Surveys in Operations Research and Management Science 21 (2016) 117–134

Contents lists available at ScienceDirect

Surveys in Operations Research and Management Science

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

Review

Classification methods applied to credit scoring: Systematic review and overall comparison

Francisco Louzada [a,*], Anderson Ara [a], Guilherme B. Fernandes [b]

[a] Department of Applied Mathematics & Statistics, University of São Paulo, São Carlos, Brazil

[b] P&D e Inovation in Analytics, Serasa-Experian, São Paulo, Brazil

ARTICLE INFO

Article history:
Received 9 April 2016
Received in revised form
27 June 2016
Accepted 5 October 2016

ABSTRACT

The need for controlling and effectively managing credit risk has led financial institutions to excel in improving techniques designed for this purpose, resulting in the development of various quantitative models by financial institutions and consulting companies. Hence, the growing number of academic studies about credit scoring shows a variety of classification methods applied to discriminate good and bad borrowers. This paper, therefore, aims to present a systematic literature review relating theory and application of binary classification techniques for credit scoring financial analysis. The general results show the use and importance of the main techniques for credit rating, as well as some of the scientific paradigm changes throughout the years.

Contents

1. Introduction 117
2. Survey methodology 118
   2.1. The main objective of the papers 119
   2.2. The main peculiarities of the credit scoring papers 119
3. The main classification methods in credit scoring 120
   3.1. Other issues related to credit scoring modeling 121
4. Results and discussion 123
   4.1. General results 123
   4.2. Results for different time periods 123
5. Is there a better method? a comparison study 129
6. Final comments 130
   Acknowledgments 132
   Appendix. 132
   References 132

1. Introduction

The need for credit analysis was born in the beginning of commerce in conjunction with the borrowing and lending of money, and the purchasing authorization to pay any debt in future. However, the modern concepts and ideas of credit scoring analysis emerged about 70 years ago with Durand [1]. Since then, traders have begun to gather information on the applicants for credit and catalog them to decide between to lend or not certain amount of money [2–4].

According to Thomas et al. [5] credit scoring is "a set of decision models and their underlying techniques that aid credit lenders in the granting of credit". A broader definition is considered in the present work: credit scoring is a numerical expression based on a level analysis of customer credit worthiness, a helpful tool for assessment and prevention of default risk, an important method in credit risk evaluation, and an active research area in financial risk management.

At the same time, the modern statistical and data mining techniques have given a significant contribution to the field of

* Corresponding author.
E-mail address: louzada@icmc.usp.br (F. Louzada).

http://dx.doi.org/10.1016/j.sorms.2016.10.001
1876-7354/

information science and are capable of building models to measure the risk level of a single customer conditioned to his characteristics, and then classify him as a good or a bad payer according to his risk level. Thus, the main idea of credit scoring models is to identify the features that influence the payment or the non-payment behavior of the customer as well as his default risk, occurring as the classification into two distinct groups characterized by the decision on the acceptance or rejection of the credit application [6].

Since the Basel Committee on Banking Supervision released the Basel Accords, specially the second accord from 2004, the use of credit scoring has grown considerably, not only for credit granting decisions but also for risk management purposes. The internal rating based approaches allow the institutions to use internal ratings to determine the risk parameters and therefore, to calculate the economic capital of a portfolio Basel III, released in 2013, render more accurate calculations of default risk, especially in the consideration of external rating agencies, which should have periodic, rigorous and formal comments that are independent of the business lines under review and that reevaluates its methodologies and models and any significant changes made to them [7,8].

Hence, the need for an effective risk management has meant that financial institutions began to seek a continuous improvement of the techniques used for credit analysis, a fact that resulted in the development and application of numerous quantitative models in this scenario. However, the chosen technique is often related to the subjectivity of the analyst or state of the art methods. There are also other properties that usually differ, such as the number of datasets applied to verify the quality of performance capability or even other validation and misclassification cost procedures. These are natural events, since credit scoring has been widely used in different fields, including propositions of new methods or comparisons between different techniques used for prediction purposes and classification.

A remarkable, large and essential literature review was presented in the paper by Hand and Henley [9], which discusses important issues of classification methods applied to credit scoring. Other literature reviews were also conducted but only focused on some types of classification methods and discussion of the methodologies, namely Xu et al. [10], Shi [11], Lahsasna et al. [12] and Nurlybayeva and Balakayeva [13]. Also, Garcia et al. [14] performed a systematic literature review, but limiting the study to papers published between 2000 and 2013, these authors provided a short experimental framework comparing only four credit scoring methods. Lessmann et al. [15] in their review considered 50 papers published between 2000 and 2014 and provided a comparison of several classification methods in credit scoring. However, it is known that there are several different methods that may be applied for binary classification and they may be encompassed by their general methodological nature and can be seen as modifications of others usual existing methods. For instance, linear discriminant analysis has the same general methodological nature of quadratic discriminant analysis. In this sense, even though Lessmann et al. [15] considered several classification methods they did not consider general methodologies as genetic and fuzzy methods.

In the most general point of view of operational research (OR) and management science (MS), regardless of the close relationship of both terms, as the use as synonyms or frequently used together OR/MS. The MS can be defined as the application of a scientific approach to solving management problems in order to help managers make better decisions [16]. On the other hand, OR can be interpreted as a mathematical and computer modeling emphasized in the approaches at the expense of systems thinking [17]. In this paper, a MS focus is considered broader, addressing the more general question of the application of the scientific method and knowledge to problems of management, while an OR focus is considered addressing mathematical decision solutions, as the studies whose objective is only to build a powerful model.

In this paper, therefore, we aim to present a more general systematic literature review over the application of binary classification techniques for credit scoring, which features a better understanding of the practical applications of credit rating and its changes over time. In the present literature review, we aim to cover more than 20 years of research (1992–2015) including 187 papers, more than any literature review already carried out so far, completely covering this partially documented period in different papers. Furthermore, we present a primary experimental simulation study under nine general methodologies, namely, neural networks, support vector machine, linear regression, decision trees, logistic regression, fuzzy logic, genetic programming, discriminant analysis and Bayesian networks, considering balanced and unbalanced databases based on three retail credit scoring datasets. We intend to summarize researching findings and obtain useful guidance for researchers interested in applying binary classification techniques for credit scoring.

The remainder of this paper is structured as follows. In Section 2 we present the conceptual classification scheme for the systematic literature review, displaying some important practical aspects of the credit scoring techniques. The main credit scoring techniques are briefly presented in Section 3. In Section 4 we present the results of the systematic review under the eligible reviewed papers, as well as the systematic review over four different time periods based on a historical economic context. In Section 5 we compare all presented methods on a replication based study. Final comments in Section 6 end the paper.

2. Survey methodology

A systematic review is an adequate alternative for identifying and classifying key scientific contributions to a field on a systematic, qualitative and quantitative description of the content in the literature. Interested readers can refer to Hachicha and Ghorbel [18] for more details on systematic literature review. It consists of an observational research method used to systematically evaluate the content of a recorded communication [19].

Overall, the procedure for conducting a systematic review is based on the definition of sources and procedures for the search of papers to be analyzed, as well as on the definition of instrumental categories for the classification of the selected papers, here based on four categories to understand the historical application of the credit scoring techniques: year of publication, title of the journal where the paper was published, name of the co-authors, and conceptual scheme based on 13 questions to be answered under each published paper. For this purpose, there is a need for defining the criteria to select credit scoring papers in the research scope. Thus, two selection criteria are used in this paper to select papers related to the credit scoring area to be included in the study:

- The study is limited to the published literature available on the following databases: ScienceDirect, Engineering Information, Reaxys and Scopus, covering 20,500 titles from 5000 publishers worldwide.
- The systematic review restricts the study eligibility to journal papers in English, especially considering ‘credit scoring’ as a keyword related to ‘machine learning’, ‘data mining’, ‘classification’ or ‘statistic’ topics. Other publication forms such as unpublished working papers, master and doctoral dissertations, books, conference proceedings, white papers and others are not included in the review. The survey horizon covers a period of almost two decades: from January 1992 to December 2015.

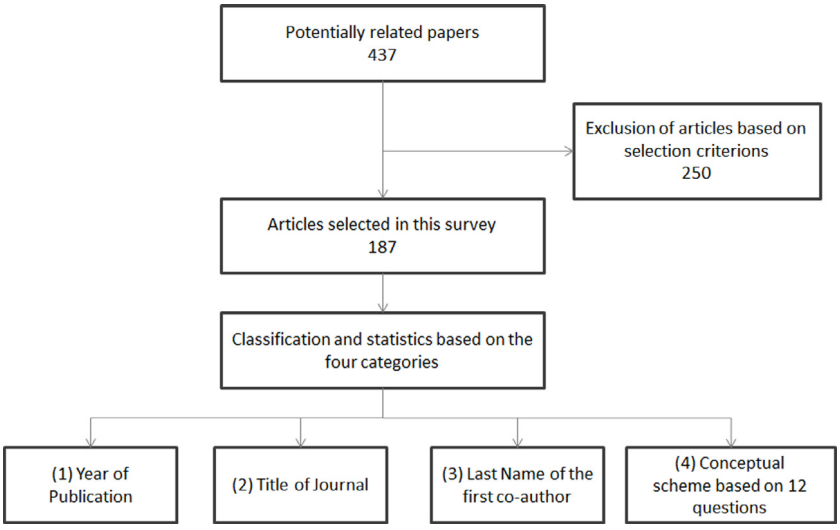

Fig. 1. Procedure of the systematic review.

The papers were selected according to the procedure shown in Fig. 1. From 437 papers eligible as potentially related to credit scoring, 250 were discarded due to not meeting the second selection criterion. The 187 papers included in the study were subjected to the systematic review, according to 12 questions on the conceptual scenario over the techniques: What is the main objective of the paper? What is the type of the main classification method? Which is the type of the datasets used? Which is the type of the explanatory variables? Does the paper perform variable selection methods? Was missing values imputation performed? What is the number of datasets used in the paper? Was exhaustive simulation study performed? What is the type of validation of the approach? What is the type of misclassification cost criterion? Does the paper use the Australian or the German datasets? Which is the principal classification method used in comparison study? Which is the principal focus of the paper concerning the decision area? The 13 questions are shown in Table A.1 in the Appendix.

2.1. The main objective of the papers

Although a series of papers are focused on the same area, they have different specific objectives. One can separate them in general similar aims. In the present work, we consider seven types of main objectives: proposing a new method for rating, comparing traditional techniques, conceptual discussions, feature selection, literature review, performance measures studies and, at last, other issues. Conceptual discussions account for papers that deal with problems or details of the credit rating analysis. In other issues, were included papers that presented low frequency objectives.

In the proposition of new methods, Lee et al. [20] introduce a discriminant neural model to perform credit rating, Gestel et al. [21] propose a support vector machine model within a Bayesian evidence framework. Hoffmann et al. [22] propose a boosted genetic fuzzy model, Hsieh and Hung [23] using a combined method that covers neural networks, support vector machine and Bayesian networks.

Shi [11] performed a systematic literature review that covers multiple criteria linear programming models applied to credit scoring from 1969 to 2010. Other literature reviews were performed by Hand and Henley [9]; Gemela [24]; Xu et al. [10]; Shi [11]; Lahsasna et al. [12]; Van Gool et al. [25].

Among the papers that perform a conceptual discussion, Bardos [26] presents tools used by the Banque de France, Banasik et al. [2] discuss how hazard models could be considered in order to investigate when the borrowers will default, Hand [27] discusses the applications and challenges in credit scoring analysis. Martens et al. [28] perform an application in credit scoring and discuss how their tool fits into a global Basel II credit risk management system. Other examples about conceptual discussion may be seen in [29,3,30].

In comparison of traditional techniques, West [31] compared five neural network model with traditional techniques. The results indicated that neural network can improve the credit scoring accuracy and also that logistic regression is a good alternative to the neural networks. Baesens et al. [32] performed a comparison involving discriminant analysis, logistic regression, logic programming, support vector machines, neural networks, Bayesian networks, decision trees and *k*-nearest neighbor. The authors concluded that many classification techniques yield performances which are quite competitive with each other. Other important comparisons may be seen in [33–52]. Also, Liu and Schumann [53]; Somol et al. [54]; Tsai [55]; Falangis and Glen [56]; Chen and Li [57]; Yu and Li [58]; McDonald et al. [59]; Wang et al. [60] handled features selection. Hand and Henley [9]; Gemela [24]; Xu et al. [10]; Shi [11]; Lahsasna et al. [12]; Van Gool et al. [25] produced their work in literature review. Yang et al. [61]; Hand [62]; Lan et al. [63]; Dryver and Sukkasem [64] worked on performance measures. There are other papers covering model selection [65], sample impact [66], interval credit [67], segmentation and accuracy [68].

2.2. The main peculiarities of the credit scoring papers

Overall the main classification methods in credit scoring are neural networks (NN) [69], support vector machine (SVM) [70], linear regression (LR) [71], decision trees (TREES) [72], logistic regression (LG) [73], fuzzy logic (FUZZY) [74], genetic programming [75], discriminant analysis (DA) [76], Bayesian networks (BN) [77], hybrid methods (HYBRID) [20], and ensemble methods (COMBINED), such as bagging [78], boosting [79], and stacking [80].

In comparison studies, the principal classification methods involve traditional techniques considered by the authors to contrast the predictive capability of their proposed methodologies. However, hybrid and ensemble methods are seldom used in comparison studies because they involve a combination of other traditional methods.

The main classification methods in credit scoring are briefly presented in Section 3 as well as other issues related to credit scoring modeling, such as, types of the datasets used in the papers (public or not public), the use of the so called Australian or

German datasets, type of the explanatory variables, feature selection methods, missing values imputation [81] number of datasets used, exhaustive simulations, validation approach, such as holdout sample, K-fold, leave one out, training/validation/test, misclassification cost criteria, such as Receiver Operating Characteristic (ROC) curve, metrics based on confusion matrix, accuracy (ACC), sensitivity (SEN), specificity (SPE), precision (PRE), false Positive Rate (FPR), and other traditional measures used in credit scoring analysis are F-Measure and two-sample K–S value.

3. The main classification methods in credit scoring

In this section, the main techniques used in credit scoring and their applications are briefly explained and discussed.

Neural networks (NN). A neural network [69] is a system based on input variables, also known as explanatory variables, combined by linear and non-linear interactions through one or more hidden layers, resulting in the output variables, also called response variables. Neural networks were created in an attempt to simulate the human brain, since it is based on sending electronic signals between a huge number of neurons. The NN structure has elements which receive an amount of stimuli (the input variables), creates synapses in several neurons (activation of neurons in hidden layers), and results in responses (output variables). Neural networks differ according to their basic structure. In general, they differ in the number of hidden layers and the activation functions applied to them. West [31] shows the mixture-of-experts and radial basis function neural network models must consider for credit scoring models. Lee et al. [20] proposed a two-stage hybrid modeling procedure to integrate the discriminant analysis approach with artificial neural networks technique. More recently, different artificial neural networks have been suggested to tackle the credit scoring problem: probabilistic neural network [82], partial logistic artificial neural network [83], artificial metaplasticity neural network [84] and hybrid neural networks [85]. In some datasets, the neural networks have the highest average correct classification rate when compared with other traditional techniques, such as discriminant analysis and logistic regression, taking into account the fact that results were very close [45]. Possible particular methods of neural networks are feedforward neural network, multi-layer perceptron, modular neural networks, radial basis function neural networks and self-organizing network.

Support vector machine (SVM). This technique is a statistical classification method and introduced by Vapnik [70]. Given a training set $\{(x_i, y_i)\}$, with $i = \{1, \ldots, n\}$, where x_i is the explanatory variable vector, and y_i represents the binary category of interest, and n denotes the number of dimensions of input vectors. SVM attempts to find an optimal hyper-plane, making it a non-probabilistic binary linear classifier. The optimal hyper-plane could be written as follows:

$$\sum_{i=1}^{n} w_i x_i + b = 0,$$

where $\mathbf{w} = w_1, w_1, \ldots, w_n$ is the normal of the hyper-plane, and b is a scalar threshold. Considering the hyper-plane separable with respect to $y_i \in \{-1, 1\}$ and with geometric distance $\frac{2}{\|\mathbf{w}\|^2}$, the procedure maximizes this distance, subject to the constraint $y_i \left(\sum_{i=1}^{n} w_i x_i + b\right) \geq 1$. Commonly, this maximization may be done through the Lagrange multipliers and using linear, polynomial, Gaussian or sigmoidal separations. Just recently support vector machine was considered a credit scoring model [86]. Li et al. [87]; Gestel et al. [21]; Xiao and Fei [88]; Yang [89]; Chuang and Lin [90]; Zhou et al. [91,92]; Feng et al. [93]; Hens and Tiwari [94]; Ling et al. [95] used support vector machine as main technique for their new method. Possible particular methods of SVM are radial basis function least squares support vector machine, linear least-squares support vector machine, radial basis function, support vector machine and linear support vector machine.

Linear regression (LR). The linear regression analysis has been used in credit scoring applications even though the response variable is a two class problem. The technique sets a linear relationship between the characteristics of borrowers $X = \{X1, \ldots, X_p\}$ and the target variable Y, as follows,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

where ϵ is the random error and independent of X. Ordinary least squares is the traditional procedure to estimate $\beta = \beta_0, \ldots, \beta_p$, being $\hat{\beta}$ the estimated vector. Once Y is a binary variable, the conditional expectation $E(Y|X) = x'\beta$ may be used to segregate good borrowers and bad borrowers. Since $-\infty < x'\beta < \infty$, the output of the model cannot be interpreted as a probability. Hand and Kelly [71] built a superscorecard model based on linear regression. Karlis and Rahmouni [96] propose the Poisson mixture models for analyzing the credit-scoring behavior for individual loans. Other authors have been working with linear regression models or its generalizations in credit scoring [71,97,96,98].

Decision trees (TREES). Classification and Regression Trees [72] is a classification method which uses historical data to construct so-called decision rules organized into tree-like architectures. In general, the purpose of this method is to determine a set of if-then logical conditions that permit prediction or classification of cases. There are three usual tree's algorithms: chi-square automatic interaction detector (CHAID), classification and regression tree (CART) and C5, which differ by the criterion of tree construction, CART uses gini as the splitting criterion, C5 uses entropy, while CHAID uses the chi-square test [99]. John et al. [100] exhibit a rule based model implementation in a stock selection. Bijak and Thomas [68] used CHAID and CART to verify the segmentation value in the performance capability. Kao et al. [101] propose a combination of a Bayesian behavior scoring model and a CART-based credit scoring model. Other possible and particular methods of decision trees are C4.5 decision trees algorithm and J4.8 decision trees algorithm.

Logistic regression (LG). Proposed by Berkson [73], the logit model considers a group of explanatory variables $X = \{X_1, \ldots, X_p\}$ and a response variable with two categories $Y = \{y_1, y_2\}$, the technique of logistic regression consists in the estimation of a linear combination between X and the logit transformation of Y. Thus, if we consider y_1 as the category of interest for analysis, the model can be represented as $\log\left(\frac{\pi}{1-\pi}\right) = X\beta$, where $\pi = P(Y = y_1)$ and β is the vector containing the model's coefficients. Alternatively, the model can be represented by,

$$\pi_i = \frac{\exp\{X_i\beta\}}{1 - \exp\{X_i\beta\}}, \tag{1}$$

where π_i is the probability of the ith individual to belong to category y_1, conditioned to X_i. The logistic regression model is a traditional method, often compared with other techniques [102,62,103,45,99,104] or it is used in technique combinations [105]. Other possible and particular methods of logistic regression are regularized logistic regression and limited logistic regression.

Fuzzy logic (FUZZY). Zadeh [74] introduced the Fuzzy Logic as a mathematical system which deals with modeling imprecise information in the form of linguistic terms, providing an approximate answer to a matter based on knowledge that is inaccurate, incomplete or not completely reliable. Unlike the binary logic, fuzzy logic uses the notion of membership to handle the imprecise information. A fuzzy set is uniquely determined by its membership function, which can be triangular, trapezoidal, Gaussian, polynomial or sigmoidal function. Hoffmann et al. [34] performed

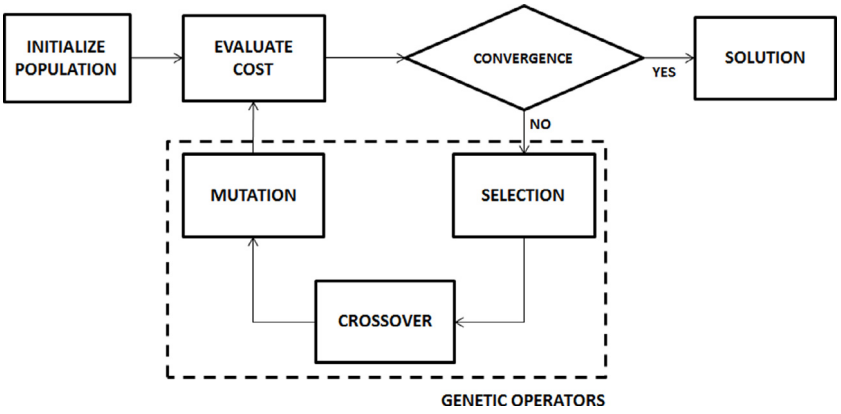

Fig. 2. Flowchart of a genetic algorithm.
Source: Adapted from Abdoun and Abouchabaka [111].

an evaluation of two fuzzy classifiers for credit scoring. Laha [106] proposes a method of building credit scoring models using fuzzy rule based classifiers. Lahsasna et al. [107] investigated the usage of Takagi–Sugeno (TS) and Mamdani fuzzy models in credit scoring. Possible methods in fuzzy logic are regularized adaptive network based Fuzzy inference systems and fuzzy Adaptive Resonance.

Genetic programming (GENETIC). Genetic Programming [75] is based on mathematical global optimization as adaptive heuristic search algorithms, its formulation is inspired by mechanisms of natural selection and genetics. Basically, the main goal of a genetic algorithm is to create a population of possible answers to the problem and then submit it to the process of evolution, applying genetic operations such as crossover, mutation and reproduction. The crossover is responsible for exchanging bit strings to generate new observations. Fig. 2 shows the optimization process of a genetic algorithm. Ong et al. [35] propose a genetic credit scoring model and compares this with traditional techniques. Huang et al. [108] introduce a two-stage genetic programming. Many other authors have investigated genetic models in application of credit scoring [29,109,49,110]. Other possible methods in genetic programming are the two stages genetic programming and genetic algorithm knowledge refinement.

Discriminant analysis (DA). Introduced by Fisher [76], the discriminant analysis is based on the construction of one or more linear functions involving the explanatory variables. Consequently, the general model is given by

$$Z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p,$$

where Z represents the discrimination score, α the intercept, β_i represents the coefficient responsible for the linear contribution of the ith explanatory variable X_i, where $i = 1, 2, \ldots, p$.

This technique has the following assumptions: (1) the covariance matrices of each classification subset are equal. (2) Each classification group follows a multivariate normal distribution. Frequently, the linear discriminant analysis is compared with other credit scoring techniques [31,21,112] or is subject of studies of new procedures to improve its accuracy [89,56]. Another possible method in discriminant analysis is quadratic discriminant analysis.

Bayesian networks (BN). A Bayesian classifier [77] is based on calculating a posterior probability of each observation belongs to a specific class. In other words, it finds the posterior probability distribution $P(Y|X)$, where $Y = (y_1, y_2, \ldots, y_k)$ is a random variable to be classified featuring k categories, and $X = (X_1, X_2, \ldots, X_p)$ is a set of p explanatory variables. A Bayesian classifier may be seen as a Bayesian network (BNs): a directed acyclic graph (DAG) represented by the triplet $(\mathbb{V}, \mathbb{E}, \mathbb{P})$, where $\mathbb{V}$ are the nodes, $\mathbb{E}$ are the edges and $\mathbb{P}$ is a set of probability distributions and its parameters. In this case, the nodes represent the domain variables and edges the relations between these variables. Giudici [113] presents a conditional Bayesian independence graph to extract insightful information on the variables association structure in credit scoring applications. Gemela [24] applied Bayesian networks in a credit database of annual reports of Czech engineering enterprises. Other authors who have investigated Bayesian nets in credit scoring models are Zhu et al. [114]; Antonakis and Sfakianakis [115]; Wu [116]. Possible methods in Bayesian networks are naive Bayes, tree augmented naive Bayes and Gaussian naive Bayes.

Hybrid methods (HYBRID). Hybrid methods combine different techniques to improve the performance capability. In general, this combination can be accomplished in several ways during the credit scoring process. Lee et al. [20] proposed a hybrid method that integrates the backpropagation neural networks with traditional discriminant analysis to evaluate credit scoring. Huang et al. [117] proposed a hybrid method that integrates genetic algorithm and support vector machine to perform feature selection and model parameters optimization simultaneously, as well as Lee et al. [20]; Lee and Chen [103]; Hsieh [118]; Huysmans et al. [119]; Shi [120]; Chen et al. [86]; Liu et al. [121]; Ping and Yongheng [122]; Capotorti and Barbanera [123]; Vukovic et al. [124]; Akkoc [112]; Pavlidis et al. [104] also work with hybrid methods. also work with hybrid methods.

Ensemble methods (COMBINED). The ensemble procedure refers to methods of combining classifiers, thereby multiple techniques are applied to solve the same problem in order to boost credit scoring performance. There are three popular ensemble methods: bagging [78], boosting [79], and stacking [80]. The Hybrid methods can be regarded as a particular case of stacking, but in this paper we consider as stacking only the methods which use this terminology. Wang et al. [125] proposed a combined bagging decision tree to reduce the influences of the noise data and the redundant attributes of data. Many other authors have chosen to deal with combined methods Hoffmann et al. [22]; Hsieh and Hung [23]; Paleologo et al. [126]; Zhang et al. [127]; Finlay [128]; Louzada et al. [105]; Xiao et al. [129]; Marques et al. [130] in credit scoring problems.

3.1. Other issues related to credit scoring modeling

Types of the datasets used. As much as nowadays the information is considered easy to access, mainly because of the modernization of the web and large data storage centers, the availability of data on the credit history of customers and businesses is still difficult to access. Datasets which contain confidential information on applicants cannot be released to third parties without careful

safeguards [27]. Not rarely, public datasets are used for the investigation of techniques and methodologies of credit rating. In this sense, the type of dataset used (public or not public) in the papers is an important issue.

Type of the explanatory variables. The explanatory variables, often known as covariates, predictor attributes, features, predictor variables or independent variables, usually guide the choice and use of a classification method. In general, the type of each explanatory variable may be continuous (interval or ratio) or categorical (nominal, dichotomous or ordinal). A common practice is to discretize a continuous attribute as done by Gemela [24]; Mues et al. [109]; Ong et al. [35]; Wu [116]. In this paper, we consider a continuous dataset to be the one that contains only interval or ratio explanatory variables—independent if a discretization method is applied or not, and a categorical dataset presents only categorical explanatory variables. A mixed dataset is composed of both types of variables.

Feature selection methods. When we use data to try to provide a credit rating, we use the number of variables that, in short, explain and predict the credit risk. Some methods provide a more accurate classification, discarding irrelevant features. Thus, it is a common practice to apply such methods when one proposes a rating model. Some authors used a variable selection procedure in their papers such as Lee et al. [20]; Verstraeten and Van Den Poel [66]; Abdou et al. [45]; Chen and Li [57]; Marques et al. [130]. Authors, who did not cite or discuss feature selection methods in their papers, were regarded as nonusers.

Missing values imputation. The presence of missing values in datasets is a recurrent statistical problem in several application areas. In credit analysis, internal records may be incomplete for many reasons: a registration poorly conducted, the customers can fail to answer to questions, or the database or recording mechanisms can malfunction. One possible approach is to drop the missing values from the original dataset, as done by Adams et al. [33]; Berger et al. [131]; Won et al. [110] or perform a preprocessing to replace the missing values, as done by Banasik et al. [97]; Baesens et al. [36]; Paleologo et al. [126]. These procedures are known as missing data imputation [81].

Number of datasets used. In general, authors must prove the efficiency of their rating methods on either real or simulated datasets. However, due to the difficulty of obtaining data in the credit area, many times the number of datasets used can be small, or even engage the use of other real datasets that prove the efficiency of the rating method. Lan et al. [63] used 16 popular datasets in the experiments that testify performance measures and which were applied in a credit card application.

Exhaustive simulations. The exhaustive simulations study is based on Monte Carlo sample replications and the statistical comparisons to assess the performance of the estimation procedure. In this sense, artificial samples with specific properties are randomly generated. Ziariet al. [65]; Hardle et al. [132]; Banasik et al. [97]; Louzada et al. [4,133] are some examples of authors who performed exhaustive simulations in credit scoring analysis.

Validation approach. Amongst the various validation procedures we point out:

Holdout sample. This validation method involves a random partition of the dataset into two subsets: the first, called training set is designed to be used in the model estimation phase. The second, called test set, is used to perform the evaluation of the respective model. Therefore, the model is fitted based on the training set aimed to predict the cases in the test set. A good performance in the second dataset indicates that the model is able to generalize the data, in other words, there is no overfitting on the training set.

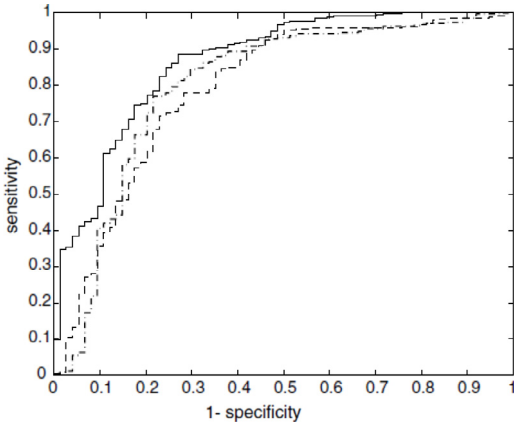

Fig. 3. The Receiver operating characteristic curves used by Gestel et al. [21] to compare support vector machine (solid line), logistic regression (dashed–dotted line) and linear discriminant analysis (dashed line).

K-fold. This method is a generalization of the hold out method, meanwhile the dataset is randomly partitioned into K subsets. Each subset is used as a test set for the model fit considering the other $K - 1$ subsets as training sets. In this approach, the entire dataset is used for both training and testing the model. Typically, a value of $K = 10$ is used in the literature [134].

Leave one out. This method is an instance of K-fold where K is equal to the size of the dataset. Each case is used as a test set for the model fit considering the other cases as training sets. In this approach, the entire dataset is used for both training and testing models. It is worth to mention that on large datasets a computational difficulty may arise.

Train/validation/test. This validation approach is an alternative of the holdout case for large datasets, the purpose is to avoid some overfitting into the validation set. The training samples are used to develop models, the validation samples are used to estimate the prediction error for the model selection, the test set is used to evaluate the generalization error of the final model chosen. For this, the performance of the selected model should be confirmed through the measuring of the performance on a third independent dataset denominated test set [135]. A common split is 50% for training, 25% each for validation and test.

Misclassification cost criteria. Amongst the various misclassification criteria we point out:

ROC curve. The Receiver Operating Characteristic curve was introduced by Zweig and Campbell [136] and may be geometrically defined as a graph for visualizing the performance of a binary classifier technique. The ROC curve is obtained by measuring the 1-specificity on the first axis and measured the sensitivity on the second axis, creating a plane. Therefore, the more the curve distances from the main diagonal, the better is the model performance. Fig. 3 shows an example of ROC Curve.

Metrics based on confusion matrix. Its aim is to compare the model's predictive outcome with the true response values in the dataset. A misclassification takes place when the modeling procedure fails to correctly allocate an individual into the correct category. A traditional procedure is to build a confusion matrix, as shown in Table 1, where M is the model prediction, D is the real value in dataset, TP the number of true positives, FP the number of false positives, FN the number of false negatives and TN the number of true negatives. Naturally, $TP + FP + FN + TN = N$, where N is the number of observations. Through the confusion matrix, some measures are employed to evaluate the performance on test samples.

Table 1
Confusion matrix.

		M	
		{1}	{0}
D	{1}	TP	FP
	{0}	FN	TN

Accuracy (ACC): the ratio of correct predictions of a model, when classifying cases into class {1} or {0}. *ACC* is defined as $ACC = (TP + TN)/(TP + TN + FN + FP)$.
Sensitivity (SEN): also known as *Recall* or *True Positive Rate* is the fraction of the cases that the technique correctly classified to the class {1} among all cases belonging to the class {1}. *SEN* is defined as $SEN = TP/(TP + FN)$.
Specificity (SPE): also known as *True Negative Rate* is the ratio of observations correctly classified by the model into the class {0} among all cases belonging to the class {0}. *SPE* is defined as $SPE = TN/(TN + FP)$.
Precision (PRE): is the fraction obtained as the number of true positives divided by the total number of instances labeled as positive. It is measured as $PRE = \frac{TN}{/}(TN + FP)$. *False Negative Rate (FNR)* also known as *Type I Error* is the fraction of {0} cases misclassified as belonging to the {1} class. It is measured as $FNR = FN/(TP + FN)$.
False Positive Rate (FPR) also known as *Type II Error* is the fraction of {1} cases misclassified as belonging to the {0} class. It is measured as $FPR = FP/(TN + FP)$. Other traditional measures used in credit scoring analysis are F-Measure and two-sample K–S value. The F-Measure combines both Precision and Recall, while the K–S value is used to measure the maximum distance between the distribution functions of the scores of the 'good payers' and 'bad payers'.

Using the Australian and German dataset. The Australian and German datasets are two public UCI [137] datasets concerning approved or rejected credit card applications. The first has 690 cases, with 6 continuous explanatory variables and 8 categorical explanatory variables. The second has 1000 instances, with 7 continuous explanatory, 13 categorical attributes. All the explanatory variables' names and values are encrypted by symbols. The use of these benchmark datasets is frequent in credit rating papers and the comparison of the overall classification performance in both cases is a common practice for the solidification of a proposed method. Ling et al. [95] show an accuracy comparison of different authors and techniques for Austrian and German datasets.

Principal methods for comparison. The principal classification methods in comparison studies involve traditional techniques considered by the authors to contrast the predictive capability of their proposed methodologies. However, hybrid and ensemble techniques are rarely used in comparison studies because they involve a combination of other traditional methods.

Principal focus of the paper concerning the decision area. The principal focus of the paper concerning the decision area refers if the paper has an OR focus or a MS focus, is based on only to build a powerful model for the first category and it comprises studies those whose objective is to develop a model to solving management problems in order to help managers make better decisions for the second category. For instance, considering the MS focus, DeYoung et al. [138] used a data source for small business loans and showed that annual increases in borrower–lender distances were slow and steady prior to 1993 but accelerated rapidly after that.

4. Results and discussion

In this section we present the general results of the reviewed papers. We discuss the classification of papers according to the year of publication, scientific journal, author and conceptual scenery. Moreover, we present a more detailed analysis and discussion of the systematic review for each time period, I, II, III and IV.

4.1. General results

On the classification of papers according to the year of publication. As indicated in Fig. 4, the number of papers published in each year from January 1992 to December 2015 ranges from 0 to 25 papers, with a evident growth in all the ranges and a fast increment after 2000, with an average of 7.8 and standard deviation of 7.6 papers by year.

In order to input in the analysis the historical occurrence we divide the studied period of time in four parts. The historical economic context expressed by Basel I, II and III encounters – 1988, 2004 and 2013, respectively – may have had an increase in the number of papers with possible time lag in reviewing and revising the submitted manuscripts. Thus, we consider the following four time period sceneries. The first scenery is obtained by considering papers published before 2006 (Year $\leq$ 2006), hereafter 'I'; the second scenery is obtained by the papers published between 2006 and 2010 (2006 $<$ Year $\leq$ 2010), hereafter 'II'; the third scenery is obtained by the papers published between 2010 and 2010 (2010 $<$ Year $\leq$ 2012), hereafter 'III'; and the last time scenery is for papers published after 2012 (Year $>$ 2012) referred to 'IV'. The respective number of papers in each time period scenery equals 45, 51, 39 and 54 papers.

On the classification of papers according to the scientific journal. The reviewed papers were published by 73 journals and the frequencies are shown in Table 2. Most of these papers are related to scientific journals of computer science, decision sciences, engineering and mathematics. As shown in Table 2, the largest number of papers were published by 'Expert Systems with Applications' and 'Journal of the Operational Research Society' which account for 27.81% and 10.70% of the 187 reviewed papers, respectively. In the four time periods, the journal 'Expert Systems with Applications' exhibits moderately an increasing number of papers in credit scoring, while the 'Journal of the Operational Research Society' exhibits a decreasing number of papers in the same context. 'Knowledge-Based Systems' amounts to an exponential increase of these papers.

On the classification of papers according to the authors. In the 187 reviewed papers, there are 525 different co-authors. The frequency of appearance of those is presented in Table 3, where only co-authors with over 4 appearances are shown. Baesens B., Vanthienen J., Hand D.J. and Thomas L.C. are the researchers who published the largest number of papers, which represented 3.0%, 1.9%, 1.5% and 1.5%, respectively. As may be seen in Table 3 these researchers are mostly from Belgium, United Kingdom, Taiwan, US, Chile and Brazil.

On the classification of papers according to conceptual scene. The twelve questions applied in the systematic review for all 187 reviewed papers are shown in the Table A.1 of Appendix. In the next section, the analysis and discussion of these results are performed, they allow us to understand the methodological progress that occurred in credit scoring analysis on the past two decades.

4.2. Results for different time periods

On the main objectives in credit scoring analysis. As shown in Fig. 5 the most common goal of the papers is the proposition of new methods in credit scoring, representing 51.3% of all 187 reviewed papers. This preference is maintained for the four time periods. Fig. 6 shows the frequencies of general techniques used as new methods in credit scoring. The hybrid is the most common method with almost 20%, followed by combined methods with almost

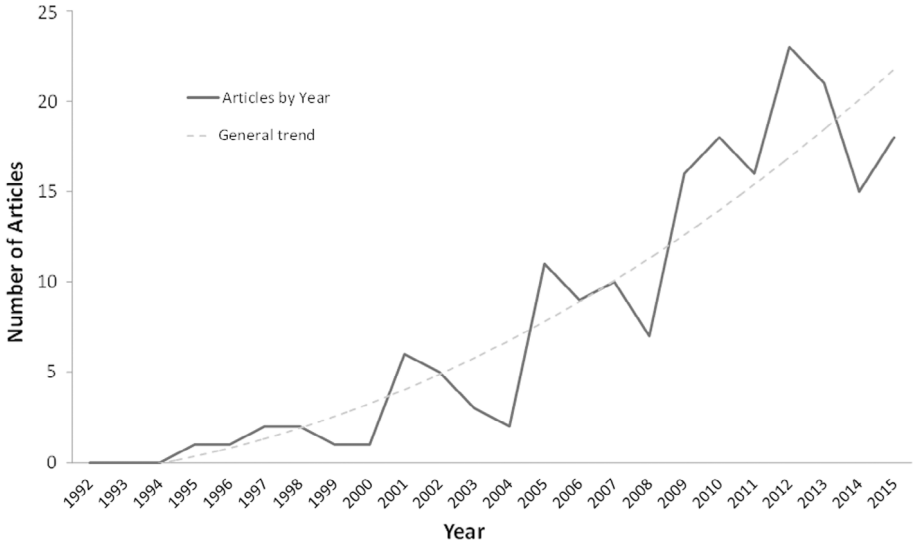

Fig. 4. Number of credit scoring papers published by year.

Table 2
Distribution of reviewed papers according to the journal title in the four time periods.

Journal	I	II	III	IV	Total	%
Expert Systems with Applications	9	16	14	13	52	27.81
Journal of the Operational Research Society	11	5	1	3	20	10.70
European Journal of Operational Research	1	6	3	6	16	8.56
Knowledge-Based Systems	0	1	2	4	7	3.74
Applied Stochastic Models in Business and Industry	4	0	0	0	4	2.14
Computational Statistics and Data Analysis	1	0	1	1	3	1.60
IMA Journal Management Mathematics	2	1	0	0	3	1.60
International Journal of Neural Systems	0	0	3	0	3	1.60
Others[a]	15	22	15	27	79	42.25
Total	43	51	39	54	187	100

[a] These include papers from ACM Trans. on Knowledge Discovery from Data, Decision Support Systems, Journal of the Royal Stat. Society, Inter. Journal of Comp. Intelligence & Applications, Applied Math. & Comp., Applied Soft Computing, Comm. in Statistics, Comp. Statistics, Credit and Banking and others.

Table 3
Distribution of reviewed papers according to the author/co-author in the four time periods.

Author	Affiliation, Country	I	II	III	IV	Total	%
Baesens, B.	Katholieke Univ. Leuven, Belgium	7	5	2	2	16	3.0
Vanthienen, J.	Katholieke Univ. Leuven, Belgium	6	4	0	0	10	1.9
Hand, D.J.	Imperial College London, UK	7	0	1	0	8	1.5
Thomas, L.C.	University of Southampton, UK	1	2	2	3	8	1.5
Mues, C.	University of Southampton, UK	1	2	3	1	7	1.3
Van Gestel, T.	Katholieke Univ. Leuven, Belgium	3	4	0	0	7	1.3
Tsai, C.-F.	Nat. Chung Cheng University, Taiwan	0	3	0	3	6	1.1
Bravo, C.	Universidade de Chile, Chile	0	0	0	4	4	0.8
Louzada, F.	Universidade de Sao Paulo, Brazil	0	0	3	1	4	0.8
Shi, Y.	University of Nebraska Omaha, US	0	3	0	1	4	0.8
Others		95	107	101	148	451	85.9
Total		120	130	112	163	525	100.0

15% and support vector machine along with neural networks with around 13%. Due to the sheer number of methods involved and different kinds of behavior in each dataset, the second most popular main objective is the comparison of traditional techniques. However, it is becoming less common in the latest years (III and IV). The third most usual main objective is the conceptual discussion, which is most common in IV time period. Other main objectives do not reach 10% of the total of papers reviewed. The performance measures studies are more common in past years (I time period). Also, there is stability in the four time periods of literature review and other issues.

The research evolution of a new field, such as credit scoring, starts with the discovery that it is poorly investigated by researchers. Moreover, the academic and professional interest in a particular research area is usually boosted by new environmental changes. In the case of credit scoring, the main environmental changes are the rapid increase of storage information and the processing capacity, combined with the creation of the Basel accords, which means a change in *why and how* to control credit risk. The conceptual discussion sets definitions, ideas and problems to be faced. The increasing number of researchers interested in credit scoring culminated in the development and adaptation of techniques for tackling the main questions. After the techniques were developed, methods for comparing those are proposed. At last, a field of research will eventually reach a state-of-the-art phase, followed by new researchers questioning the paradigm,

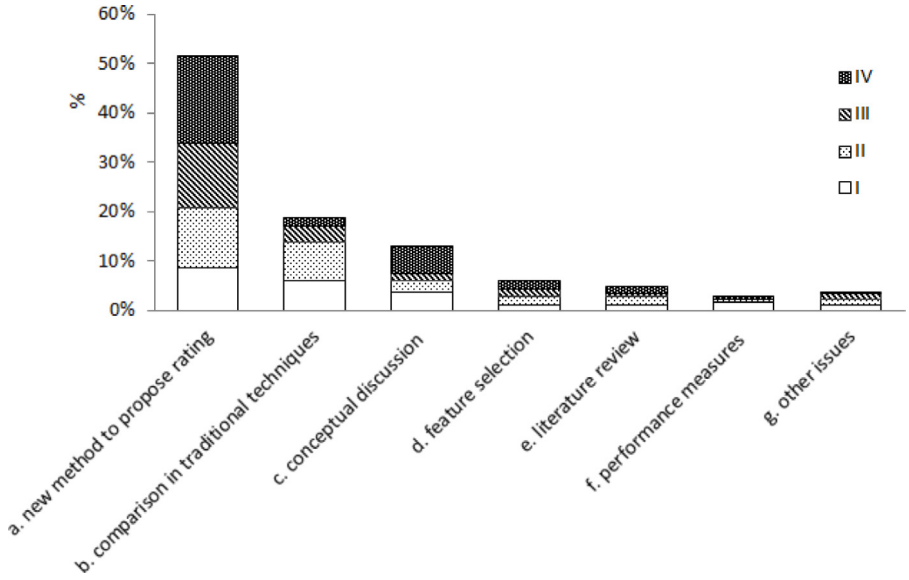

Fig. 5. Main objectives of the credit scoring analysis.

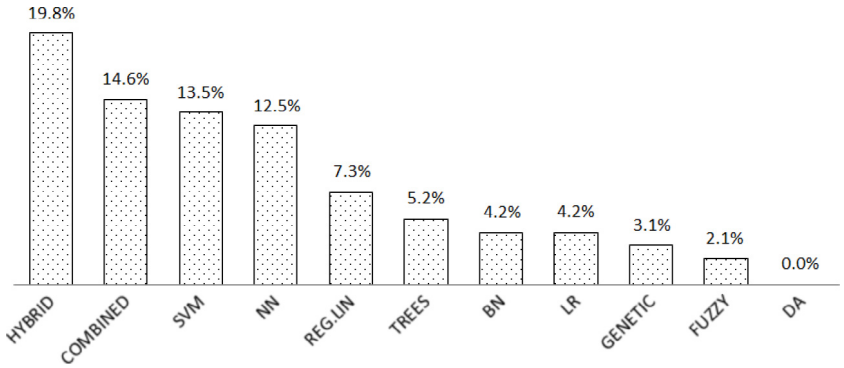

Fig. 6. The principal techniques in proposition of new methods in credit scoring.

ideas and disrupting the status quo of the credit scoring area. Currently, credit scoring is going through the process of tools development, as shown in Fig. 4.

On the main classification techniques. As a classification technique is applied as a credit scoring model, the choice of technique is often related to the subjectivity of the analyst or to state-of-the-art methods. Ideally, a precise prediction indicates whether a credit extended to an applicant will probably result in profit for the lending institution. Fig. 7 shows the circular bar plots concerning the main classification techniques applied in all considered periods as well as their utilization over time, this figure only considers the techniques indicated in Section 2.2. In general, the neural networks and support vector machine are the most common used techniques in credit scoring (17.6%), the discriminant analysis remained as a rarely used technique (1.7%). In the first time period analyzed, the most common technique is the neural network (20.6%).

However, neural networks and hybrid methods remained at constant use in all following periods considered with a higher frequency. Support vector machine was most used between 2006 and 2010 II time period (21.4%), this method is the fourth most commonly used in general, although with a fast increasing in past and decreasing its participation over recent years. The trees, Bayesian net, linear regression and logistic regression techniques had this same percentage in this period. However, logistic regression was most used (15.2%) in recent years and matching the use of neural networks in IV time period. In addition, there is a strong decrease in the use of the genetic, fuzzy and discriminant analysis methods and a remarkable growth of combined techniques which are the most used methods in recent years, IV time period, with 21.2%. Hybrid methods have always been highly used, but were not the highlights in any time period. In comparison with Fig. 6, the hybrid and combined methods are mostly used in new methods to propose rating in credit scoring, followed by support vector machine and neural networks.

On the datasets used in credit scoring. Fig. 8(a) and (b) show information about the datasets used in credit scoring reviewed papers. As indicated by Fig. 8(a), the most common type of datasets is private in all time periods, followed by public dataset and lastly the use of both types. In other words, the authors usually employ only private datasets in their credit scoring applications. This fact seems to be independent of the time period. As indicated by Fig. 8(b), authors prefer to use datasets that have continuous and discrete variables. However, in I the datasets with only discrete variables were more common than those with only continuous variables. Discarding Lan et al. [63], which used 16 datasets in their work, Table 4 shows the basic statistics of the number of datasets used in reviewed papers. In general, the papers consider an average of 2.18 datasets in their content. Fig. 9 shows the behavior of the number of datasets in the four times periods, and indicates a growth in the number of datasets used in the periods I, II and III and an average decrease in IV with a growth in the standard deviation.

On the preprocessing data methods in credit scoring. In regard to preprocessing methods in credit scoring, this review covers

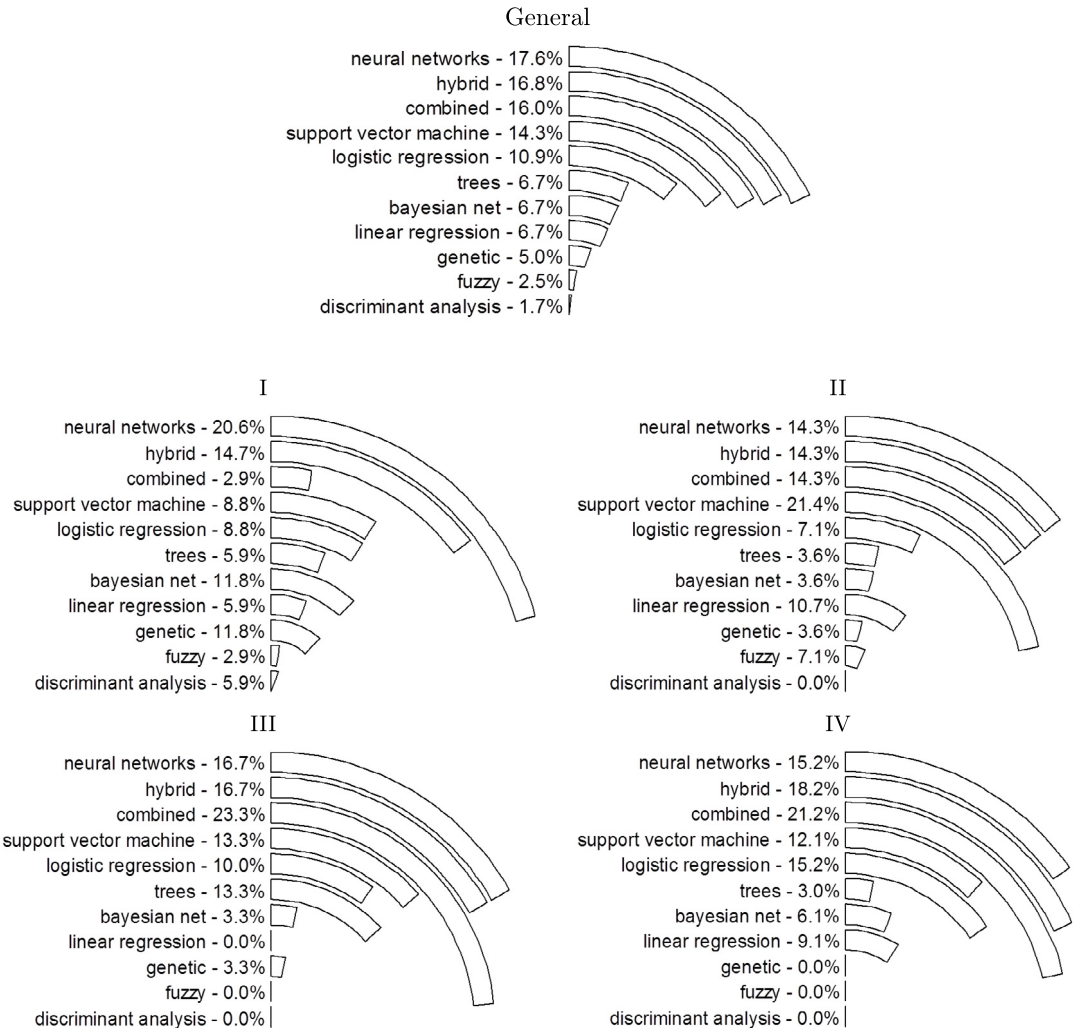

Fig. 7. The main classification techniques in credit scoring.

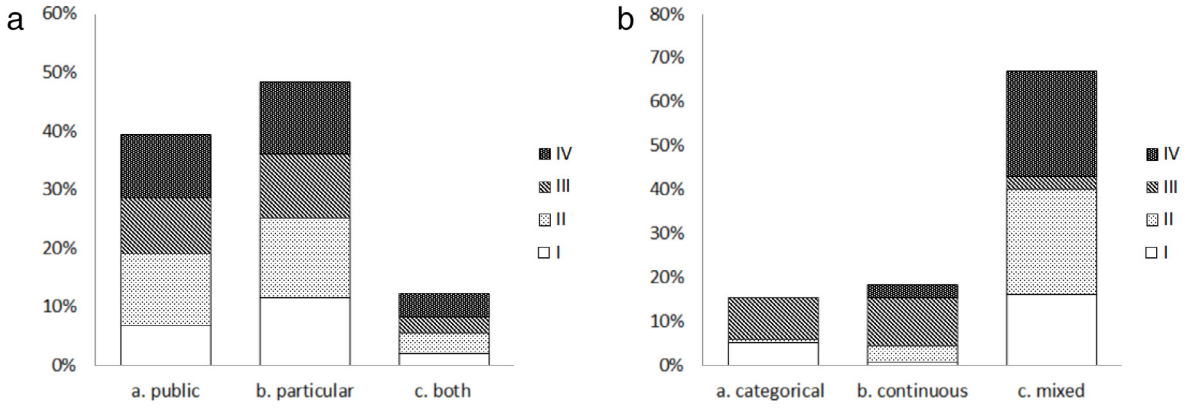

Fig. 8. (a) The type of used datasets and (b) the type of variables in datasets.

Table 4
Statistical summary of the number of used datasets.

Time period	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Sdv
I	1.00	1.00	1.00	1.80	2.00	8.00	1.69
II	1.00	1.00	2.00	2.05	2.00	7.00	1.40
III	1.00	1.00	2.00	2.55	3.00	8.00	1.85
IV	1.00	1.00	1.00	2.31	3.00	10.00	2.32
General	1.00	1.00	1.00	2.18	3.00	10.00	1.84

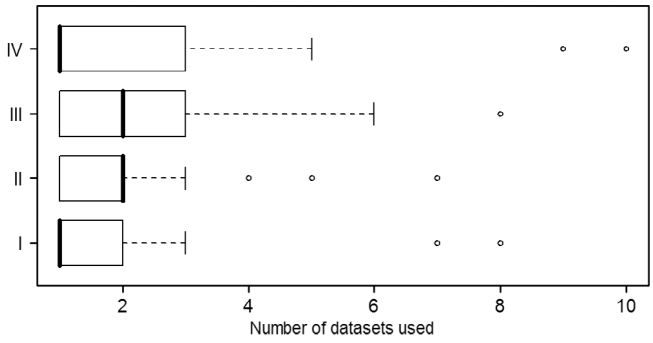

Fig. 9. The behavior of the number of dataset used in credit scoring studies.

two relevant aspects: the feature (variable) selection and missing data procedures. Fig. 10(a) shows that, independently of the time period, the feature selection is performed in most studies. However, in about 49% of the papers this procedure is not used. Fig. 10(b) shows that the missing data imputation is a procedure often not used in credit scoring analysis (90%).

On the validation of the approaches. The validation of the approaches is a part of the procedures that ensure the authors of the examination of the performance and comparability of their methods. In general, as indicated by Fig. 11(a), more than 80% of the papers do not consider exhaustive simulations in their procedures. Likewise, as indicated in Fig. 11(b), almost 45% of all reviewed papers consider the Australian or German credit dataset, and during the II time period it became an even more common practice. Table 5 shows the overall classification performance on Australian and German credit datasets for 30 reviewed papers. Concerning the splitting of the datasets, Fig. 12 shows that K-Fold cross validation and holdout methods were more common in general, and in more recent time periods, the K-fold cross validation became the most widely used method. The splitting of the dataset in three parts (train/validation/test) is more used than the leave-one-out procedure.

On the misclassification cost criterion. Fig. 13 shows that to measure the misclassification cost, the most common criteria used in the reviewed papers are the metrics based on confusion matrix (45%). Although this criterion was not in use solely in the I time period, it was widely used in others. The utilization of the ROC Curve was more common in the past period and about 10% of all the reviewed papers used both or other criteria.

On the classification methods used in comparison studies. Regarding the traditional techniques used in comparison studies, Fig. 15 shows the circular bar plots concerning techniques applied in all considered periods. The most used technique in comparison studies is logistic regression (23.14%) which has always had a high frequency of use in all considered periods. The neural networks is the second most used technique (21.0%) with a high usage in II time period. The support vector machine was widely and recently used in comparison studies, but in general it is the fourth most frequently used technique (14.8%). The trees remained as the third most used technique in all periods. In the reviewed papers, no study performs comparisons using combined techniques.

On the principal focus of the paper concerning the decision area. Regarding the principal focus concerning the decision area, Fig. 14

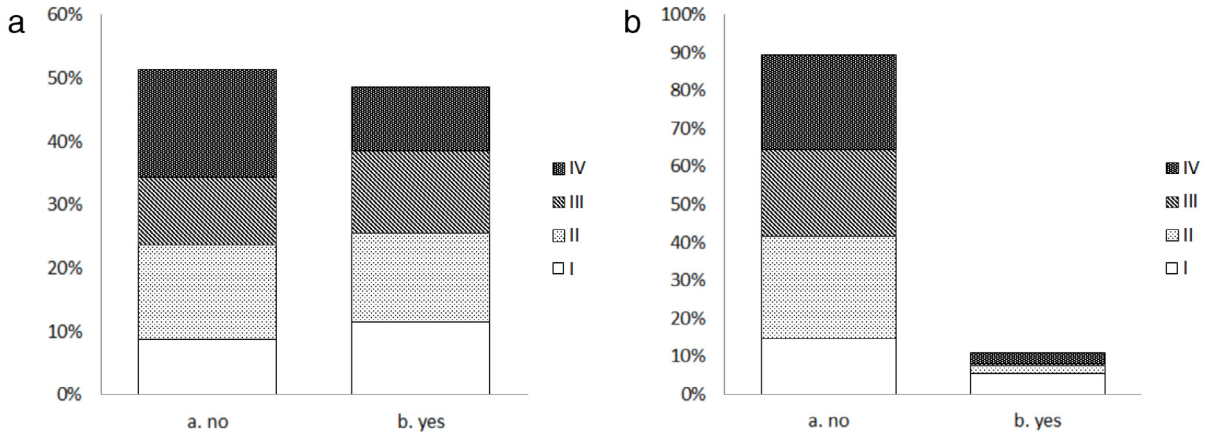

Fig. 10. (a) The using of feature selection and (b) the using of missing data imputation.

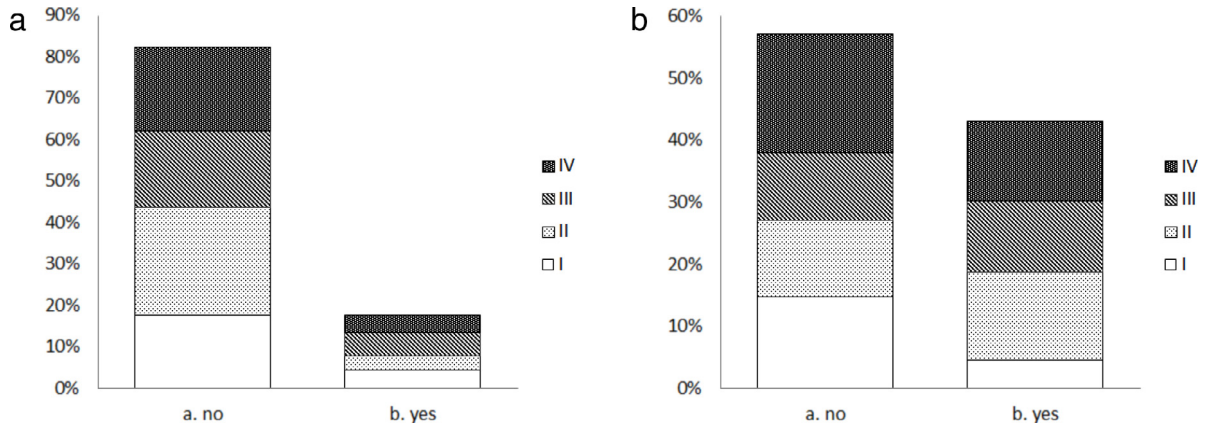

Fig. 11. (a) The using of exhaustive simulations and (b) the using of the Australian or German credit dataset.

Table 5
Overall classification performance on Australian and German credit datasets.

Paper	AUS	GER	Paper	AUS	GER
Baesens et al. [32]	90.40	74.60	Nieddu et al. [139]	87.30	79.20
Hsieh [118]	98.00	98.50	Marcano-Cedeno et al. [84]	92.75	84.67
Somol et al. [54]	92.60	83.80	Ping and Yongheng [122]	87.52	76.60
Lan et al. [63]	86.96	74.40	Yu and Li [58]	85.65	72.60
Hoffmann et al. [22]	85.80	73.40	Chang and Yeh [140]	85.36	77.10
Huang et al. [117]	87.00	78.10	Wang et al. [125]	88.17	78.52
Tsai and Wu [141]	97.32	78.97	Hens and Tiwari [94]	85.98	75.08
Tsai [44]	90.20	79.11	Vukovic et al. [124]	88.55	77.40
Tsai [55]	81.93	74.28	Marques et al. [142]	86.81	76.60
Luo et al. [47]	86.52	84.80	Ling et al. [95]	87.85	79.55
Lahsasna et al. [107]	88.60	75.00	Sadatrasoul et al. [143]	84.83	73.51
Chen and Li [57]	86.52	76.70	Zhang et al. [144]	88.84	73.20
Zhang et al. [127]	91.97	81.64	Liang et al. [145]	86.09	74.16
Liu et al. [121]	86.84	75.75	Tsai et al. [146]	87.23	76.48
Wang et al. [52]	86.57	76.30	Zhu et al. [147]	86.78	76.62

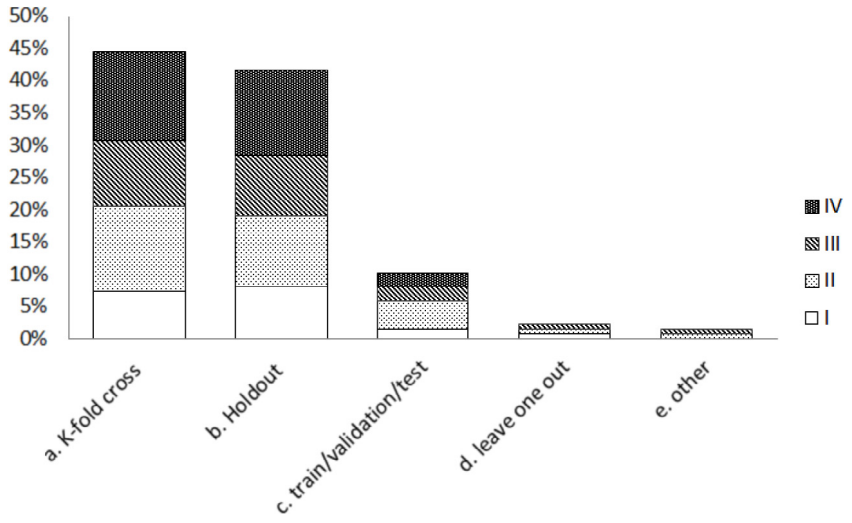

Fig. 12. The type of validation methods.

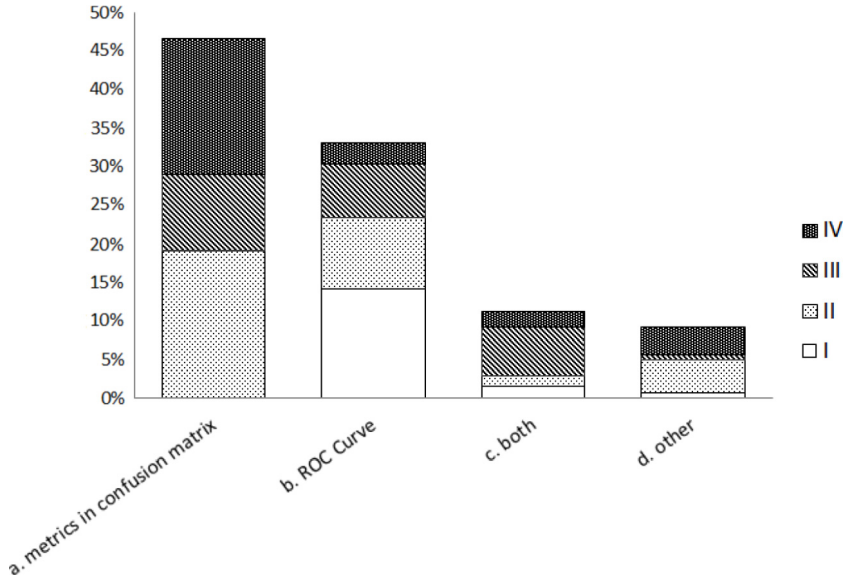

Fig. 13. The misclassification criteria.

shows the cumulative bar plots to all considered periods. In general, OR focus is the most used (59.2%) with a low use in the first time period considered and superior at the others. The MS focus was more used in the first time period (19.4%). In this sense, we can state that there is a recent increased interest in OR as opposed to MS in the credit scoring area.

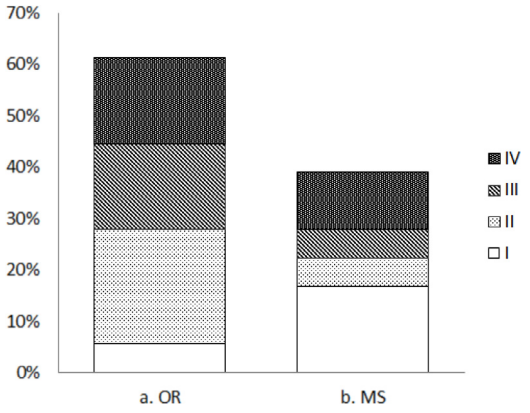

Fig. 14. Principal focus of the paper concerning the decision area.

5. Is there a better method? a comparison study

In this section, all presented methods are compared using two frameworks, marked out by two predictive performance measures, AC (Approximate Correlation) and FM (F1-score Measure) for three different benchmark datasets: (A) Australian Credit, (B) Credit German and (C) Japanese Credit, available in UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/). For each dataset we performed 1000 replications in a handout validation approach (70% training sample and 30% test sample) under a balanced base ($p = 0.5\%$, 50% of bad payers) and an unbalanced base ($p = 0.1\%$, 10% of bad payers). The methods were implemented in Sofware R 3.0.2 through RBase with the packages: *nnet*, *MASS*, *rpart*, *rgp*, *e*1071 and *frbs* on a HP Pavilion PC i7-3610QM 2.30 GHz CPU, RAM 8 GB, Windows 7 64-bit.

Taking into account the all comparisons, Figs. 16 and 17, we noticed the highlight of two methods, SVM and FUZZY, that permeate this comparison study as the two best techniques of

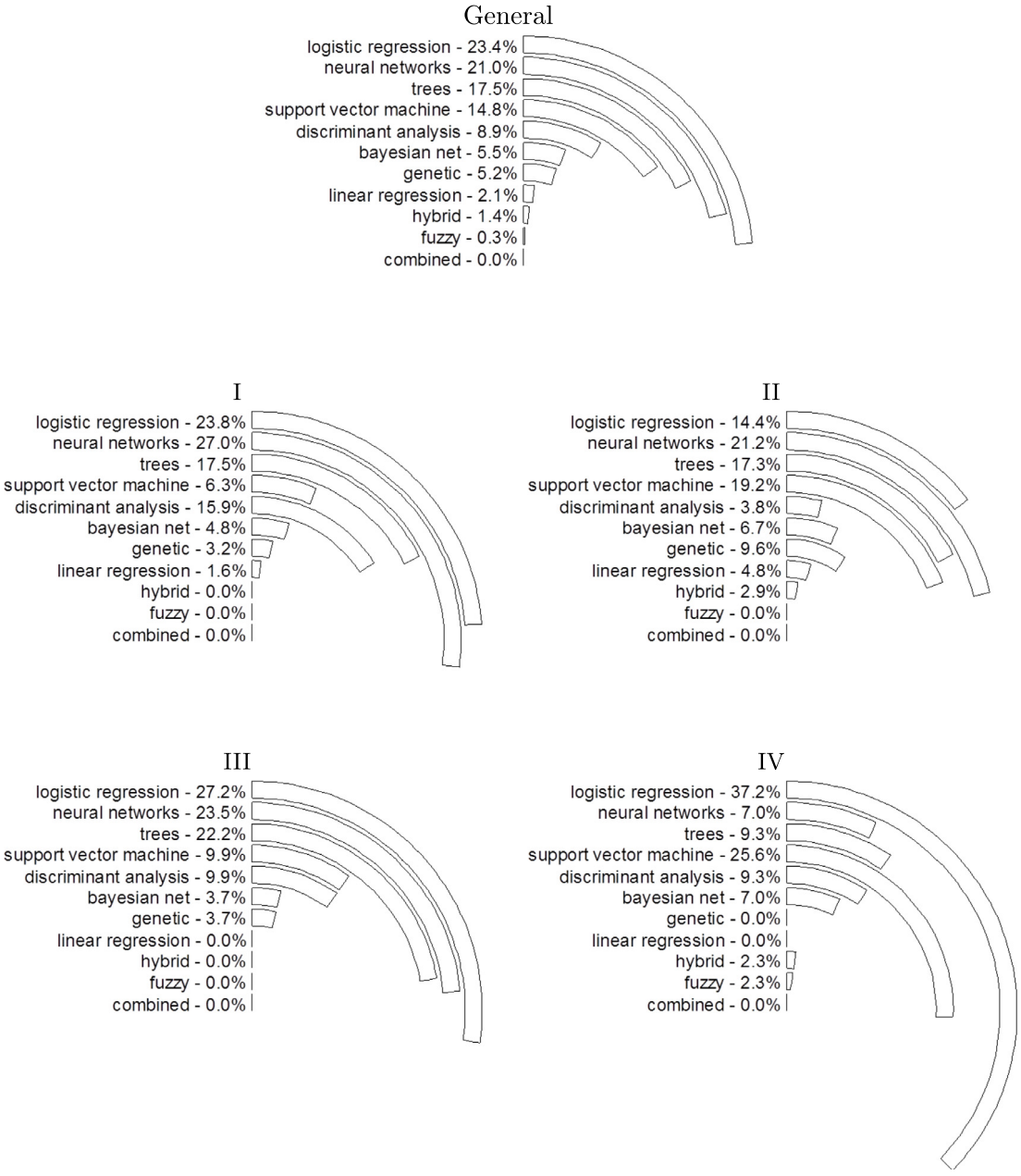

Fig. 15. The techniques used in the paper's comparison studies.

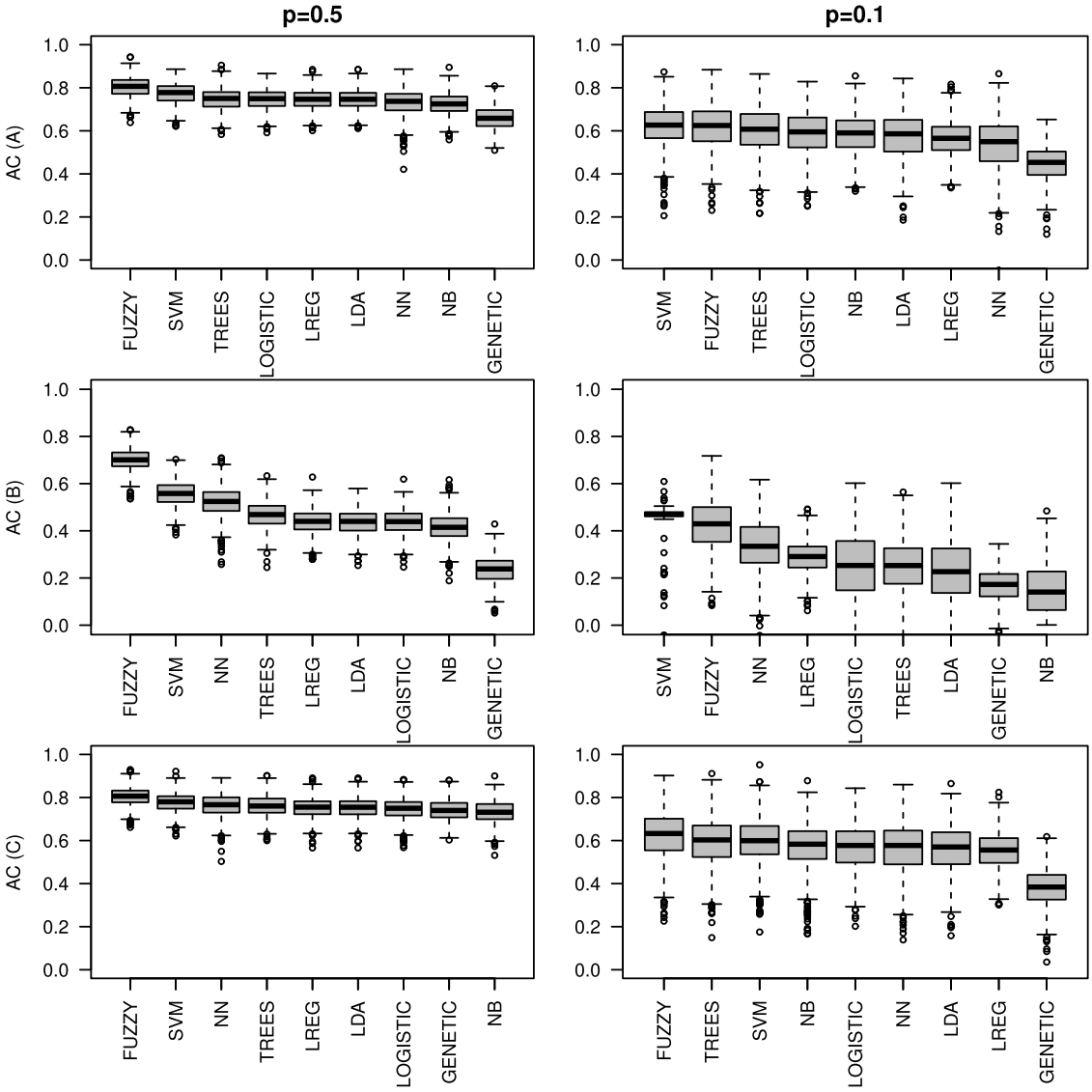

Fig. 16. Approximate correlation results.

greater predictive performance for both measures evaluated, this fact is confirmed by the Kruskal–Wallis test at a significance level of 5% ($p\text{-}values < 2e - 16$).

However we noticed that in most cases there is a shift of the predictive performance of both when unbalance in the number of bad payers occurs. For $p = 0.5$ FUZZY is given as the method with greater predictive performance, with SVM as the second method. For $p = 0.1$ SVM is given as the method with greater predictive performance, with FUZZY as the second method. Alternatively, TREES is often the third best method and it is independent of the unbalance. In addition, we noticed most often that NN lost the predictive performance when there is imbalance. The LOGISTIC, NB and LDA methods do not seem to present any standard, with predictive performance behavior between the median methods. GENETIC and LREG are considered as the smaller predictive performance when there is imbalance.

Table 6 displays computational time (in seconds) for the implementation of methods for each replication. Among the methods with greater predictive performance, SVM (0.37 s) has a much lower computational effort than FUZZY (48.92 s). GENETIC and FUZZY are the methods with higher computational effort. In summary among the analyzed methods, SVM stands out as a method of high predictive performance and low computational effort than others.

6. Final comments

We present in this paper a methodologically structured systematic literature review of binary classification techniques for credit scoring financial analysis. 187 papers on credit scoring published in scientific journals during the two last decades (1992–2015) were analyzed and classified. Based on the survey, we observed an increasing number of papers in this area and noticed that credit scoring analysis is a current and significant financial area, a plenteous area for the application of statistical and data mining tools.

Although, regardless of the time period, the most common main objective of the revised papers is to propose a new method for rating in credit scoring, especially with hybrid techniques, a similarity between the predictive performance of the methods is observed. This result is corroborated by Hand [148]. Moreover, comparison with traditional techniques was rarely performed in recent time periods. This fact shows that, though the researchers are giving up to compare techniques, the pursuit of a general

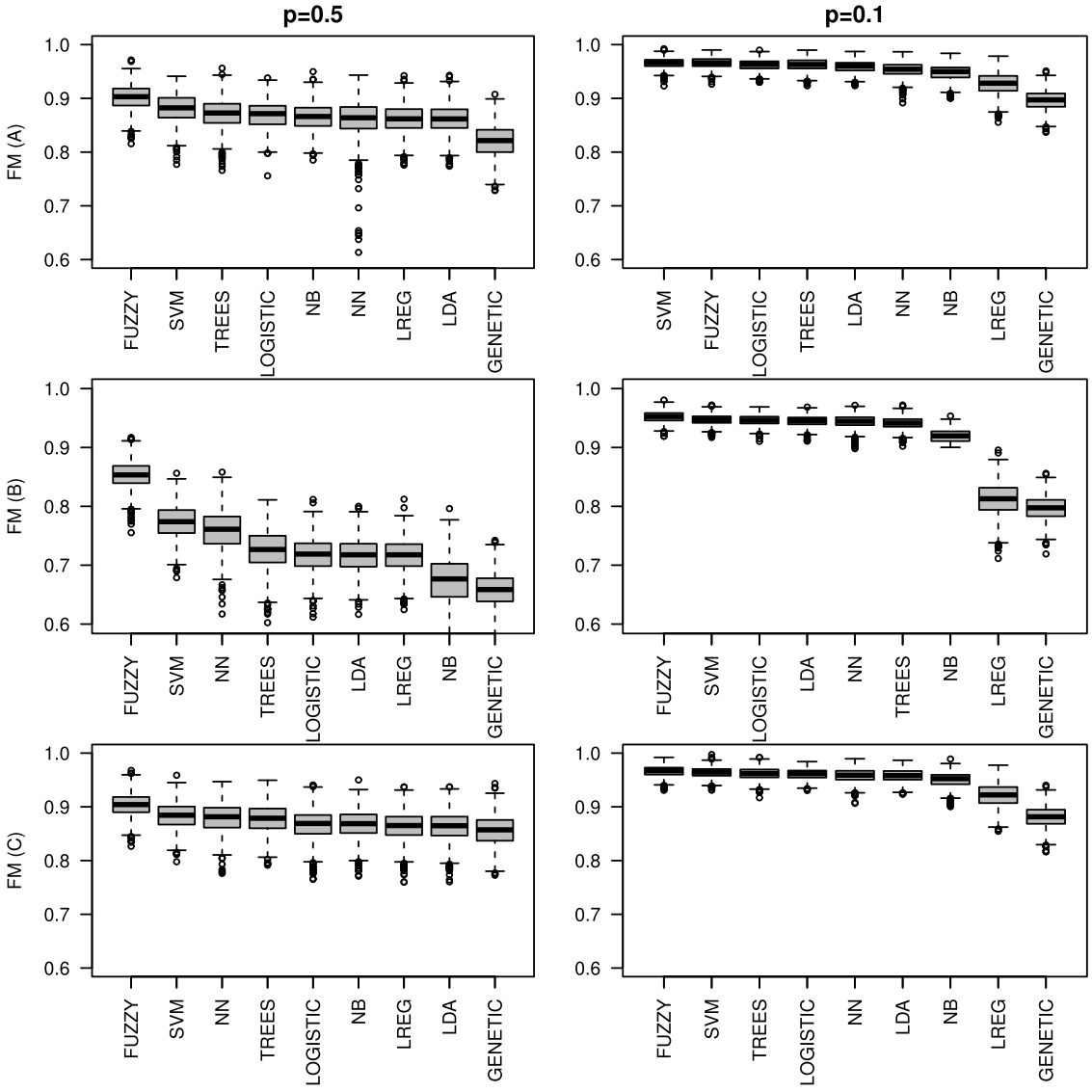

Fig. 17. $F1$-score measure results.

Table 6
Time in seconds for each method's replication.

Dataset	GENETIC	FUZZY	NN	LOGISTIC	SVM	NB	TREES	LDA	LREG
(A)	41.69	39.69	0.66	0.24	0.34	0.34	0.25	0.30	0.23
(B)	42.58	24.31	0.41	0.23	0.31	0.33	0.27	0.23	0.23
(C)	146.35	82.77	1.38	0.74	0.47	0.45	0.28	0.25	0.24
Average	76.87	48.92	0.82	0.40	0.37	0.37	0.27	0.26	0.23

method with a high predictive performance continues. On the other hand, other types of researches in credit scoring are required as conceptual discussions based on data quality, database enrichment, time dependence, classes type and so on.

While knowing these mishaps, for the moment, neural networks, support vector machine, hybrid and combined techniques appear as the most common main tools. The logistic regression, trees and also neural networks are mostly used in comparisons of techniques as standards that must be overcome. In general, support vector machine appears as a method of high predictive performance and low computational effort than other methods. Regarding datasets for credit scoring, the number has been increasing as well as the presence of a mixture of continuous and discrete variables. The majority of datasets however are private and there is a wide usage of the well known German and Australian datasets. This fact shows how difficult it is to obtain datasets on the credit scoring scenario, since there are issues related to maintenance of confidentiality of credit scoring databases.

The K-fold cross validation and holdout are the most common validation methods. Care should be taken when interpreting the results of both methods, because they are different methods and subject to subjectivity of the random distribution of the database. The use of ROC Curve as unique misclassification criterion has decreased significantly in the articles over the years. More recently the use of metrics based on confusion matrix is most common. Also, there are a small number of papers that handled missing data in credit scoring analysis and a high frequency of papers that applied feature selection procedures as pre-proceeding method. Moreover,

Table A.1
List of questions and of possible responses to the proposed systematic review.

1. *Which is the main objective of the paper?*	6. *Was missing values imputation performed?*
a. Proposing a new method for rating	a. Yes
b. Comparing traditional techniques	b. No
c. Conceptual discussion	
d. Feature selection	7. *What is the number of datasets used in the paper?*
e. Literature review	
f. Performance measures	8. *Was exhaustive simulation study performed?*
g. Other issues	a. Yes
	b. No
2. *What is the type of the main classification method?*	
a. Neural networks	9. *What is the type of validation of the approach?*
b. Support vector machine	a. K-fold cross
c. Linear regression	b. Handout
d. Trees	c. Train/validation/test
e. Logistic regression	d. Leave one out
f. Fuzzy	e. Other
g. Genetic	
h. Discriminant Analysis	10. *What is the type of misclassification cost criterion?*
i. Bayesian net	a. ROC curve
j. Hybrid	b. Metrics based on confusion matrix
k. Combined	c. Both
l. Others	d. Others
3. *Which is the type of the datasets used?*	11. *Does the paper use the Australian or the German datasets?*
a. Public	a. Yes
b. Particular	b. No
c. Both	
	12. *Which is the principal classification method used in comparison study?*
4. *Which is the type of the explanatory variables?*	a. Neural networks
a. Categorical	b. Support vector machine
b. Continuous	c. Linear regression
c. Mixed	d. Trees
	e. Logistic regression
5. *Does the paper perform variable selection methods?*	f. Fuzzy
a. Yes	g. Genetic
b. No	h. Discriminant analysis
	i. Bayesian net
	j. Others
	11. *Which is the principal focus of the paper concerning the decision area?*
	a. OR
	b. MS

in the credit scoring area, there is a recent increased interest in OR focus as opposed to MS one.

Although our systematic literature review is exhaustive, some limitations still persist. First, the findings were based on papers published in English and in scientific journals inside the following databases: ScienceDirect, Engineering Information, Reaxys and Scopus. Although such databases cover more than 20,000 journal titles, other databases may be hereafter included in the survey. Secondly, as pointed out in Section 2, we did not include in the survey other forms of publication such as unpublished working papers, master and doctoral dissertations, books, conference proceedings, white papers and others. Moreover, high quality research is eventually published in scientific journals, other forms of publication may be included in this list in future investigations. Notwithstanding these limitations, our systematic review provides important insights into the research literature on classification techniques applied to credit scoring and how this area has been moving over time.

Acknowledgments

This research was sponsored by the Brazilian organizations CNPq and FAPESP and by Serasa Experian, through their research grant programs.

Appendix

See Table A.1.

References

[1] D. Durand, Risk elements in consumer instalment financing, in: National Bureau of Economics, New York, 1941.
[2] J. Banasik, J.N. Crook, L.C. Thomas, Not if but when will borrowers default, J. Oper. Res. Soc. 50 (12) (1999) 1185–1190.
[3] D. Marron, 'Lending by numbers': Credit scoring and the constitution of risk within American consumer credit, Econ. Soc. 36 (1) (2007) 103–133.
[4] F. Louzada, P.H. Ferreira-Silva, C.A.R. Diniz, On the impact of disproportional samples in credit scoring models: An application to a Brazilian bank data, Expert Syst. Appl. 39 (9) (2012) 8071–8078.
[5] L.C. Thomas, D. Edelman, J. Crook, Credit Scoring and its Applications, in: Monographs on Mathematical Modeling and Computation, SIAM, 2002.
[6] Jiawei Han, Micheline Kamber, Jian Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2006.
[7] V.M. Rohit, S. Kumar, J. Kumar, Basel ii to basel iii ï the way forward. Technical Report, Infosys, 2013.
[8] R.B.N.Z. Staff, Statement of principles: Bank registration and supervision financial stability, Banking System Handbook, 2013.
[9] D.J. Hand, W.E. Henley, Statistical classification methods in consumer credit scoring: A review, J. Roy. Statist. Soc. Ser. A 160 (3) (1997) 523–541.
[10] X. Xu, C. Zhou, Z. Wang, Credit scoring algorithm based on link analysis ranking with support vector machine, Expert Syst. Appl. 36 (2 PART 2) (2009) 2625–2632.
[11] Y. Shi, Multiple criteria optimization-based data mining methods and applications: A systematic survey, Knowl. Inf. Syst. 24 (3) (2010) 369–391.
[12] A. Lahsasna, R.N. Ainon, T.Y. Wah, Credit scoring models using soft computing methods: A survey, Int. Arab J. Inf. Technol. 7 (2) (2010) 115–123.
[13] K. Nurlybayeva, G. Balakayeva, Algorithmic scoring models, Appl. Math. Sci. 7 (9–12) (2013) 571–586. cited By 3.
[14] V. Garcia, A.I. Marques, J.S. Sanchez, An insight into the experimental design for credit risk and corporate bankruptcy prediction systems, J. Intell. Inf. Syst. 44 (1) (2014) 159–189. cited By 0.
[15] S. Lessmann, B. Baesens, H.-V. Seow, L.C. Thomas, Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research, European J. Oper. Res. 247 (1) (2015) 124–136. cited By 2.

[16] Bernard W. Taylor, C.R. Bector, S.K. Bhatt, Earl Saul Rosenbloom, Introduction to Management Science, Prentice Hall, New Jersey, 1996.
[17] John Mingers, Leroy White, A review of the recent contribution of systems thinking to operational research and management science, European J. Oper. Res. 207 (3) (2010) 1147–1161.
[18] Wafik Hachicha, Ahmed Ghorbel, A survey of control-chart pattern-recognition literature (1991–2010) based on a new conceptual classification scheme, Comput. Ind. Eng. 63 (1) (2012) 204–222.
[19] R. Kolbe, M. Brunette, Content analysis research: An examination of applications with directives for improving research, reliability and objectivity, J. Consum. Res. 18 (2) (1991) 243–250.
[20] T.-S. Lee, C.-C. Chiu, C.-J. Lu, I.-F. Chen, Credit scoring using the hybrid neural discriminant technique, Expert Syst. Appl. 23 (3) (2002) 245–254.
[21] T.V. Gestel, B. Baesens, J.A.K. Suykens, D. Van den Poel, D.-E. Baestaens, M. Willekens, Bayesian kernel based classification for financial distress detection, European J. Oper. Res. 172 (3) (2006) 979–1003.
[22] F. Hoffmann, B. Baesens, C. Mues, T. Van Gestel, J. Vanthienen, Inferring descriptive and approximate fuzzy rules for credit scoring using evolutionary algorithms, European J. Oper. Res. 177 (1) (2007) 540–555.
[23] N.-C. Hsieh, L.-P. Hung, A data driven ensemble classifier for credit scoring analysis, Expert Syst. Appl. 37 (1) (2010) 534–545.
[24] J. Gemela, Financial analysis using Bayesian networks, Appl. Stoch. Models Bus. Ind. 17 (1) (2001) 57–67.
[25] J. Van Gool, W. Verbeke, P. Sercu, B. Baesens, Credit scoring for microfinance: Is it worth it? Int. J. Financ. Econ. 17 (2) (2012) 103–123.
[26] M. Bardos, Detecting the risk of company failure at the banque de france, J. Bank. Finance 22 (10–11) (1998) 1405–1419.
[27] D.J. Hand, Modelling consumer credit risk, IMA J. Manag. Math. 12 (2) (2001) 139–155.
[28] D. Martens, T. Van Gestel, M. De Backer, R. Haesen, J. Vanthienen, B. Baesens, Credit rating prediction using ant colony optimization, J. Oper. Res. Soc. 61 (4) (2010) 561–573.
[29] M.-C. Chen, S.-H. Huang, Credit scoring and rejected instances reassigning through evolutionary computation techniques, Expert Syst. Appl. 24 (4) (2003) 433–441.
[30] L.C. Thomas, Consumer finance: Challenges for operational research, J. Oper. Res. Soc. 61 (1) (2010) 41–52.
[31] D. West, Neural network credit scoring models, Comput. Oper. Res. 27 (11–12) (2000) 1131–1152.
[32] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, J. Vanthienen, Benchmarking state-of-the-art classification algorithms for credit scoring, J. Oper. Res. Soc. 54 (6) (2003) 627–635.
[33] N.M. Adams, D.J. Hand, R.J. Till, Mining for classes and patterns in behavioural data, J. Oper. Res. Soc. 52 (9) (2001) 1017–1024.
[34] F. Hoffmann, B. Baesens, J. Martens, F. Put, J. Vanthienen, Comparing a genetic fuzzy and a neurofuzzy classifier for credit scoring, Int. J. Intell. Syst. 17 (11) (2002) 1067–1083.
[35] C.-S. Ong, J.-J. Huang, G.-H. Tzeng, Building credit scoring models using genetic programming, Expert Syst. Appl. 29 (1) (2005) 41–47.
[36] B. Baesens, T. Van Gestel, M. Stepanova, D. Van Den Poel, J. Vanthienen, Neural network survival analysis for personal loan data, J. Oper. Res. Soc. 56 (9) (2005) 1089–1098.
[37] 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.
[38] T.-S. Lee, C.-C. Chiu, Y.-C. Chou, C.-J. Lu, Mining the customer credit using classification and regression tree and multivariate adaptive regression splines, Comput. Statist. Data Anal. 50 (4) (2006) 1113–1130.
[39] Y.-M. Huang, C.-M. Hung, H.C. Jiau, Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem, Nonlinear Anal. RWA 7 (4) (2006) 720–747.
[40] W. Xiao, Q. Zhao, Q. Fei, A comparative study of data mining methods in consumer loans credit scoring management, J. Syst. Sci. Syst. Eng. 15 (4) (2006) 419–435.
[41] T. Van Gestel, D. Martens, B. Baesens, D. Feremans, J. Huysmans, J. Vanthienen, Forecasting and analyzing insurance companies' ratings, Int. J. Forecast. 23 (3) (2007) 513–529.
[42] D. Martens, B. Baesens, T. Van, Gestel, J. Vanthienen, Comprehensible credit scoring models using rule extraction from support vector machines, European J. Oper. Res. 183 (3) (2007) 1466–1476.
[43] Y.-C. Hu, J. Ansell, Measuring retail company performance using credit scoring techniques, European J. Oper. Res. 183 (3) (2007) 1595–1606.
[44] C.-F. Tsai, Financial decision support using neural networks and support vector machines, Expert Syst. 25 (4) (2008) 380–393.
[45] H. Abdou, J. Pointon, A. El-Masry, Neural nets versus conventional techniques in credit scoring in Egyptian banking, Expert Syst. Appl. 35 (3) (2008) 1275–1292.
[46] A.P. Sinha, H. Zhao, Incorporating domain knowledge into data mining classifiers: An application in indirect lending, Decis. Support Syst. 46 (1) (2008) 287–299.
[47] S.-T. Luo, B.-W. Cheng, C.-H. Hsieh, Prediction model building with clustering-launched classification and support vector machines in credit scoring, Expert Syst. Appl. 36 (4) (2009) 7562–7566.
[48] S. Finlay, Are we modelling the right thing? The impact of incorrect problem specification in credit scoring, Expert Syst. Appl. 36 (5) (2009) 9065–9071.
[49] H.A. Abdou, Genetic programming for credit scoring: The case of Egyptian public sector banks, Expert Syst. Appl. 36 (9) (2009) 11402–11417.
[50] Y.-C. Hu, J. Ansell, Retail default prediction by using sequential minimal optimization technique, J. Forecast. 28 (8) (2009) 651–666.
[51] S. Finlay, Credit scoring for profitability objectives, European J. Oper. Res. 202 (2) (2010) 528–537.
[52] G. Wang, J. Hao, J. Ma, H. Jiang, A comparative assessment of ensemble learning for credit scoring, Expert Syst. Appl. 38 (1) (2011) 223–230.
[53] Y. Liu, M. Schumann, Data mining feature selection for credit scoring models, J. Oper. Res. Soc. 56 (9) (2005) 1099–1108.
[54] P. Somol, B. Baesens, P. Pudil, J. Vanthienen, Filter- versus wrapper-based feature selection for credit scoring, Int. J. Intell. Syst. 20 (10) (2005) 985–999.
[55] C.-F. Tsai, Feature selection in bankruptcy prediction, Knowl.-Based Syst. 22 (2) (2009) 120–127.
[56] K. Falangis, J.J. Glen, Heuristics for feature selection in mathematical programming discriminant analysis models, J. Oper. Res. Soc. 61 (5) (2010) 804–812.
[57] F.-L. Chen, F.-C. Li, Combination of feature selection approaches with svm in credit scoring, Expert Syst. Appl. 37 (7) (2010) 4902–4909.
[58] J.-L. Yu, H. Li, On performance of feature normalization in classification with distance-based case-based reasoning, Recent Pat. Comput. Sci. 4 (3) (2011) 203–210.
[59] R.A. McDonald, M. Sturgess, K. Smith, M.S. Hawkins, E.X.M. Huang, Non-linearity of scorecard log-odds, Int. J. Forecast. 28 (1) (2012) 239–247.
[60] J. Wang, A.-R. Hedar, S. Wang, J. Ma, Rough set and scatter search metaheuristic based feature selection for credit scoring, Expert Syst. Appl. 39 (6) (2012) 6123–6128.
[61] Z. Yang, Y. Wang, Y. Bai, X. Zhang, Measuring scorecard performance, Lect. Notes Comput. Sci. 3039 (2004) 900–906. (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).
[62] D.J. Hand, Good practice in retail credit scorecard assessment, J. Oper. Res. Soc. 56 (9) (2005) 1109–1117.
[63] Y. Lan, D. Janssens, G. Chen, G. Wets, Improving associative classification by incorporating novel interestingness measures, Expert Syst. Appl. 31 (1) (2006) 184–192.
[64] A.L. Dryver, J. Sukkasem, Validating risk models with a focus on credit scoring models, J. Stat. Comput. Simul. 79 (2) (2009) 181–193.
[65] H.A. Ziari, D.J. Leatham, P.N. Ellinger, Development of statistical discriminant mathematical programming model via resampling estimation techniques, Am. J. Agric. Econ. 79 (4) (1997) 1352–1362.
[66] G. Verstraeten, D. Van Den Poel, The impact of sample bias on consumer credit scoring performance and profitability, J. Oper. Res. Soc. 56 (8) (2005) 981–992.
[67] M. Rezac, Advanced empirical estimate of information value for credit scoring models, Acta Univ. Agric. Silviculturae Mendelianae Brun. 59 (2) (2011) 267–274.
[68] K. Bijak, L.C. Thomas, Does segmentation always improve model performance in credit scoring? Expert Syst. Appl. 39 (3) (2012) 2433–2442.
[69] B.D. Ripley, Pattern Recognition and Neural Networks, Cambridge University Press, 1996.
[70] V. Vapnik, Statistical Learning Theory, 1998.
[71] D.J. Hand, M.G. Kelly, Superscorecards, IMA J. Manag. Math. 13 (4) (2002) 273–281.
[72] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, Charles J. Stone, Classification and Regression Trees, Wadsworth & Brooks, Monterey, CA, 1984.
[73] Joseph Berkson, Application of the logistic function to bio-assay, J. Amer. Statist. Assoc. 39 (227) (1944) 357–365.
[74] Lotfi A. Zadeh, Fuzzy sets, Inf. Control 8 (3) (1965) 338–353.
[75] John R. Koza, Genetic programming: on the programming of computers by means of natural selection, Complex Adapt. Syst. (1992).
[76] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugenics 7 (1986) 179–188.
[77] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Mach. Learn. 29 (2–3) (1997) 131–163.
[78] Leo Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.
[79] Robert E. Schapire, The strength of weak learnability, Mach. Learn. 5 (2) (1990) 197–227.
[80] David H. Wolpert, Stacked generalization, Neural Netw. 5 (2) (1992) 241–259.
[81] Roderick J.A. Little, Donald B. Rubin, Statistical Analysis with Missing Data, 2002.
[82] S.-L. Pang, Study on credit scoring model and forecasting based on probabilistic neural network, Xitong Gongcheng Lilun yu Shijian/Syst. Eng. Theory Pract. 25 (5) (2005) 43–48.
[83] P.J.G. Lisboa, T.A. Etchells, I.H. Jarman, C.T.C. Arsene, M.S.H. Aung, A. Eleuteri, A.F.G. Taktak, F. Ambrogi, P. Boracchi, E. Biganzoli, Partial logistic artificial neural network for competing risks regularized with automatic relevance determination, IEEE Trans. Neural Netw. 20 (9) (2009) 1403–1416.
[84] A. Marcano-Cedeno, A. Marin-De-La-Barcena, J. Jimenez-Trillo, J.A. Pinuela, D. Andina, Artificial metaplasticity neural network applied to credit scoring, Int. J. Neural Syst. 21 (4) (2011) 311–317.
[85] C.-L. Chuang, S.-T. Huang, A hybrid neural network approach for credit scoring, Expert Syst. 28 (2) (2011) 185–196.
[86] W. Chen, C. Ma, L. Ma, Mining the customer credit using hybrid support vector machine technique, Expert Syst. Appl. 36 (4) (2009) 7611–7616.
[87] S.-T. Li, W. Shiue, M.-H. Huang, The evaluation of consumer loans using support vector machines, Expert Syst. Appl. 30 (4) (2006) 772–782.
[88] W.-B. Xiao, Q. Fei, A study of personal credit scoring models on support vector machine with optimal choice of Kernel function parameters, Xitong Gongcheng Lilun yu Shijian/Syst. Eng. Theory Pract. 26 (10) (2006) 73–79.

[89] Y. Yang, Adaptive credit scoring with Kernel learning methods, European J. Oper. Res. 183 (3) (2007) 1521–1536.
[90] C.-L. Chuang, R.-H. Lin, Constructing a reassigning credit scoring model, Expert Syst. Appl. 36 (2 PART 1) (2009) 1685–1694.
[91] L. Zhou, K.K. Lai, J. Yen, Credit scoring models with auc maximization based on weighted svm, Int. J. Inf. Technol. Decis. Mak. 8 (4) (2009) 677–696.
[92] L. Zhou, K.K. Lai, L. Yu, Least squares support vector machines ensemble models for credit scoring, Expert Syst. Appl. 37 (1) (2010) 127–133.
[93] L. Feng, Y. Yao, B. Jin, Research on credit scoring model with svm for network management, J. Comput. Inf. Syst. 6 (11) (2010) 3567–3574.
[94] A.B. Hens, M.K. Tiwari, Computational time reduction for credit scoring: An integrated approach based on support vector machine and stratified sampling method, Expert Syst. Appl. 39 (8) (2012) 6774–6781.
[95] Y. Ling, Q. Cao, H. Zhang, Credit scoring using multi-kernel support vector machine and chaos particle swarm optimization, Int. J. Comput. Intell. Appl. 11 (3) (2012) 12500198:1–12500198:13.
[96] D. Karlis, M. Rahmouni, Analysis of defaulters' behaviour using the Poisson-mixture approach, IMA J. Manag. Math. 18 (3) (2007) 297–311.
[97] J. Banasik, J. Crook, L. Thomas, Sample selection bias in credit scoring models, J. Oper. Res. Soc. 54 (8) (2003) 822–832.
[98] S. Efromovich, Oracle inequality for conditional density estimation and an actuarial example, Ann. Inst. Statist. Math. 62 (2) (2010) 249–275.
[99] B.W. Yap, S.H. Ong, N.H.M. Husain, Using data mining to improve assessment of credit worthiness via credit scoring models, Expert Syst. Appl. 38 (10) (2011) 13274–13283.
[100] G.H. John, P. Miller, R. Kerber, Stock selection using rule induction, IEEE Expert-Intell. Syst. Appl. 11 (5) (1996) 52–58.
[101] L.-J. Kao, C.-C. Chiu, F.-Y. Chiu, A Bayesian latent variable model with classification and regression tree approach for behavior and credit scoring, Knowl.-Based Syst. 36 (2012) 245–252.
[102] H.G. Li, D.J. Hand, Direct versus indirect credit scoring classifications, J. Oper. Res. Soc. 53 (6) (2002) 647–654.
[103] T.-S. Lee, I.-F. Chen, A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines, Expert Syst. Appl. 28 (4) (2005) 743–752.
[104] N.G. Pavlidis, D.K. Tasoulis, N.M. Adams, D.J. Hand, Adaptive consumer credit classification, J. Oper. Res. Soc. 63 (12) (2012) 1645–1654.
[105] F. Louzada, O. Anacleto-Junior, C. Candolo, J. Mazucheli, Poly-bagging predictors for classification modelling for credit scoring, Expert Syst. Appl. 38 (10) (2011) 12717–12720.
[106] A. Laha, Building contextual classifiers by integrating fuzzy rule based classification technique and k-nn method for credit scoring, Adv. Eng. Inf. 21 (3) (2007) 281–291.
[107] A. Lahsasna, R.N. Ainon, T.Y. Wah, Enhancement of transparency and accuracy of credit scoring models through genetic fuzzy classifier, Maejo Int. J. Sci. Technol. 4 (1) (2010) 136–158.
[108] J.-J. Huang, G.-H. Tzeng, C.-S. Ong, Two-stage genetic programming (2sgp) for the credit scoring model, Appl. Math. Comput. 174 (2) (2006) 1039–1053.
[109] C. Mues, B. Baesens, C.M. Files, J. Vanthienen, Decision diagrams in machine learning: An empirical study on real-life credit-risk data, Expert Syst. Appl. 27 (2) (2004) 257–264.
[110] C. Won, J. Kim, J.K. Bae, Using genetic algorithm based knowledge refinement model for dividend policy forecasting, Expert Syst. Appl. 39 (18) (2012) 13472–13479.
[111] Otman Abdoun, Jaafar Abouchabaka, A comparative study of adaptive crossover operators for genetic algorithms to resolve the traveling salesman problem, Int. J. Comput. Appl. 31 (11) (2011) 49–57.
[112] S. Akkoc, An empirical comparison of conventional techniques, neural networks and the three stage hybrid adaptive neuro fuzzy inference system (anfis) model for credit scoring analysis: The case of Turkish credit card data, European J. Oper. Res. 222 (1) (2012) 168–178.
[113] P. Giudici, Bayesian data mining, with application to benchmarking and credit scoring, Appl. Stoch. Models Bus. Ind. 17 (1) (2001) 69–81.
[114] H. Zhu, P.A. Beling, G.A. Overstreet, A Bayesian framework for the combination of classifier outputs, J. Oper. Res. Soc. 53 (7) (2002) 719–727.
[115] A.C. Antonakis, M.E. Sfakianakis, Assessing naïve Bayes as a method for screening credit applicants, J. Appl. Stat. 36 (5) (2009) 537–545.
[116] W.-W. Wu, Improving classification accuracy and causal knowledge for better credit decisions, Int. J. Neural Syst. 21 (4) (2011) 297–309.
[117] C.-L. Huang, M.-C. Chen, C.-J. Wang, Credit scoring with a data mining approach based on support vector machines, Expert Syst. Appl. 33 (4) (2007) 847–856.
[118] N.-C. Hsieh, Hybrid mining approach in the design of credit scoring models, Expert Syst. Appl. 28 (4) (2005) 655–665.
[119] J. Huysmans, B. Baesens, J. Vanthienen, T. Van Gestel, Failure prediction with self organizing maps, Expert Syst. Appl. 30 (3) (2006) 479–487.
[120] Y. Shi, Current research trend: Information technology and decision making in 2008, Int. J. Inf. Technol. Decis. Mak. 8 (1) (2009) 1–5.
[121] X. Liu, H. Fu, W. Lin, A modified support vector machine model for credit scoring, Int. J. Comput. Intell. Syst. 3 (6) (2010) 797–803.
[122] Y. Ping, L. Yongheng, Neighborhood rough set and svm based hybrid credit scoring classifier, Expert Syst. Appl. 38 (9) (2011) 11300–11304.
[123] A. Capotorti, E. Barbanera, Credit scoring analysis using a fuzzy probabilistic rough set model, Comput. Statist. Data Anal. 56 (4) (2012) 981–994.
[124] S. Vukovic, B. Delibasic, A. Uzelac, M. Suknovic, A case-based reasoning model that uses preference theory functions for credit scoring, Expert Syst. Appl. 39 (9) (2012) 8389–8395.
[125] G. Wang, J. Ma, L. Huang, K. Xu, Two credit scoring models based on dual strategy ensemble trees, Knowl.-Based Syst. 26 (2012) 61–68.
[126] G. Paleologo, A. Elisseeff, G. Antonini, Subagging for credit scoring models, European J. Oper. Res. 201 (2) (2010) 490–499.
[127] D. Zhang, X. Zhou, S.C.H. Leung, J. Zheng, Vertical bagging decision trees model for credit scoring, Expert Syst. Appl. 37 (12) (2010) 7838–7843.
[128] S. Finlay, Multiple classifier architectures and their application to credit risk assessment, European J. Oper. Res. 210 (2) (2011) 368–378.
[129] J. Xiao, L. Xie, C. He, X. Jiang, Dynamic classifier ensemble model for customer classification with imbalanced class distribution, Expert Syst. Appl. 39 (3) (2012) 3668–3675.
[130] A.I. Marques, V. Garcia, J.S. Sanchez, Two-level classifier ensembles for credit risk assessment, Expert Syst. Appl. 39 (12) (2012) 10916–10922.
[131] A.N. Berger, W.S. Frame, N.H. Miller, Credit scoring and the availability, price, and risk of small business credit, J. Money Credit Bank. 37 (2) (2005) 191–222.
[132] W. Hardle, E. Mammen, M. Muller, Testing parametric versus semiparametric modeling in generalized linear models, J. Amer. Statist. Assoc. 93 (444) (1998) 1461–1474.
[133] F. Louzada, V.G. Cancho, M. Roman, J.G. Leite, A new long-term lifetime distribution induced by a latent complementary risk framework, J. Appl. Stat. 39 (10) (2012) 2209–2222.
[134] Tom M. Mitchell, Machine Learning, Vol. 1997, McGraw Hill, Burr Ridge, IL, 1997, p. 45.
[135] Christopher M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995.
[136] M.H. Zweig, G. Campbell, Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine, Clin. Chem. 29 (1993) 561–577.
[137] K. Bache, M. Lichman, UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.
[138] R. DeYoung, W.S. Frame, D. Glennon, P. Nigro, The information revolution and small business lending: The missing evidence, J. Financ. Serv. Res. 39 (41306) (2011) 19–33.
[139] L. Nieddu, G. Manfredi, S. D'Acunto, K. la, Regina, An optimal subclass detection method for credit scoring, World Acad. Sci. Eng. Technol. 75 (2011) 349–354.
[140] S.-Y. Chang, T.-Y. Yeh, An artificial immune classifier for credit scoring analysis, Appl. Soft Comput. J. 12 (2) (2012) 611–618.
[141] C.-F. Tsai, J.-W. Wu, Using neural network ensembles for bankruptcy prediction and credit scoring, Expert Syst. Appl. 34 (4) (2008) 2639–2649.
[142] A.I. Marques, V. Garcia, J.S. Sanchez, Exploring the behaviour of base classifiers in credit scoring ensembles, Expert Syst. Appl. 39 (11) (2012) 10244–10250.
[143] S. Sadatrasoul, M. Gholamian, K. Shahanaghi, Combination of feature selection and optimized fuzzy apriori rules: The case of credit scoring, Int. Arab J. Inf. Technol. 12 (2) (2015) 138–145. cited By 1.
[144] Z. Zhang, G. Gao, Y. Shi, Credit risk evaluation using multi-criteria optimization classifier with Kernel, fuzzification and penalty factors, European J. Oper. Res. 237 (1) (2014) 335–348. cited By 8.
[145] D. Liang, C.-F. Tsai, H.-T. Wu, The effect of feature selection on financial distress prediction, Knowl.-Based Syst. 73 (1) (2014) 289–297. cited By 1.
[146] C.-F. Tsai, Y.-F. Hsu, D.C. Yen, A comparative study of classifier ensembles for bankruptcy prediction, Appl. Soft Comput. J. 24 (2014) 977–984. cited By 4.
[147] X. Zhu, J. Li, D. Wu, H. Wang, C. Liang, Balancing accuracy, complexity and interpretability in consumer credit decision making: A c-topsis classification approach, Knowl.-Based Syst. 52 (2013) 258–267. cited By 1.
[148] D.J. Hand, Classifier technology and the illusion of progress, Statist. Sci. 21 (9) (2006) 1–14.