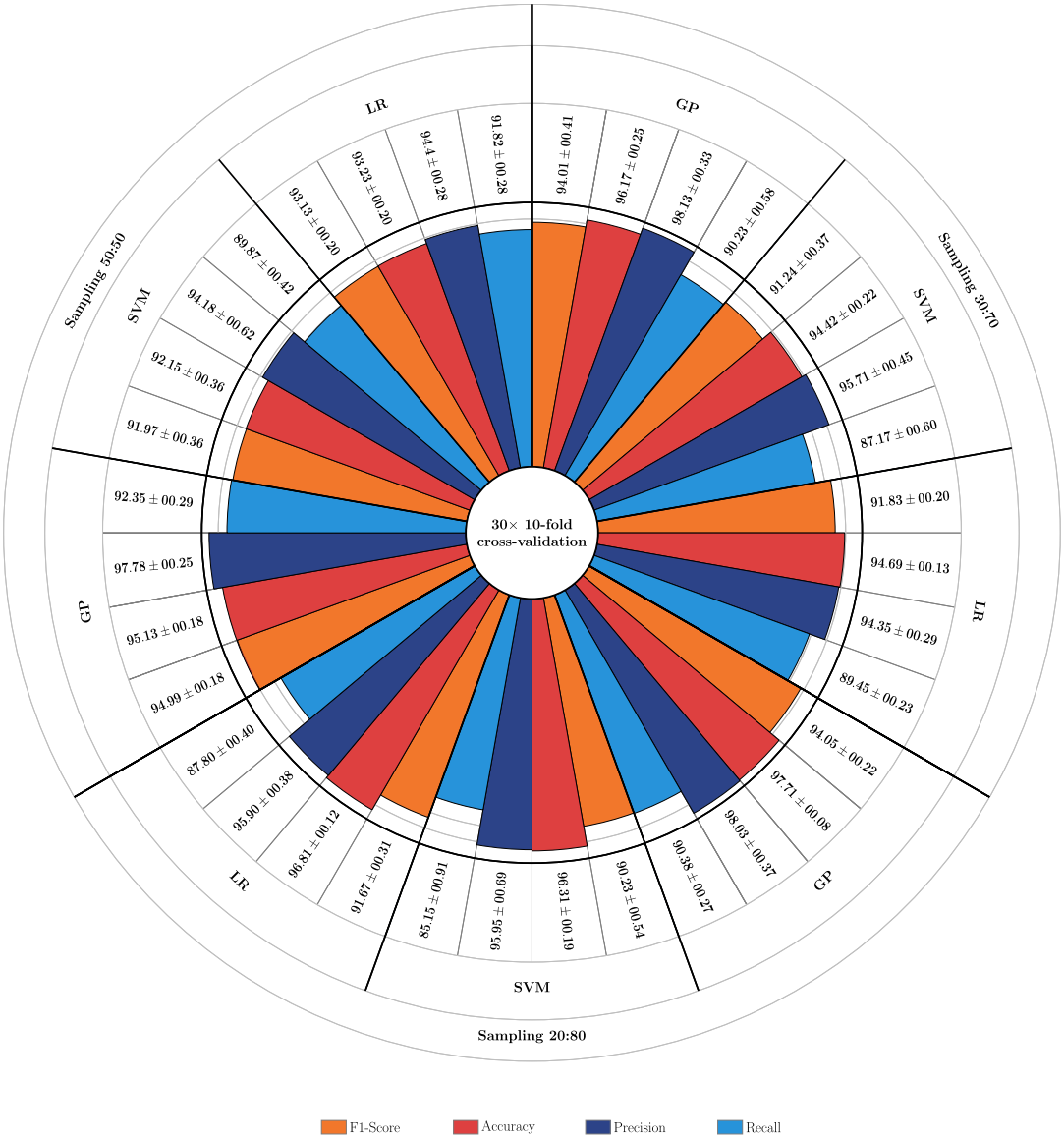

Fig. 3. Performance evaluation (%) in DIANE for Bankruptcy prediction.

a straightforward way to compare the performance of different classifiers.

5. Results and discussion

In this section we present and discuss the results by comparing the GP, SVM and LR models. Our design is divided in two major sets of experiments, in particular, model selection and parameter selection.

5.1. Model selection

In a first set of experiments we conduct several model runs according to different parameter configurations, whose results are presented in Fig. 2. Generally, we observe that, with respect to all the presented metrics, the GP outperforms the SVM for the selected kernel parameters, which we forwardly refer as parameter scenarios. We selected four parameter configuration scenarios based on multiple preliminary experiments. For each metric, each colored pair of bars corresponds a single parameter configuration for both models. For instance, regarding the F1-score, the configuration of the parameters for the second scenario is $\sigma_f = 20$, $\sigma_1 = \sigma_2 = 2$ and $C = 1$, where C corresponds to the SVM trade-off error margin constant. Here we used the standard SE as the kernel function for the GP to make the graphical illustrations possible. The LR is not included in this comparison, as it is not a kernel-based model. For scenarios 2 and 3, it is interesting to see that although both models achieved, respectively, similar values for accuracy and F1-score, the same did not occurred to the Type I and Type II errors. In scenario 2, the SVM reached a Type I Error around 3% against 2.5% for the GP. This means that the SVM has 0.5% more false alarms than the GP. However, such difference is much more evident in scenario 3. In this case, the error increased to 6%, while the GP approximately maintained its former performance. Conversely, for the false positive rate, similar conclusions are drawn. Recall that this second type of misclassification represents the rate of missed positives, i.e., companies that go bankrupt and which are incorrectly classified as healthy. The difference between the models is most noticeable in scenario 2 ($\sigma_f = 20$, $\sigma_1 = \sigma_2 = 2$ and $C = 1$) evidencing the superiority of GP for all metrics. The GP was slightly outperformed by the SVM only in scenario 1, by 2%, on the Type II error. In scenario 4 ($\sigma_f = 10$, $\sigma_1 = \sigma_2 = 0.5$ and $C = 10$), both models show competitive performance. It is also wor-

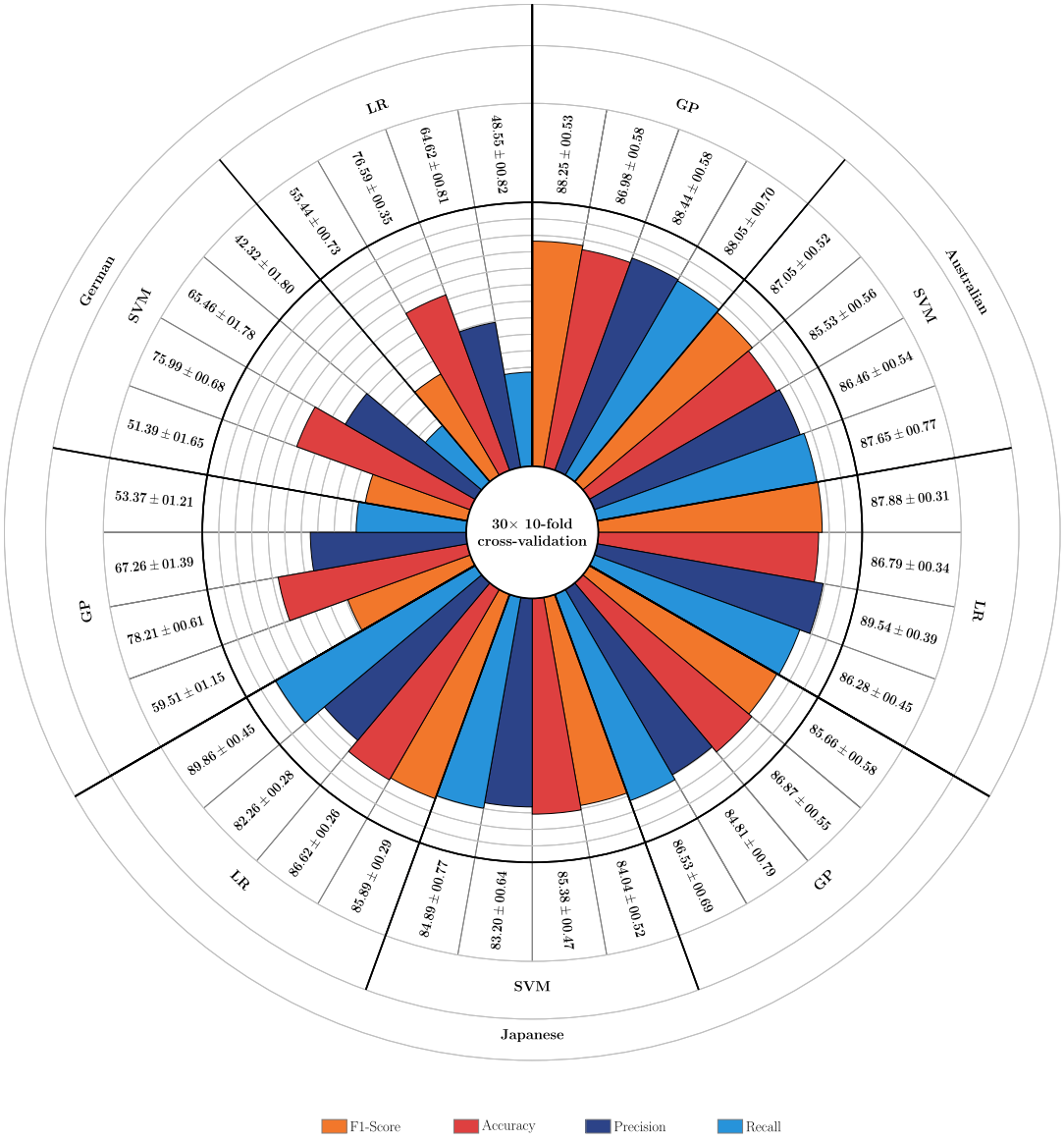

Fig. 4. Performance evaluation (%) in the Australian, German and Japanese credit risk datasets.

Table 3
Type I and Type II errors obtained with GP, SVM and LR for DIANE Bankruptcy datasets.

Dataset	Error	Models		
	(%)	GP	SVM	LR
50:50	Type I	02.09 ± 00.24	05.55 ± 00.63	05.35 ± 00.29
	Type II	07.64 ± 00.29	10.12 ± 00.42	08.17 ± 00.28
30:70	Type I	00.85 ± 00.15	01.95 ± 00.21	02.68 ± 00.14
	Type II	09.76 ± 00.58	12.82 ± 00.60	10.54 ± 00.23
20:80	Type I	00.45 ± 00.08	00.89 ± 00.16	00.93 ± 00.08
	Type II	09.61 ± 00.27	14.84 ± 00.91	12.19±00.40

Table 4
Type I and Type II errors obtained with GP, SVM and LR for credit risk datasets.

Dataset	Error	Models		
	(%)	GP	SVM	LR
Australian	Type I	14.35 ± 00.78	17.12 ± 00.76	12.57 ± 00.52
	Type II	11.94 ± 00.70	12.34 ± 00.77	13.71 ± 00.45
German	Type I	11.13 ± 00.60	09.57 ± 00.67	11.39 ± 00.35
	Type II	46.62 ± 01.21	57.67 ± 01.80	51.44 ± 00.82
Japanese	Type I	12.84 ± 00.77	14.21 ± 00.64	16.06 ± 00.29
	Type II	13.46 ± 00.69	15.10 ± 00.77	10.13 ± 00.45

thy of note that, for the accuracy and F1-score metrics, both models present similar behavior across the four scenarios.

We now proceed to the core set of experiments of our study, according to the scheme represented in Fig. 1. In Figs. 3 and 4 we present the mean and standard deviation values of the selected measures obtained from 30 runs, for the different datasets in study. To complement these results, Tables 3 and 4 summarize the Type I and Type II errors generated from the same runs.

For the DIANE data, we observe the superiority of the GP over the SVM and the LR for all the key performance indicators, in all the settings (see Fig. 3). Particularly for the 30:70 sampling and with respect to the SVM, the GP improved the accuracy and F1-score by 1.75% and 2.77%, respectively. For the LR, the GP improved the same metrics by 1.48% and 2.18%, respectively. We can also observe that the latter model presents a similar performance to the SVM, although a slightly better one. In Table 3, we can observe that, in terms of the Type I error, the GP averaged 00.85%, while the SVM and

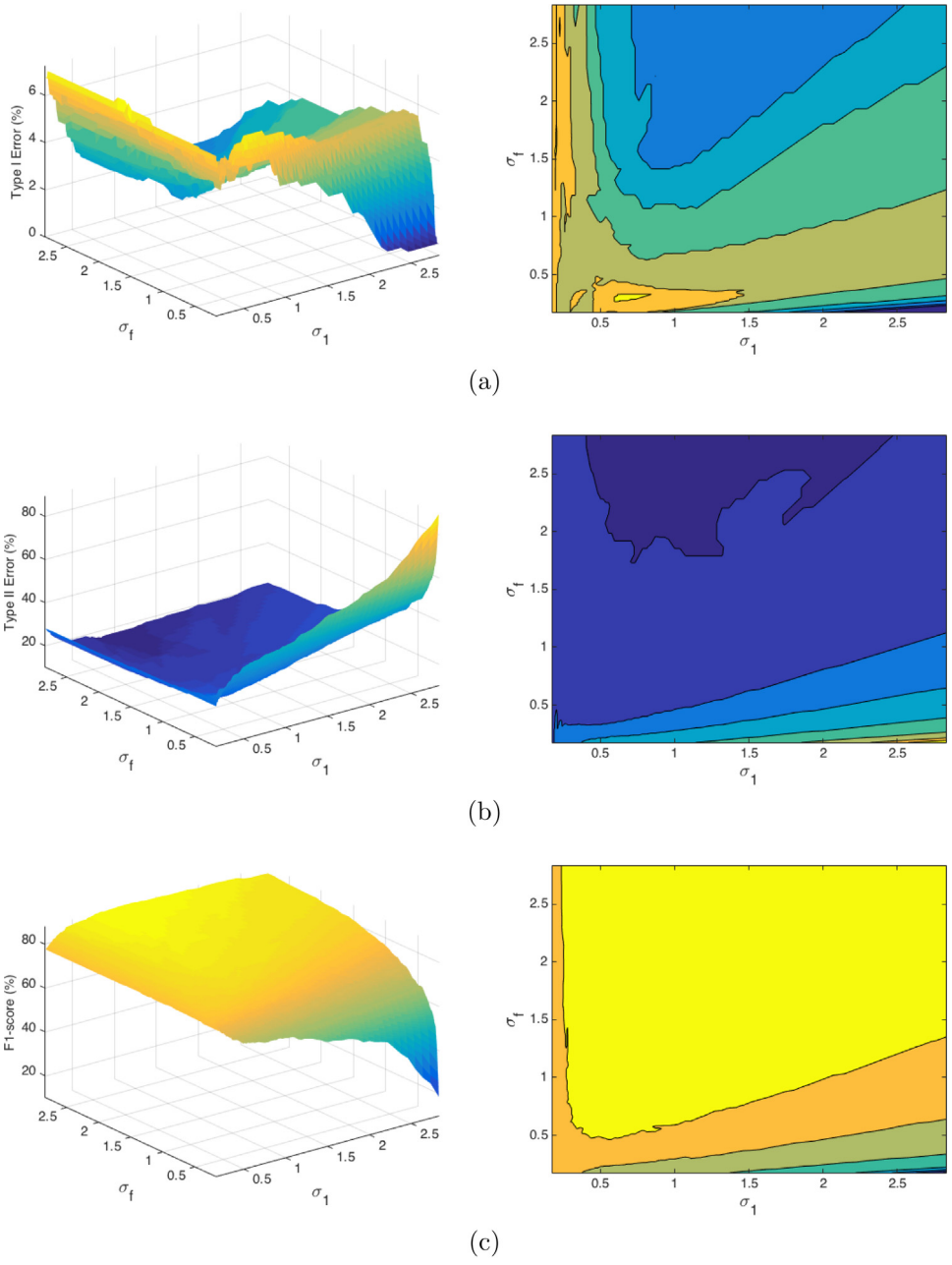

Fig. 5. Type I error (a), Type II error (b), F1-score (c) and the corresponding contours plots of the GP w.r.t. σ_1 and σ_f for the DIANE 30:70 sampling.

LR reached 1.95% and 2.68%, respectively. This difference is more evident for the second type of misclassification errors, where the GP attained 9.76% against 12.82% from the SVM and 10.54% from the LR.

Within the 50:50 sampling and in the GP case, we observe that the F1-score increased 1%, while the accuracy decreased about the same percentage. For the SVM, a comparable pattern occurred as the F1-score slightly increased from 91.24% to 91.97%, while the Accuracy decreased almost 2.30%, from 94.42% to 92.15%. For the misclassification errors we observe that the GP still clearly outperforms the other models. For the remaining dataset sampling, similar conclusions can be drawn. In the case of the LR, a similar behavior occurred.

The effect of the class balance in the prediction performances is quite clear, specially in the misclassification errors, as seen in Tables 3 and 4. Recovering the interpretation of these misclassification errors from Section 4.4, although both type of errors are equally important, Type II error should be analyzed with particular care as it translates to the classification of a default company as healthy and, consequently, to potentially disastrous investments. On the other hand, Type I can be viewed as an opportunity loss. Notice that the Type II error increased as the sampling ratio evolved unfavorably w.r.t. the positive (default) class. Nevertheless, the GP generally showed a more stable performance behavior. In the 30:70 and 20:80 data samplings the Type II error remained virtually unchanged, while for the SVM and LR models it increased around 2%. Even with less training points for the positive class, the GP was able to maintain a very good classification performance and quite comparable to the balanced dataset. We can additionally observe that the Type I error decreased as the proportion of default companies increased. This was an expected behavior, since an increased number of default observations also means more training points

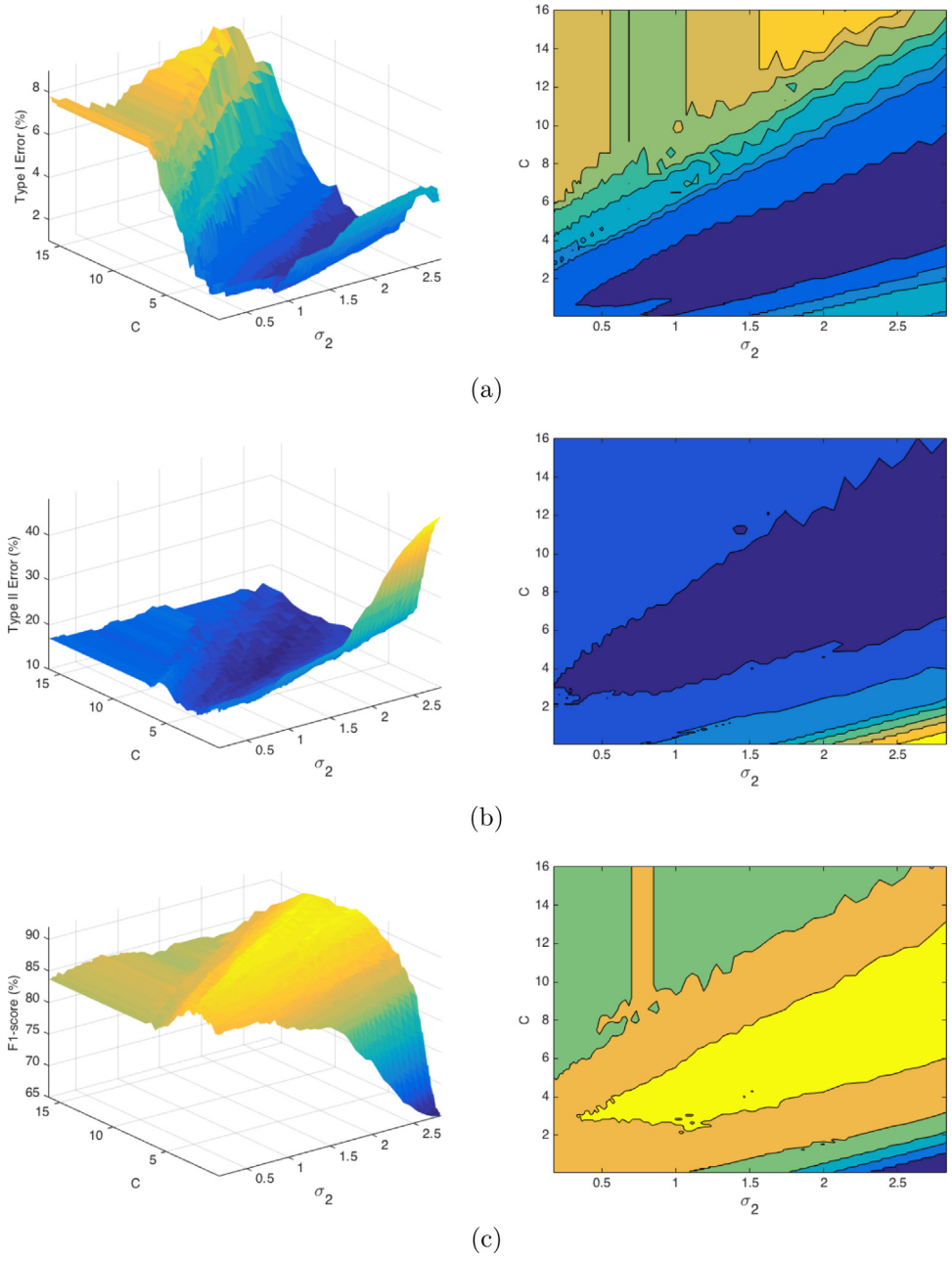

Fig. 6. Type I error (a), Type II error (b), F1-score (c) and the corresponding contours plots of the SVM w.r.t. σ_2 and C for the DIANE 30:70 sampling.

and, therefore, an improved classification performance for that particular class.

For the Australian, German and Japanese datasets, the performances were generally worse, as seen in Fig. 4, although the GP continued to show superior performance across all the depicted metrics. By observing Table 4, we conclude that the magnitudes of the misclassification errors are weakly comparable to the ones attained for the DIANE data. In any case, the GP generally achieved the lowest values. However, it was outperformed by the SVM in the Type II, with a difference of 1.56%, in the German dataset, and by the LR in terms of the Type I and Type II errors, for the Australian and Japanese datasets, respectively.

5.2. Parameter selection

In the second set of experiments we analyzed the effect of the variation of kernel parameters on the classification performance according to a 70–30% training-testing split. For space saving reasons, this study was only conducted only for the DIANE 30:70 sampling. Figs. 5 and 6 illustrate the variations of the different parameters for GP and SVM, respectively, for a given parameter range. This range was approximately based on a mesh-grid of discrete steps of powers of two, where $(\sigma_f, \sigma_1) \in [0, 3]^2$ for the GP, and $(\sigma_2, C) \in [0, 3] \times [0, 20]$ for the SVM. Again, for sake of graphical representation, we used the standard SE for both models.

Although GP and SVM are based on the same kernel function, it is a rather challenging task to compare both methods. In fact, they are fundamentally different in their inner conceptual structure. On the one hand, in the GP we use purely a probabilistic and Bayesian approach based on the generalization of the multivariate linear regression and, on the other hand, in the SVM, we have a non-probabilistic model (in its original form) based on a maximum margin optimization problem. Moreover, there is one specific parameter for each model. For the former we have σ_f and, for the

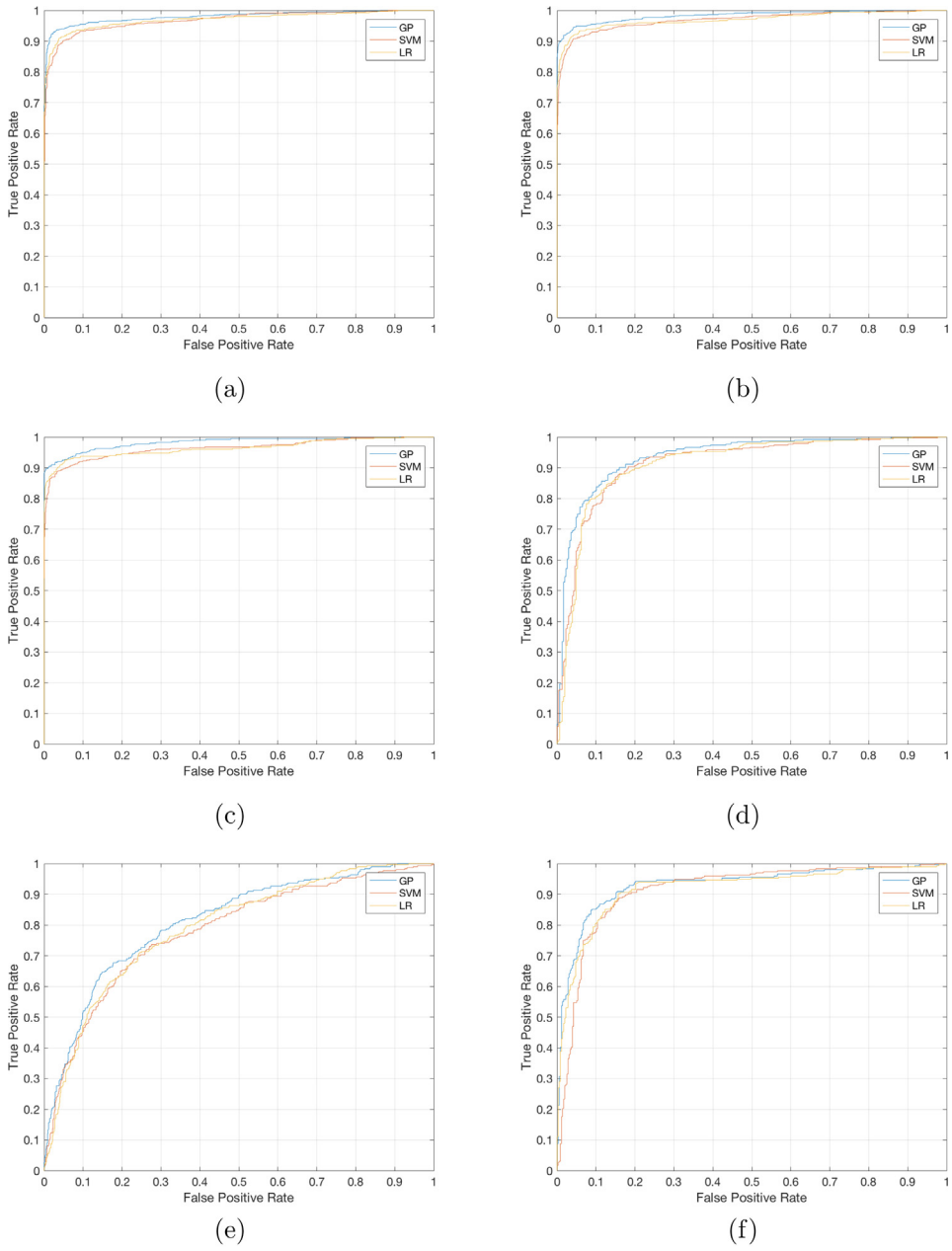

Fig. 7. ROC curves for the DIANE database samplings, 50:50 (a), 30:70 (b) and 20:80 (c), and for the Australian (d), German (e) and the Japanese (f) datasets.

latter, we have the parameter controlling the margin, C, besides the SE parameter which is shared by both.

It is interesting to observe the different behaviors across the different metrics. For the Type I error, note that the lowest values are achieved within the range of $1.5 < \sigma_1 < 3$ and σ_f close to zero for the GP, while for the SVM the best parameters lie around $1 < \sigma_2 < 3$ and $0 < C < 8$. Regarding the second type of misclassification errors we observe similar patterns. For the GP, this error tends to increase as σ_1 increases and σ_f is close to zero. In the SVM case, a similar conclusion is trivially obtained.

Finally, for the F1-score a similar analysis applies, although the GP shows a more stable and uniform behavior than the SVM. The worst performance values are also reached in the parameters range consistent with all the other metrics.

In Fig. 7 we can observe an overall performance of the three classifiers presented in this work for all the studied datasets. For each cut-off value of the ROC curves, the estimated proportion of the companies incorrectly classified as default (false positive) versus the corresponding proportion of true positives is given. The results show the superiority of the GP over the SVM and LR. Nevertheless, these two latter models still achieve excellent results. Again, we can graphically observe that the three models performed noticeably worse in the Australian, German and Japanese datasets.

5.3. Tests for statistical significance

It is important that bankruptcy prediction is built over effective and coherent financial models. Thus, in order to ensure consistent decisions and conclusions, the obtained results should be statistically significant.

As pointed out by Demšar [52], the widely used t-test it not appropriate for performance comparison when dealing with several classifiers over multiple datasets. Instead, the non-parametric tests, such like the Friedman test [53], should be conducted. This is

Table 5
Friedman test at 5% with statistical variables F1-score, Type I error and Type II error, and accuracy. For three classifiers and six datasets the corresponding critical value is 42. The null-hypothesis is rejected for the cases in bold.

Variable	Friedman statistic
F1-score	**62.00**
Type I error.	26.00
Type II error.	**50.00**
Accuracy	**72.00**

Table 6
Absolute pairwise rank differences from the multiple comparisons post-hoc test at 5% with statistical variables F1-score, Type II error and accuracy with critical values 3.15, 4.67 and 0, respectively.

Variable	GP – SVM	GP – LR	SVM – LR
F1-score	11	4	7
Type II error	10	5	5
Accuracy	12	6	6

a kind of omnibus statistical test, where the null hypothesis states that all classifiers performed equally w.r.t. a specific classification metric. The rejection of this hypothesis means that at least one pair of classifiers performed significantly different from the others. Then, in case of rejection, a post-hoc test should be conducted to further determine which pairs are significantly different in terms of performance.

In this paper, we used a Friedman test implementation for small size samples [54], which applies a Multiple Comparisons post-hoc test, closely following Conover [55]. We conducted these tests for several variables of interest, namely, Accuracy, Type I and Type II errors, and F1-score. Table 5 presents the results for the Friedman test. For F1-score, Type II error and Accuracy, the corresponding obtained Friedman statistic is greater than its critical value at a significance level of 5%, which lead us to reject the null-hypothesis in these cases. On the other hand, for our experimental setting, we can see that the test was not able to detect significant differences for the Type I error measure. This may be related with the fact that only three classifiers are being compared [55]. Recall that the Type I error refers to the misclassification of a healthy company. It is not desirable for a model to yield high values for this type of error, as it will keep the potential investors away from investing in a good company. Conversely, a high value of Type II error can itself be regarded as a credit or cost risk, as it means that the stakeholders are investing in what appears to be an healthy company. In our study, a false positive is more desirable than a false negative. While Type I errors will keep investors away from good business opportunities, Type II errors may lead to bad investments.

Following the rejection of the null-hypothesis, Table 6 shows the post-hoc results for F1-score, Type II error and accuracy. For each pair of classifiers, their performances are considered statistically different if their corresponding absolute rank differences (provided by the Friedman test) is greater than the associated critical value. We can thus conclude that the performances of the studied models are statistically different in terms of F1-score, Type II error and accuracy.

5.4. Probability analysis

As mentioned earlier, the GP allows the classification of a new test point (new company to be classified) in terms of the estimated probability, given by $\bar{\pi}_*$, of it belonging to the positive class. Thus, each prediction comes in a form of a probability distribution, providing us with a probabilistic classification, rather than "hard" point-wise prediction with no associated uncertainty. This constitutes a strong point of the GP classification framework within the business analytics applications, as we forwardly present.

Table 7
Average probability confusion matrices for the DIANE, Australian, German and Japanese datasets over 30 runs.

Dataset	Real class	Predicted class	
		Default	Healthy
DIANE			
50:50	Default	0.9343	0.2165
	Healthy	0.3181	0.9235
30:70	Default	0.9440	0.1959
	Healthy	0.3344	0.9506
20:80	Default	0.9380	0.1592
	Healthy	0.3406	0.9703
Australian	Default	0.8986	0.2811
	Healthy	0.2525	0.8687
German	Default	0.6858	0.2803
	Healthy	0.3657	0.8214
Japanese	Default	0.8682	0.2528
	Healthy	0.2911	0.9015

To analyze the probability sensitivity of the GP in our bankruptcy prediction context, we compute a probabilistic confusion entropy matrix based on the model estimated probabilities, following a similar approach of [56]. For our matrix, we replaced the usual values of true/false positives/negatives by their corresponding averaged probabilities at each test point, as summarized in Table 7. In our case the values correspond to averages computed over the 30 cross-validation runs and also over each 10-fold cross-validation scheme. Thus, each entry of the matrix depicted in Table 7 is an average of averaged probabilities.

Regarding the DIANE 30:70 dataset, we can observe that the correct classifications yield an average probability between 0.94 and 0.95, whereas for the false alarms and false negatives, it decreased to approximately 0.33 and 0.19, respectively. For the financial decision maker, this is an important property as it allows to sense how confident the GP is about its predictions. Note that high values of $p(y_* = +1|X, \mathbf{y})$ mean greater probability of the new test point belonging to the positive class. This provides an additional level of information that enhances both the decision making and taking processes. Remember that this constitutes a conditional probability with respect to the observed data.

In the 50:50, case a similar pattern occurred. However, we can observe that the probability of the correct classification of the healthy companies decreased slightly (about 2.8 %) in comparison with the 30:70 sample. On the other hand, in the 20:80 dataset the same probability raised to about 0.97. In terms of the misclassifications of default companies (as healthy), note that the probability increased from 0.15 (20:80) to 0.19 (30:70) and 0.21 (50:50), thus showing that the class balance has a significant effect on the attained class probabilities. This is rather expected phenomenon as the proportion of healthy and default data points has a direct impact on the model training and on its generalization capabilities. Across all the samplings, the correct classification of bankrupted companies present rather similar probabilities.

For the remaining datasets we observe that the attained probabilities are in accordance with the classification performances shown in Fig. 4 and Table 4. We can conclude that higher classification performances are generally associated with higher probabilities values for the correct classifications and lower ones for the misclassifications. Again, the effect of the class balance is visible. Note that in the German dataset, where the class balance is near a 30:70 ratio, favoring the healthy class, the correct classifications achieved a probability of 0.82 and 0.68 for the healthy and default classes, respectively. On the other hand, for the class balanced Australian dataset, the same probabilities raised to 0.86 and

0.89, respectively. A similar behavior is observed for the Japanese dataset. We also observe that the probability corresponding to the Type I error is significantly bigger in the German dataset than in the remaining ones, while for the Type II error they are quite comparable. Conversely, the probabilities of the true positive classifications are the lowest in this dataset.

6. Managerial implications

Probabilistic classification represents a powerful tool for bankruptcy prediction, particularly for more conservative investors whose main objective is, by all means, to minimize the risk and gain control of the uncertainty associated with their future investments. One of the great advantages of the GP framework is that each prediction comes in the form of a predictive distribution, which contrasts with pointwise prediction outputs from other types of classification models such as, for instance, SVM or neural networks. In fact, for any given company it is possible not only to guess its future financial situation through a pointwise forecast, but also to infer the probability of this prediction. Note that the generalization step from the training cases to any test point inevitably involves a certain degree of uncertainty, which in turn is encoded in the form of a probability distribution. It is perfectly desirable that, in a highly competitive and volatile business environment, each prediction is generated in a probabilistic form that somehow quantifies the uncertainty associated with both the generalization process and the market complex behavior. Therefore, the GP framework is able to provide additional insights, enclosing them into a probability distribution which then handles the uncertainty present in the available financial data. This ability can prove to be crucial for both decision making and taking processes, potentially protecting the stakeholders from undesirable situations of bankruptcy or credit loss.

Another clear advantage of the GPs is its ability to extract the kernel parameters directly from the training data in a rather effortless way. Such extraction is conducted through the optimization of the likelihood function, constituting a clear inherited characteristic of the Bayesian formalism. The optimal parameters are the ones that maximize the likelihood for each training block. This fact also facilitates the incorporation of more complex kernel functions. A straightforward generalization of the SE is its ARD variant, which is able to determine the most relevant features during the training process, by assigning a unique length-scale to each of the input dimensions. Within the GP framework we can easily avoid, at least in a first approach, the usual steps of feature selection and extraction, naturally involved in dimensionality reduction processes. Thus, the GP can simultaneously and effectively combine the flexibility of a kernel-based machine learning technique with an automatic relevance determination approach.

We also noticed that the GP seems to be more stable with respect to the class balance, in comparison with the remaining analyzed approaches. Even decreasing the number of default cases, the GP was able to maintain acceptable values of the F1-score and Type II error, making it a reliable bankruptcy prediction tool, even when facing unbalanced or even shortage of data.

7. Conclusions and future work

As a result of the recent world-wide financial crisis and economic recession, the demand for bankruptcy prediction models and financial risk analysis have gained strong attention. The inability to accurately predict both bankruptcy and credit risk can cause devastating socio-economic effects. Therefore, it is important to provide financial decision makers with effective predictive power to anticipate these loss scenarios.

In this work we empirically compared three different classification models, namely, Gaussian Process (GP), support vector machines (SVM) and logistic regression (LR), in a setting of real-world bankruptcy data from the French market. The data is composed by 30 financial ratios that are used as explanatory variables which can account for the state of bankruptcy within a certain time period. We focused our study on three data samplings randomly extracted from the DIANE database. These datasets contain financial records from companies with business activity between years 2006 and 2007. Two of these samplings were purposely set to be imbalanced favoring the healthy class, not only to represent realistic real-world scenarios but also to mimic the possible lack of data that often occurs in this research domain. Moreover, we added three publicly available credit risk datasets to our experiments in order to further extend our study.

The probabilistic GP classifier used in this paper showed to be superior in comparison with both SVM and LR methods in a broad range of studied scenarios and datasets. Moreover, regarding the DIANE data, the GP proved to be less sensitive to the class balance, maintaining a comparable performance to that of the balanced dataset. For the credit risk datasets, the results were generally worse across all the models. However, in the majority of the cases, the GP proved to have higher classification performance. Although we believe that the research design held in this work allows for a fair model comparison, it proved to be quite a demanding task due to the fact that the considered models, mainly the GP and SVM, represent different modeling paradigms and mathematical structures. We addressed the problem of parameter selection which has a distinctive impact in the final performance behavior in the kernel-based models. Furthermore, we generated a visual representation for each design experiment which includes all the sampling datasets and all the metrics (mean and standard deviation) in a meaningful graphical visualization that allows to understand the trade-off between the metrics and the attained model performances.

As a future line of work we plan to extend and deepen the present study. It would be interesting to try another set of kernel functions, which may embed distinct similarity between data points and additional likelihood functions. Another extension is to consider not only random datasets extracted from the DIANE database but to actually perform the same set of experiences in distinct financial datasets with different ratios of healthy/bankrupted companies. Although seemingly not an easy task, we would like to apply the same experimental setup to more domains and consider mismatch distributions, for instance, another set of companies from other countries, within another economic zones. This would be a good assessment of the generalization power we believe the GP model has.

Acknowledgements

We gratefully acknowledge the useful comments and suggestions of the associate editor and the anonymous reviewers that helped to improve the paper.

References

[1] N. Chen, B. Ribeiro, A. Chen, Financial credit risk assessment: a recent review, Artif. Intell. Rev. (2015) 1–23, ISSN 0269-2821.
[2] T. Pena, S. Martinez, B. Abudu, Bankruptcy prediction: a comparison of some statistical and machine learning techniques, in: H. Dawid, W. Semmler (Eds.), Computational Methods in Economic Dynamics, vol. 13 of Dynamic Modeling and Econometrics in Economics and Finance, Springer, Berlin, Heidelberg, 2011, pp. 109–131, ISBN 978-3-642-16942-7.
[3] C.E. Rasmussen, C. Williams, Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning), The MIT Press, 2005.
[4] C. Bishop, Model-based machine learning, Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. 371 (1984).

[5] M.N. Seidu, Predicting Bankruptcy Risk: A Gaussian Process Classification Model, Linköping University, 2015 (Master's thesis).
[6] T.C. Lim, L. Yun, S. Gan, H. Jiang, Bankruptcy prediction: theoretical framework proposal, Int. J. Manag. Sci. Bus. Res. 1 (9) (2012) 69.
[7] P.J. FitzPatrick, A comparison of the ratios of successful industrial enterprises with those of failed companies, J. Acc. Res. 10 (1932) 598–605.
[8] E.I. Altman, Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, J. Finance 23 (4) (1968) 589–609.
[9] A.F. Atiya, Bankruptcy prediction for credit risk using neural networks: a survey and new results, IEEE Trans. Neural Netw. 12 (4) (2001) 929–935.
[10] C.-F. Tsai, J.-W. Wu, Using neural network ensembles for bankruptcy prediction and credit scoring, Expert Syst. Appl. 34 (4) (2008) 2639–2649.
[11] T. Van Gestel, B. Baesens, J.A. Suykens, D. Van den Poel, D.-E. Baestaens, M. Willekens, Bayesian kernel based classification for financial distress detection, Eur. J. Oper. Res. 172 (3) (2006) 979–1003.
[12] W.-Y. Lin, Y.-H. Hu, C.-F. Tsai, Machine learning in financial crisis prediction: a survey, IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) 42 (4) (2012) 421–436.
[13] E. Kirkos, Assessing methodologies for intelligent bankruptcy prediction, Artif. Intell. Rev. 43 (1) (2015) 83–123.
[14] J. Galindo, P. Tamayo, Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications, Comput. Econ. 15 (1–2) (2000) 107–143.
[15] P.R. Kumar, V. Ravi, Bankruptcy prediction in banks and firms via statistical and intelligent techniques – a review, Eur. J. Oper. Res. 180 (1) (2007) 1–28.
[16] A. Marqués, V. García, J. Sanchez, A literature review on the application of evolutionary computing to credit scoring, J. Oper. Res. Soc. 64 (9) (2012) 1384–1399.
[17] M. Kouki, A. Elkhaldi, Toward a predicting model of firm bankruptcy: evidence from the Tunisian context, Middle Eastern Finance Econ. 14 (2011) 26–43.
[18] W. Kwak, Y. Shi, G. Kou, Bankruptcy prediction for Korean firms after the 1997 financial crisis: using a multiple criteria linear programming data mining approach, Rev. Quant. Finance Acc. 38 (4) (2012) 441–453, ISSN 0924-865X.
[19] N. Chen, B. Ribeiro, A. Vieira, A. Chen, Clustering and visualization of bankruptcy trajectory using self-organizing map, Expert Syst. Appl. 40 (1) (2013) 385–393, ISSN 0957-4174.
[20] T. Korol, Early warning models against bankruptcy risk for Central European and Latin American enterprises, Econ. Modell. 31 (2013) 22–30, ISSN 0264-9993.
[21] S. Chakraborty, S.K. Sharma, Prediction of corporate financial health by Artificial Neural Network, Int. J. Electron. Finance 1 (4) (2007) 442–459.
[22] H. Li, J. Sun, J. Wu, Predicting business failure using classification and regression tree: an empirical comparison with popular classical statistical methods and top classification mining methods, Expert Syst. Appl. 37 (8) (2010) 5895–5904, ISSN 0957-4174.
[23] D. Delen, C. Kuzey, A. Uyar, Measuring firm performance using financial ratios: a decision tree approach, Expert Syst. Appl. 40 (10) (2013) 3970–3983, ISSN 0957-4174.
[24] C.-S. Park, I. Han, A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction, Expert Syst. Appl. 23 (3) (2002) 255–264, ISSN 0957-4174.
[25] H. Jo, I. Han, H. Lee, Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis, Expert Syst. Appl. 13 (2) (1997) 97–108, ISSN 0957-4174.
[26] J.K. Bae, Predicting financial distress of the South Korean manufacturing industries, Expert Syst. Appl. 39 (10) (2012) 9159–9165, ISSN 0957-4174.
[27] B. Ribeiro, C. Silva, N. Chen, A. Vieira, J. Neves, Enhanced default risk models with SVM+, Expert Syst. Appl. 39 (2012) 10140–10152.
[28] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (3) (1999) 264–323, ISSN 0360-0300.
[29] C.-F. Tsai, W. Eberle, C.-Y. Chu, Genetic algorithms in feature and instance selection, Knowl. Based Syst. 39 (2013) 240–247, ISSN 0950-7051.
[30] M.-Y. Chen, A hybrid ANFIS model for business failure prediction utilizing particle swarm optimization and subtractive clustering, Inf. Sci. 220 (2013) 180–195.
[31] M.-Y. Chen, Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches, Comput. Math. Appl. 62 (12) (2011) 4514–4524.
[32] M.-Y. Chen, Visualization and dynamic evaluation model of corporate financial structure with self-organizing map and support vector regression, Appl. Soft Comput. 12 (8) (2012) 2274–2288.
[33] A. Chaudhuri, K. De, Fuzzy support vector machine for bankruptcy prediction, Appl. Soft Comput. 11 (2) (2011) 2472–2486, ISSN 1568-4946.
[34] Y. Wang, S. Wang, K.K. Lai, A new fuzzy support vector machine to evaluate credit risk, IEEE Trans. Fuzzy Syst. 13 (6) (2005) 820–831.
[35] L. Cleofas-Sánchez, V. García, A. Marqués, J. Sánchez, Financial distress prediction using the hybrid associative memory with translation, Appl. Soft Comput. 44 (2016) 144–152.
[36] A. Marqués, V. García, J.S. Sánchez, Exploring the behaviour of base classifiers in credit scoring ensembles, Expert Syst. Appl. 39 (11) (2012) 10244–10250.
[37] J. Sun, H. Li, Financial distress prediction using support vector machines: ensemble vs. individual, Appl. Soft Comput. 12 (8) (2012) 2254–2265.
[38] P. du Jardin, A two-stage classification technique for bankruptcy prediction, Eur. J. Oper. Res. 254 (1) (2016) 236–252.
[39] J. Heo, J.Y. Yang, AdaBoost based bankruptcy forecasting of Korean construction companies, Appl. Soft Comput. 24 (2014) 494–499, ISSN 1568-4946.
[40] C.-F. Tsai, Y.-F. Hsu, D.C. Yen, A comparative study of classifier ensembles for bankruptcy prediction, Appl. Soft Comput. 24 (2014) 977–984, ISSN 1568-4946.
[41] A. Verikas, Z. Kalsyte, M. Bacauskiene, A. Gelzinis, Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey, Soft Comput. 14 (9) (2010) 995–1010.
[42] D. Hernández-Lobato, V. Sharmanska, K. Kersting, C.H. Lampert, N. Quadrianto, Mind the nuisance: Gaussian process classification using privileged noise, in: Advances in Neural Information Processing Systems, 2014, pp. 837–845.
[43] Q. Yu, Y. Miche, E. Séverin, A. Lendasse, Bankruptcy prediction using extreme learning machine and financial expertise, Neurocomputing 128 (2014) 296–302.
[44] P. Du Jardin, Predicting bankruptcy using neural networks and other classification methods: the influence of variable selection techniques on model accuracy, Neurocomputing 73 (10) (2010) 2047–2060.
[45] B. Ribeiro, N. Lopes, Deep belief networks for financial prediction, in: B.-L. Lu, L. Zhang, J. Kwok (Eds.), Neural Information Processing, vol. 7064 of Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2011, pp. 766–773, ISBN 978-3-642-24964-8.
[46] B. Ribeiro, A. Vieira, J.C. das Neves, Sparse Bayesian models: bankruptcy-predictors of choice? in: IEEE International Joint Conference on Neural Networks, Vancouver, 2006, pp. 3377–3381.
[47] S.-C. Huang, Using Gaussian process based kernel classifiers for credit rating forecasting Expert Syst. Appl. 38 (7) (2011) 8607–8611.
[48] Z. Bitvai, T. Cohn, Predicting peer-to-peer loan rates using Bayesian non-linear regression, in: Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015, pp. 2203–2209.
[49] M. Lichman, UCI Machine Learning Repository, 2013 http://archive.ics.uci.edu/ml.
[50] MATLAB Release R2015a, 2015.
[51] H. Ooghe, C. Spaenjers, A note on performance measures for failure prediction models, Working Papers of Faculty of Economics and Business Administration, Ghent University, Belgium 06/405, Ghent University, Faculty of Economics and Business Administration, 2006.
[52] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (Jan) (2006) 1–30.
[53] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, J. Am. Stat. Assoc. 32 (200) (1937) 675–701.
[54] G. Cardillo, MYFRIEDMAN: Friedman Test for Non Parametric Two way ANalysis Of VAriance, 2009.
[55] W. Conover, Practical Nonparametric Statistics, 3rd ed., John Wiley and Sons (WIE), Kirjastus, 1998.
[56] X.-N. Wang, J.-M. Wei, H. Jin, G. Yu, H.-W. Zhang, Probabilistic confusion entropy for evaluating classifiers, Entropy 15 (11) (2013) 4969–4992.