Expert Systems with Applications 42 (2015) 5737–5753

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. A correlated-adjusted decision forest proposal

Raquel Florez-Lopez *, Juan Manuel Ramon-Jeronimo

University Pablo Olavide of Seville, Department of Financial Economics and Accounting, Utrera Road, km. 1, 41013 Seville, Spain

ARTICLE INFO

Article history:
Available online 5 March 2015

Keywords:
Ensemble strategies
Credit scoring
Decision forests
Diversity
Gradient boosting
Random forests

ABSTRACT

Credit risk assessment is a critical topic for finance activity and bankruptcy prediction that has been broadly explored using statistical models and Machine Learning methods. Recently, studies have suggested the use of ensemble strategies to enhance credit modelling performance. However, accuracy is obtained at the expense of interpretability, leading to the reluctance of financial industry to employ ensemble models in favour of simpler models. In this work we introduce an ensemble approach based on merged decision trees, the correlated-adjusted decision forest (CADF), to produce both accurate and comprehensible models. As main innovation, our proposal explores the combination of complementary sources of diversity as mechanisms to optimise model's structure, which leads to a manageable number of comprehensive decision rules without sacrificing performance. We evaluate our approach in comparison to individual classifiers and alternative ensemble strategies (gradient boosting, random forests). Empirical results suggest CADF is an encouraging solution for credit risk problems, being able to compete in accuracy with much complex proposals while producing a rule-based structure directly useful for managerial decisions.

1. Introduction

Credit risk assessment is critical for the survival of financial and non-financial firms. As the current global financial crisis has revealed, inadequate decision making in credit grant process does not only affect profitability but often threatens firm solvency (Kestens, Van Cauwenberge, & Vauwedhe, 2012). Too restrictive, a credit granting policy will reduce sales and benefits, but too permissive will result in unpaid accounts and insolvency. In the financial industry, the increasing number of bank collapses and massive losses had lead to international banking regulations being demanding to develop more appropriate credit risk models for scoring their financial loan portfolios (Basel Committee on Banking Supervision (BCBS), 2011). A common method for making credit risk decisions is through a credit scorecard, which forecasts the probability that the customer will exhibit a certain payment behaviour departing on a group of risk drivers.

To be useful, credit scoring models should reach a good balance between accuracy and interpretability. Accuracy refers to building models with a strong classification performance that minimizes prediction error; interpretability focuses on building models being comprehensible by human users (Crook, Edelman, & Thomas, 2007; Hand, 2006). Accuracy has being largely perceived as the primary focus of credit scoring, since even a fraction of a percent improvement could leave to significant future savings and profits (Crook et al., 2007; Derelioglu & Gürgen, 2011; Henley & Hand, 1997). A large body of literature has been devoted to the evaluation of techniques to increase the accuracy of credit predictions, departing on traditional statistical models such as linear discriminant analysis (LDA) and logistic regression (LR) (Altman, 1968). Later on, non-parametric Machine Learning (ML) techniques were considered to reach a higher accuracy in presence of complex credit risk datasets (Baesens et al., 2003; Brown & Mues, 2012; Crook et al., 2007; Kruppa, Schwarz, Arminger, & Ziegler, 2013). Applications of ML techniques include k-nearest neighbours (knn) (Henley & Hand, 1997), neural networks (NN) (West, 2000), or support vector machines (SVM) (Danenas & Garsva, 2015; Harris, 2015; Huang, Chen, & Wang, 2007), multivariate adaptive regression splines (Lee & Chen, 2005), or genetic algorithms (Ong, Huang, & Tzeng, 2005) among others.[1] More recently, literature has focused on the suitability of ensemble strategies for credit risk scoring, based on combining the decisions of multiple classifiers to deliver a final aggregated output. As far as ensemble members are a broad set of diverse and accurate classifiers, the ensemble will be more robust and will exhibit a stronger classification performance than any individual member. Interest in ensemble strategies has increased significantly over the last decade (Abellan & Mantas, 2014; Finlay, 2011;

* Corresponding author. Tel.: +34 954 349 854; fax: +34 954 348 353.
E-mail addresses: rflorez@upo.es (R. Florez-Lopez), jmramjer@upo.es (J.M. Ramon-Jeronimo).

[1] For an exhaustive review of methods and applications of credit scoring we refer the reader to Baesens et al. (2003), Crook et al. (2007) or Harris (2015).

http://dx.doi.org/10.1016/j.eswa.2015.02.042
0957-4174/

Marques, Garcia, & Sanchez, 2012a; Nanni & Lumini, 2009), since literature has demonstrated their potential to outperform stand-alone accuracy from 5% to 75% (Breiman, 1996).

Besides accuracy, model comprehensibility is of vital importance in credit scoring domains. First, managers need interpretable models to justify the reasons for the denial of a credit, a banking supervision obligation in many countries (Crook et al., 2007; Hand, 2006; Tomczak & Zieba, 2015). Second, comprehensible model reduce managers' reluctance to use statistical techniques for credit decision making (Feldman & Gross, 2005; Hand, 2006; Sun, Li, Huang, & He, 2014). Finally, as far as managers understand the information they receive, they gain insight into factors that affect credit default so can combine both statistical scores and expert judgement to make proper credit decisions (Chen & Cheng, 2013; Finlay, 2011). Different techniques have been used to develop comprehensible credit risk models, as scoring tables (Tomczak & Zieba, 2015), decision trees (Daubie, Levecq, & Meskens, 2002), decision diagrams (Mues, Baesens, Files, & Vanthienen, 2004), or rule-based reasoning systems (Kim, 1993). However, accuracy and comprehensibility are two properties that can be hardly balanced, which is indicated as the accuracy-interpretability dilemma: As long as credit models gain in interpretability, they lose in accuracy, and vice versa (Chen & Cheng, 2013; Crook et al., 2007; Härdle, Moro, & Schafer, 2005). As a result, a gap emerges between credit risk research and practice-oriented needs: While literature goes on developing lots of very complex proposals, financial industry needs comprehensible models to be used in practice, so the empirical usefulness of complex learners is reduced (Chen & Cheng, 2013; Finlay, 2011; Hsieh & Hung, 2010; Mues et al., 2004).

Different authors have recognised the need to reconcile interpretability and accuracy by extracting comprehensible rules from strong classifiers as SVM (Martens, Baesens, Van Gestel, & Vanthienen, 2007; Wu & Hu, 2012), NN (Baesens et al., 2003; Derelioglu & Gürgen, 2011; Mues et al., 2004; Setiono, Baesens, & Mues, 2011), or rough sets (Chen & Cheng, 2013). While these rule-extraction models do not fully reveal the decision criterion of the original classifier (Derelioglu & Gürgen, 2011), they provide a direct mode to explain the main input-relationships of the model. In presence of ensembles of classifiers, the need of such a balance is even higher considering the potential gain that their use represents in terms of prediction accuracy and financial profits (Hsieh & Hung, 2010). However, proposals on improving the interpretability of ensemble strategies are still reduced and, as a result of model's complexity, largely focused on estimating variable importance scores instead of real knowledge (Breiman, 2001; De Bock & Van den Poel, 2012; Kruppa et al., 2013). As a result, developing ensemble models that hold the characteristics of interpretation, explanation, and understanding is one of the most significant future research topics in financial default prediction (Sun et al., 2014).

In this paper we approach the problem of model's interpretability in terms of diversity, a key prerequisite for building adequate ensemble techniques (Breiman, 1996; Dietterich, 2000; Sun et al., 2014; Zhou, Lai, & Yu, 2010). Literature has pointed two main strategies for inducing diversity (Sun et al., 2014): multiple data partitions (instance and feature diversity), and multiple learning algorithms (classifier diversity). Besides, diversity may be enhanced by selecting an appropriate combination function to merge base learners (Zhou et al., 2010). Departing on their different nature and purpose, synergistic results are expected by using diversity sources in conjunction, since the variety produced with a method can be improved with the diversity produced by other method. However, the strategy of including multiple sources of diversity has been scarcely used in practice, and approaches have focused on exploiting accuracy gains. Instead of searching for a better performance, our proposal tries to exploit diversity to optimise model's structure. We depart on a simple idea: since diversity increases the accuracy of a fixed number of merged learners, diversity could also reduce the number of base learners that must be merged to maintaining the initial accuracy rate (Zhou et al., 2010). If ensemble models use comprehensible base learners as decision trees (so-called decision forests), such a complexity reduction would directly enhance model's interpretability in terms of decision rule extraction.

This paper introduces a new ensemble proposal, the correlated-adjusted decision forest (CADF), which tries to balance the superior accuracy of ensemble strategies with a high level of interpretability. Our proposal departs on decision trees as base learners, including complementary sources of diversity while controlling model's complexity. First, a multiple classifier strategy is considered that merges five different inductive models from a single dataset; since each model implements a different wrapper-feature selection process, feature diversity is also introduced in the proposal. Besides, instance diversity is included by using 10-fold cross validation for tree building, while bootstrapping samples are used for out-of-sample estimates. Finally, diversity is enhanced by introducing a new pseudo-R2 penalty function that combines decision trees using a correlation-adjusted weighted voting scheme.

For testing and illustration purposes, CADF is applied to the German credit risk dataset from UCI repository. Different scoring techniques are applied as benchmarking references including single statistical models (LDA, LR), ML classifiers (knn, NN, linear SVM, 2-degree polynomial SVM), and decision trees (ChAID, Assistant, C4.5, CART univariate, CART oblique). Besides, we are particularly interested in the comparison to alternative ensembles of decision trees (gradient-boosting and random forests), to test if CADF is able to obtain a similar accuracy than multiple data partition ensembles but departing on much reduced and better comprehensible rules. Models are evaluated in terms of their accuracy and interpretability. First is computed in terms of the accuracy rate, type I error, and type II error, which are particularly interesting to analyse sensitivity to data imbalance (Li, Tsang, & Chaudhari, 2012; Marques et al., 2012a). Models are also evaluated using the area under the receiver operating characteristic curve (AUC), a measure of discriminatory power that is independent of class distribution or misclassification cost (Hand, 2009; Henley & Hand, 1997). To make inferences from differences in accuracy, we use non-parametric tests for the statistical comparison of accuracy rates (McNemar and Wilcoxon paired tests); differences in AUC are tested using the Friedman test, and the post hoc Nemenyi Bonferroni–Dunn test. Complexity-based and semantic-based interpretability measures are also included (Gacto, Alcala, & Herrera, 2011).

The remainder of this paper is organised as follows. In Section 2, we present the background of ensemble models and decision forests. A critical literature review about credit risk ensemble models based in terms of diversity and interpretability is also conducted. In Section 3 we introduce CAFD methodology, discussing its main stages and parameters. In Section 4, we describe the empirical set up of our study, with their results in Section 5. Finally, in Section 6 we present the conclusions, limitations, and discuss the future research directions.

2. Background

2.1. Ensemble of classifiers. A diversity overview

An ensemble of classifiers is a ML paradigm generated by training a set of individual (base) classifiers for the same task, and combining their decisions using a certain fusion rule. Instead of learning one hypothesis for training data, the ensemble of classifiers produce a set of hypotheses and combine them, which lead to higher accuracy than base models (Nanni & Lumini, 2009; Paleologo, Eliseeff, &

Antonini, 2010; Wang & Ma, 2011). Such a superior performance relies on the bias-variance trade-off: a combination of forecasts exhibits a smaller error variance than any of the individual methods if base classifiers are both accurate and as diverse as possible. In particular, diversity appears as the critical point to build ensemble models (Breiman, 1996; Dietterich, 2000; Yu, Yue, Wang, & Lai, 2010). Different strategies have been proposed to induce diversity, which allow to classify ensemble methods into multiple data partition strategies and multiple learning algorithms (Kuncheva & Whitaker, 2003; Paleologo et al.; 2010; Sun & Li, 2012; Sun et al., 2014). Besides, a good combination strategy can be used to enhance the diversity of merged models (Zhou et al., 2010).

2.1.1. Data partition diversity

The most common strategy to enhance diversity is inducing multiple data partitions to develop different models to be combined from a single algorithm; both instance-partitioning methods and feature-partitioning methods can be used.

Bagging and boosting are the most widely used ensemble methods based on multiple partitions of the training instances. *Bootstrap aggregating* ('bagging') is one of the earliest ensemble learning algorithms (Breiman, 1996), also one of the most simple to implement. Diversity is obtained by using bootstrapping replicas of the training set. Each subset is used to induce a base classifier by the same learning algorithm (Breiman, Freidman, Olshen, & Stone, 1984), and predictions are combined by simple majority vote. Bagging is well-known to produce small changes in data partitions, so it requires unstable base classifiers to induce diversity enough for increasing accuracy (Sun et al., 2014). *Boosting* also induces diversity by a resampling strategy but, instead of just sub-sampling the training dataset, the method starts from a weak model and then misclassified observations are given more weight in successive iteration by a multi-stage adjustment strategy (Schapire, 1990); final prediction is obtained by weighted majority vote based on base classifiers' performance. Wang and Ma (2012) provide a critical review of bagging and boosting methods on the basis of diversity. In bagging, the only diversity factor is the different proportion of instances in the training sample, so a large number of base classifiers are needed to produce accurate results. Whilst boosting induces a higher diversity through different weights being introduced in subsequent multi-stages, the use of a single classifier algorithm reduces real variety so inclusion of alternative diversity sources is encouraged (Wang et al., 2014).

Alternatively, random subspace and rotation forests encourage diversity by inducing multiple data partitions on features instead of training instances. In *random subspace* (Ho, 1998), multiple subsets are obtained based on randomly sampled features (without replacement) from the full dataset. Each subset is used to induce a base classifier belonging to the same algorithm family, which outputs are combined using a simple majority voting scheme. Following a different strategy, *rotation forests* (Rodriguez, Kuncheva, & Alonso, 2006) induce a bootstrapping-based ensemble of base classifiers trained with different sets of PCA (principal component analysis) transformed attributes, in order to reduce feature redundancy.

2.1.2. Learning algorithm diversity

Diversity can also be induced using different algorithms trained on the same dataset, or using a single algorithm with different parameters applied to the same data.

In its simplest form, two or more different classifiers are developed independently in parallel and its output being combined to deliver a final classification decision (Dietterich, 2000). While a large number of combination functions are available for this *multi-classifier system* (Hsieh & Hung, 2010) simple majority or weighted majority vote are often favoured. To maximise accuracy, individual classifiers must be based on different theoretical concepts (Finlay, 2011; Hsieh & Hung, 2010; Hung & Chen, 2009). *Stacking* ('stacked generalisation') is a more complex, multi-stage method that combines base learners built by different learning algorithms using a high level classifier (Wolpert, 1992). In the first step, distinct learners are built on subsets of training data generated by cross-validation (Wang, Hao, Ma, & Jiang, 2011). Base classifiers are then applied to data excluded from the original training subsets, and their predictions are treated as new data. In the second-stage, a learning algorithm is applied to the new data in order to build the final meta-classifier. In spite of its accuracy, the model provides low-interpretable results due to the two-stage process (Wang et al., 2011).

Alternatively, diversity can be induced by training one algorithm with *different parameter values* (Yu et al., 2010). However, such a strategy usually does not produce enough diversity, so it is frequently combined with multiple data partition strategies and/or multi-classifier systems (Sun et al., 2014).

2.1.3. Combination functions

Diversity could also be enhanced in the ensembling stage of base classifiers: the more diverse merged learners are the higher the accuracy and the smaller the complexity of the final model (Sun et al., 2014; Zhou et al., 2010).

Different combination functions have been proposed for merging individual classifiers, such as voting methods, reliability-based measures, ranking methods, Bayesian methods, or adaptive weighting methods (Finlay, 2011; Twala, 2010; Zhou et al., 2010). Multivariate combination functions have been also proposed, even if no significant increases of accuracy are reported (Fedorova, Gilenko, & Dovzhenko, 2013; Finlay, 2011). From them, voting methods are mostly employed due to their simplicity and easy implementation (Abellan & Mantas, 2014; Finlay, 2011; Li et al., 2012; Yu et al., 2010). In the simple majority vote strategy, each classifier makes a prediction and the instance is assigned to the class with the highest number of votes; however, it does not make full use of the information for each base learner (Zhou et al., 2010). As an alternative, the weighted majority vote assigns a weight to each classifier based on its performance, such that the final class value is obtained by majority vote of the weighted decisions. However, voting methods depart on the assumption of independence of classifiers, leading to bias in presence of correlated classifiers (Hsieh & Hung, 2010; Yu, Wang, & Lai, 2008). To face it, a correlation minimization strategy should be included to select the group of classifiers to be combined. Preliminary results suggest that merging performance-weighted individual classifiers below a prespecified correlation threshold increases the diversity of model and outperforms alternative voting approaches (Zhou et al., 2010).

2.2. Ensemble of decision trees: decision forests

Ensembles of decision trees (also called decision forests -DF-) are one of the most accepted ensemble strategies (Abellan & Masegosa, 2012; Kruppa et al., 2013). Due to the simple rule-based structure of decision trees, decision forests represent promising ensemble techniques in terms of the accuracy vs. interpretability dilemma (Abellan & Mantas, 2014; Paleologo et al., 2010).

As base components of DF, decision trees (DT) are recursive acyclical models that represent rules underlying the distribution of data through a hierarchical structure in a form of a tree (Quinlan, 1993); each route connecting the root and leaves defines a logical 'if-then' decision rule that encodes data splitting criteria along several nodes. DT can be classified into two broad groups: univariate trees, which use a single feature to split data at each node; and oblique trees, which use a linear combination of attributes for data splitting at each node (Breiman et al., 1984) (Fig. 1).

DT strengths include computational efficiency in dealing with datasets with a large number of explanatory variables relative to the number of cases; mixed data types; missing data; and different relationships between variables in different parts of the measurement space (Feldman & Gross, 2005). Such characteristics are commonplace in most credit scoring datasets. Besides, DT are relatively immune to the presence of outliers (Twala, 2010), while producing rules whose semantics are clear to domain experts, (Marques, Garcia, & Sanchez, 2012b; Martens et al., 2007). But their recursive nature, which is the source of its transparency, is also a drawback since a local optimization on single variables is performed at a time (Feldman & Gross, 2005). As a result, DT have been found to be instable (Abellan & Mantas, 2014) and sensitive to noise and redundant attributes (Marques et al., 2012b).

Departing on DT drawbacks, the aim of DF is to synergistically explode the intuitive rule representation of base trees, together the superior accuracy of ensemble models, producing both accurate and comprehensible classifiers. Any ensemble strategy could be used to merge DF for classification tasks, including multiple data partition, multiple learning algorithms, and alternative combination functions (Wang et al., 2012, 2014). Since DT instability induces high diversity on ensemble models, DF could enhance final accuracy over alternative approaches (Abellan & Mantas, 2014). As a result, DF have been found to generate the highest improvement of accuracy over other ensembles of classifiers (Marques et al., 2012a; Wang et al., 2011; 2012).

Credit scoring literature has used different variants of boosting, bagging, and multiple attribute subsets to build decision forests. From them, gradient boosting (Friedman, 1999a; Friedman, 1999b) and random forests (Breiman, 2001) are the most popular DF. Even if they have yet to be fully researched in the context of credit scoring, up-to-date applications suggest, however, that while multiple data partition DF are certainly accurate, use to combine a large number of trees leading on a lack of interpretability. Even if most authors do not report the number of base learners, Kim, Kang, and Kim (2015) inform on the use of 25 classifiers; Finlay (2011) suggests at least 50 partitions to reduce the misclassification rate of bagging and boosting ensembles of decision trees; Paleologo et al. (2010) and Wang et al. (2011) merge 100 data partitions; Abellan and Masegosa (2012) merge between 100 and 500 decision trees; Kruppa et al. (2013) combine 500 decision trees; and Brown and Mues (2012) and Tsai, Hsu, and Yen (2014) considers up to 1000 individual trees for building ensembles of decision trees. Therefore, DF have missed its original interpretability advantage in spite of higher classification accuracy. Consequently, improving the interpretability of DF is considered as a much important but yet largely understudied research direction (Wang et al., 2011).

2.2.1. Gradient boosting

Stochastic gradient boosting (Friedman, 1999a; Friedman, 1999b) is an ensemble algorithm that applies one sole source of diversity (boosting-based error minimization strategy) to combine decision trees. After the initial learner is grown, different data subsets are generated so that a classifier is trained on each subset and tested by the rest of data. A weighing function is then introduced to give preference to subsets in which previous classifiers have failed. The model is built on the notion of committees of experts, in a way similar to a long expansion series, i.e., a sum of factors that become progressively more accurate as the expansion continues, as

$$F(t) = F_0 + \beta_1 C_1(t) + \cdots + \beta_M C_M(t),$$

where C_i are individual trees, β_i are coefficients for the respective trees computed by the gradient boosting algorithm, $t \in T$. For binary problems, the final score $F(t)$ is obtained through a weighted voting strategy that starts with the sample mean F_0 and then progressively adjusts it upwards or downwards, depending on the predicted response of each C_i. In order to avoid overfitting, gradient boosting requires selecting an adequate tuning of the stopping rule (number of iterations or error minimization), the maximum number of merged decision trees, and the maximum branch size used to split base classifiers.

2.2.2. Random forests

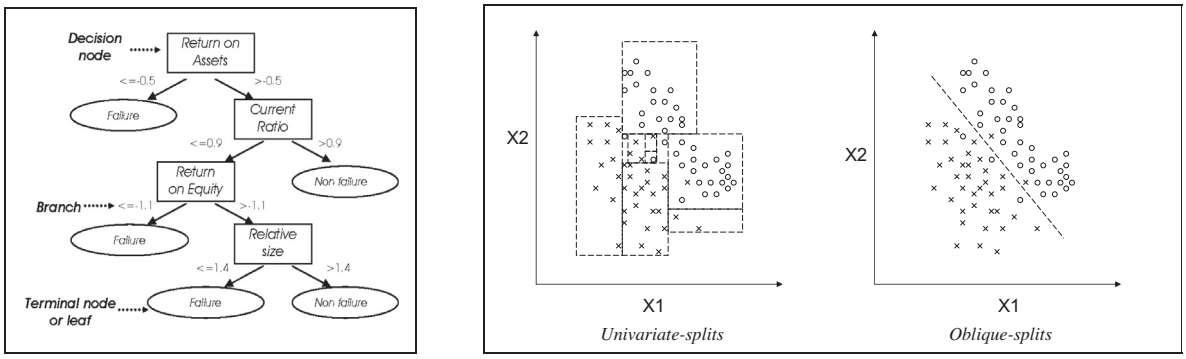
Random forests (RF) are dual diversity DF that combine bagging and random subspace feature selection to merge individual decision trees (Breiman, 2001). Randomness is explicitly introduced in two steps: First, T subsets are generated, which each subset randomly selecting N (sample size) data from the original sample. Secondly, an un-pruned tree is built from each subset using random subspace feature selection to generate splits, which reduces correlation between trees in the forest. Each tree casts a unit vote for the most popular class at point $t \in T$, with the final class obtained by majority rule. Partial interpretability options are available with RF, such as the estimation of the most representative tree, the selection of the most important variables, or the calculation of proximities between subjects (Breiman, 2001; Kruppa et al., 2013). However, RF has been observed to be ineffective in presence of a large number of irrelevant features, leading to a much large number of base learners but reduced accuracy gains (Rodriguez et al., 2006).

2.3. A critical review of the credit risk literature on ensemble strategies

In this Section, a review of recent ensemble strategies is performed, discussing methods in terms of diversity, accuracy and interpretability[2] (Appendix A). Based on such a review, we have classified ensemble approaches in three main groups: proposals based on just one source of diversity without including interpretability measures; proposals that include diverse sources of diversity but no interpretability measures; and attempts to balance accuracy and comprehensibility departing on single or multiple diversity sources.

A large body of the credit risk literature introduces just one source of diversity in building the ensemble approach, without reporting any measure of interpretability. Abellan and Masegosa (2012) use a bagging ensemble of imprecise DT (credal DT) to model both credit risk and not-credit risk datasets. Even if authors report a reduction in the average number of rules per individual tree, the large number of classifiers included in the ensemble (up to 500 DT) dramatically reduces the possibility of model's interpretability. Later on, Abellan and Mantas (2014) compare credal-based DF to alternative ensemble strategies (bagging and random subspace), reporting a higher relative accuracy but no measures of model's comprehensibility. Twala (2010) confirms the superior accuracy of 20 ensembles of classifiers over base individual models, using four credit risk datasets with different noise levels. However, no interpretability measures are reported. Wang et al. (2011) compare different sources of diversity (boosting, bagging, stacking) for building ensemble models, even if just one source is considered simultaneously. Results report the high accuracy of stacking for dealing with credit risk models, but model's complexity prevents interpretability. Tsai et al. (2014) compare ensembles of DT by boosting with other base classifiers, reporting accuracy gains that are larger as far as the number of embedded DT is incremented. While ensembles of DT are found to produce the lowest computational cost, no interpretability measures are reported. More recently, Geng, Bose, and Chen (2015) use a multi classifier system to combine outputs from

[2] See Nanni and Lumini (2009) for earlier references of ensemble strategies in bankruptcy and credit scoring fields.



(a) Decision tree (univariate) (b) Univariate and oblique splits

Fig. 1. Structure of decision trees.

three different classifiers (NN, SVM, DT). Results report slightly, not-generalised accuracy gains with respect to NN. No interpretability measures are provided, even if authors confirm that introducing DT in the final model provides somewhat comprehensibility to financial prediction. Besides, Kim et al. (2015) propose a variant of SVM to build boosting ensemble models in presence of imbalanced datasets. While authors do not analyse it explicitly, they introduce a geometric accuracy measure that seems to induce high diversity by overweighting minority class instances. Result suggests the superiority of the ensemble proposals, but no interpretability analysis is performed.

Another body of literature has begun to combine multiple sources of diversity for synergistic accuracy gains, even if no interpretability issues are considered. Yu et al. (2010) develop a multiagent SVM ensemble, and analyse the effect of two diversity sources (bagging and parameter diversity) on its generalisation performance. Alternative combination functions are also analysed, including majority-vote and nonlinear NN-based functions. Results support the dominance of the SVM NN-based merged model over individual classifiers, so the superiority of multi-diversity strategies over simpler models. Similarly, Zhou et al. (2010) generate diverse least square SVM ensembles by mixing data-partition and parameter diversity sources. Besides, a correlation-based penalty function is introduced to maintain a high diversity of the classifiers. Results suggest that merging a reduced number of diverse base learners, while cannot always achieve better performance than alternative models, do as good as the best with 5% level of confidence. However, model simplicity is not explored in terms of interpretability. Finlay (2011) investigates different ensemble strategies including boosting, error trimmed boosting, bagging, multi-classifier system, and local accuracy dynamic systems. While just one source of diversity is introduced, authors analyse the potential of including multivariate combination functions. Results confirm the superiority of the highest diversity strategies; error trimmed boosting and multi-classifier systems with multivariate combinational functions outperform alternative models. However, no interpretability measures are introduced even if author recognises their practical usefulness. Sun and Li (2012) develop one of the first proposals that explicitly introduce several sources of diversity in a SVM ensemble: different algorithm parameters, different feature subsets, and a weighted majority voting combination rule. As a result of high diversity, authors report accuracy gains with a reduced number of base learners (between 2 and 15), but no attempts are done to improve model's interpretability. Considering SVM as base models, Wang and Ma (2012) combine two diversity sources, bagging and random subspace, to produce the ensemble of SVM classifiers. Using enterprise credit risk datasets, authors find their proposal to be more accurate than any alternative model, but no interpretability measures are provides. Wang et al. (2012) use the same dual strategy to build ensembles of DT; their findings corroborate the superior accuracy of the multiple-diversity ensembles. Li et al. (2012) introduce relevant vector machines (RVM) as base classifiers of a boosting-ensemble approach. Besides, soft margin boosting is introduced to increase diversity when assembling base models. Diversity is thirdly increased by using different parameters in base model learning. Results suggest the superior accuracy of the RVM-soft margin boosting approach over other classifiers. Whilst authors report up to 200 base models being assembled, no information is provided to interpret model's decisions. Marques et al. (2012a) perform an extensive research that evaluates the performance of seven base models when used as members of five ensemble methods (including multiple data partition on instances and features). Besides, Marques et al. (2012b) propose two-level ensemble methods that explicitly introduce dual diversity by combining data resampling and feature selection methods. Results confirm that the two-level ensemble strategies outperform single ensembles, even if no interpretability concerns are discussed. Brown and Mues (2012) compare gradient boosting and random forests ensembles of DT with individual classifiers for dealing with data imbalance. As expected, results confirm the highest performance of random forest as a dual-diversity method, but no interpretability measures are reported. Kruppa et al. (2013) confirm that random forests of probabilistic DT consistently outperform alternative individual classifiers. While authors recognise the importance of interpretability, the final model includes 500 DT so no logical rules are obtained but estimates of feature importance. In this line, Wang et al. (2014) present a proposal that integrates two sources of diversity into the ensemble of decision trees: boosting and feature selection. Even if a higher performance is reported, no interpretability measures are generated.

Finally, a reduced number of references have tried to reconcile interpretability with strong classification accuracy in credit scoring. Hsieh and Hung (2010) proposal is one of the firsts that deal with balance between accuracy and interpretability in ensemble models. Authors propose a multi-stage model that hierarchically combines two sources of diversity, bagging and multi-classifier systems. Diversity is also enhanced through feature selection for reducing correlation between variables; clustering techniques for refining the quality of samples; and confidence-weighted voting for combining base learners. Results confirm a high performance, but no comparisons are performed with alternative models to report accuracy gains. Authors also use base Bayesian networks classifiers to extract partial decision rules on the model's decision process. Paleologo et al. (2010) compare a variant of bagging ensembles (subagging) with single models based on alternative classifiers. Results confirm that ensembles of DT provide a superior accuracy while keeping the model simple and somewhat comprehensible. Authors report the

Table 1
Diversity sources comparison in decision forests: gradient boosting, random forests, and CADF.

	Diversity sources			
	Diversity on training sets	Diversity on attribute subsets	Diversity on learning strategy	Ensemble combination function
Gradient boosting	Boosting	–	One single learning method (CART univariate)	Weighted majority voting
Random forests	Bagging	Random subspace	One single learning method (CART univariate)	Simple majority voting
CADF	Five ensemble classifiers trained on the same dataset. 10-fold cross validation used for tree size selection. Bootstrapping subsamples used for out-of-sample estimates	Upto five split criteria (univariate and oblique splits)	Static and parallel approach (five DT)	Correlated-based weighted majority voting

average depth of each DT, and a relatively reduced number of base classifiers (between 20 and 50). However, no rules are extracted but simple scores of variable importance based on the frequency of use. De Bock and Van den Poel (2012) combine boosting and random subspace methods to increase diversity in an ensemble of generalised additive models. Results suggest the highest accuracy of the proposed method compared to alternative methods but random forests. Their proposal also produces importance scores and average trends for predictive features, even if no decision rules are obtained. Besides, Fedorova et al. (2013) propose a multi-stage ensemble method where boosting is applied as a combination function instead of a data partition method. Individual classifiers are used to select features being used for building boosting ensembles of NN. While results report some accuracy gains, interpretability is limited to the identification of important variables. Finally, Tomczak and Zieba (2015) use a variant of Boltzmann Machines to produce a simple scoring table based on relevancy weights of binarised-feature inputs. While authors do not develop an ensemble proposal, they compare their method with individual classifiers and ensembles of DT (boosting, bagging, random forests). Results provide some evidence on the superiority of the new method but, much interestingly, explicitly report the superior interpretability of the generated scoring table in front of more complex models.

3. The proposed correlated-adjusted decision forest (CADF). A methodological approach

In this section, we present the correlated-adjusted decision forest (CADF) proposal, which tries to optimise the balance between accuracy and interpretability of ensemble strategies by enhancing diversity. Instead of exploiting one source of diversity, CADF includes four complementary strategies to guarantee that merged DT are both accurate and mutually low-correlated (Table 1): learning algorithm diversity to induce base learners; feature diversity to select relevant attributes; instance-diversity to build accurate individual classifiers and estimate generalisation performance; and combination function diversity to merge classifiers. Such diversity strategies are reflected in the architecture of the proposed framework, which consists of three levels (Fig. 2). Firstly, bootstrap training subsets are obtained in order to estimate a post-distribution of model parameters, also to compute out-of-sample accuracy. Secondly, five different inductive algorithms are used to bring diversity on both attribute selection and base learning strategies. Finally, individual results are merged into a final output using a pseudo-R^2 penalty function that produces adjusted-weighted votes through a mixed accuracy-correlation ranking scheme.

3.1. Data partition

CADF uses a multi-classifier strategy, where all members are diverse in nature but trained on the same dataset. Since DT are over-sensitive to redundant attributes and noise, a 10-fold cross validation approach is included in the training process for optimal architecture selection (Hsieh & Hung, 2010): while each model is trained on the same dataset, cross-validation data are used to establish the optimal tree size.

Due to data scarcity applies to many credit risk applications, the full dataset is not split into disjoint training (in-sample) and test (out-of-sample) sets. Instead, a .632 bootstrapping strategy is used for learning and testing purposes (Efron & Tibshirani, 1995). Let $\hat{F}$ be the empirical distribution function for the full sample T with mass $1/n$ on $t_1,\ldots,t_n$; let T^* be a random sample of size n taken iid with replacement from $\hat{F}$, where t_i^* is a single random observation, $t_i^* = (x_i, y_i)$. True error is estimated through independent bootstrap training sets $T_1^*,\ldots,T_B^*$; for each T_b^*, a prediction model is built $\hat{f}[T_b^*, x_i]$ and the expected .632 bootstrapping true error rate is obtained as,

$$\begin{aligned}\hat{Err}_{.632E} &= 0.368 \times \overline{err} + 0.632 \times E_0 \\ &= 0.368 \times \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{f}[T_n, x_i]| + 0.632 \\ &\quad \times \frac{\sum_{b=1}^{B}\sum_{A_B}|y_i - \hat{f}[T_b^*, x_i]|}{\sum_B |A_B|},\end{aligned}$$

where $\overline{err}$ is the error on the full sample (resubstitution error), E_0 is the bootstrapping out-of-sample average error, and $A_b = \{i | p_i^* = 0\}$ is the number of training vectors not included in the bth bootstrap sample, also called boostrapping test vectors; $B = 50$ are often sufficient to achieve robust results (Efron & Tibshirani, 1993). Similarly, the .632 AUC estimate can be obtained as follows (Hanczar et al., 2010),

$$AUC_{.632} = 0.368 \times \overline{AUC} + 0.632 \times AUC_0.$$

Departing on bootstrapping error estimates, nonparametric percentile intervals are built as,

$$\left[\hat{E}_{0\%low}, \hat{E}_{0\%up}\right] \approx \left[\hat{E}_{0B}^{(\alpha)}, \hat{E}_{0B}^{(1-\alpha)}\right],$$

where $\hat{E}_{0B}^{(\alpha)}$ is the $100 \cdot \alpha$th empirical percentile of the error values, that is, the $B \cdot \alpha$th value in the ordered list of B replications of $\hat{E}_0$.

3.2. Base classifiers

Diverse models with much disagreement are more likely to achieve a good generalisation accuracy in terms of the principle of bias-variance trade-off (Yu et al., 2010). For DT, several inductive algorithms have been investigated, departing on different parameters, architectures, pruning, and stopping criteria. Each algorithm includes its particular splitting rule for attribute selection, inducing diversity by wrapper feature selection.[3] This is crucial since the

[3] Generally, two types of featuring methods are available: filter and wrapper. While filter strategies carry out feature selection as an independent process of the base classifiers, wrapper strategies consider particular characteristics of each model (Sun et al., 2014).

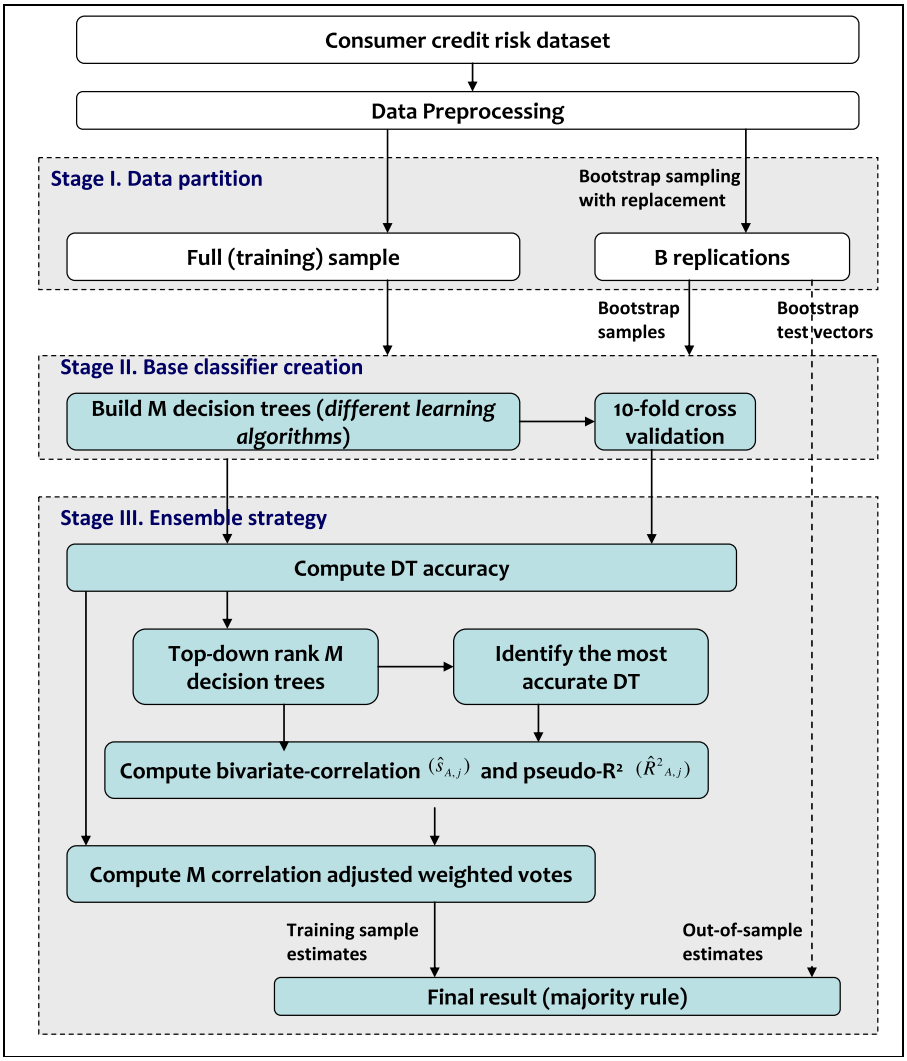


Fig. 2. Stages of CAFD proposal.

choice of feature subsets influences the selection of an appropriate classifier and vice versa (Hsieh & Hung, 2010; Sun et al., 2014). In CADF, five different DT algorithms are used as base classifiers: ChAID, Assistant, C4.5, CART univariate, and CART oblique.

Chi-2 or ChAID algorithm (Kass, 1980) develops a sequential process in three states, merging (combination of classes), splitting (node partition), and stopping (user-given value). At each node, the splitting rule selects the feature with the smallest p value using a Chi-square test.

Assistant (Cestnik, Kononenko, & Bratko, 1987) is a constructive algorithm based on Quinlań's ID3 (Quinlan, 1979) that includes the binarization of attributes, decision tree pruning, handling of incomplete data, and use of a naïve Bayesian classifier. The splitting rule is the 'gain criterion' based on the concept of information entropy, so that attribute x with the highest information gain is used to split data,

$$gain(x) = \text{info}(T) - \text{inf}\, o_x(T),$$

$$\text{inf}\, o(T) = -\sum_{k=1}^{K} p_{j.} \log_2(p_{j.}), \quad \text{inf}\, ox(T) = \sum_{l=1}^{L} \frac{|T_l|}{|T|} \times \text{info}(T_i),$$

where K is the number of classes, p_j is the proportion of the jth class in the (sub)tree T, L is the number of subtrees generated if x is used to split the node (L being the number of attributes of x), $|T_l|$ is the number of examples classified by subtree T_l.

C4.5 (Quinlan, 1993) is a descendant and improved version of ID3 algorithm that considers two univariate split criteria for each node, the previous 'gain criterion' and the 'gain ratio criterion' as,

$$gain\_ratio(x) = gain(x) \Big/ \left(-\sum_{l=1}^{L} \frac{|T_l|}{|T|} \times \log_2\left(\frac{|T_l|}{|T|}\right)\right),$$

where 'gain ratio criterion' is a normalised variant of the gain criterion that avoids bias in favour of high-dimensional attributes. C4.5 uses a stopping rule based on the Chi-square test, and a pessimistic error-based pruning criterion to reduce the size of the tree.

CART (Classification and Regression Trees) (Breiman et al., 1984) uses two binary split criteria, the 'Gini' criterion (or Gini index) and the 'Twoing' index, based on Bayesian risk. The Gini criterion is the most broadly used rule, based on searching for the attribute and splitting value that solve the following problem,

$$\underset{x_j \leqslant x_j^R,\, j=1,\ldots,M}{\arg\max} \left[-\sum_{k=1}^{K} p^2(k|t_p) + P_l \sum_{k=1}^{K} p^2(k|t_l) + P_r \sum_{k=1}^{K} p^2(k|t_r) \right],$$

where $x_j \leqslant x_j^R, j = 1..,M$ is any of the possible splits, $p(k|t)$ is the conditional probability of class k provided we are in node t; t_p, t_l, t_r are parent, left, and right nodes, respectively; P_l, P_r are probabilities of right and left nodes. Even if Gini criterion works well for noisy data, it could generate quite imbalanced trees. As an alternative, the 'Twoing' index searches for the attribute that generates two 'superclasses' with maximum separation (among 2^{K-1} possible splits) as,

$$\arg\max_{x_j \leqslant x_j^R,\ j=1,\ldots,M} \frac{P_l P_r}{4}\left[\sum_{k=1}^{K}|p(k|t_l)-p(k|t_r)|\right]^2.$$

Though there are just slight differences between the Gini and Twoing-based trees, Gini splits appear to perform better and faster (Breiman et al., 1984). A pruning strategy is also included to reduce over fitting.

3.3. Combination functions

Depended upon the work done in previous steps, a set of base DT are generated, each one outputting its own predicted class for a specified problem. The subsequent task is to combine these base learners into an aggregated classifier using an appropriate ensemble strategy. In order to enhance diversity among the models to be combined, CAFD introduces a modified weighted voting strategy that penalises high-correlated decision trees. This strategy follows Hsieh and Hung (2010) suggestion on using a hierarchical combination of heterogeneous classifiers to enhance accuracy.

The pseudo-R^2 correlation estimate is obtained based on the volume of overlapping regions labelled with the same class by any pair of classifiers (Ho, 1998). Given two classifiers, C_i and C_j, let their class decisions for a point $t \in T$ be $c_i(t)$ and $c_j(t)$, respectively. Considering a set of N observations and assuming a uniform distribution, the estimated correlation ($\hat{s}_{i,j}$) can be obtained as,

$$\hat{s}_{i,j} = \frac{1}{N}\sum_{k=1}^{N} f(t_k),\quad f(t_k) = \begin{cases} 1 & \text{if } c_i(t_k) = c_j(t_k) \\ 0 & \text{otherwise} \end{cases}.$$

Similarly, for a group of M decision trees ($C_1, \ldots, C_M$) being top-down ranked in terms of their accuracy measured on the same training sample ($p_1, \ldots, p_M$), a correlation estimate could be obtained between the most accurate tree ($C_A = C_1$) and any other classifier as,

$$\hat{s}_{A,j} = \frac{1}{N}\sum_{k=1}^{N} f(t_k),\quad f(t_k) = \begin{cases} 1 & \text{if } c_A(t_k) = c_j(t_k) \\ 0 & \text{otherwise} \end{cases}.$$

Subsequently, the pseudo-square correlation coefficient (pseudo-R^2) can be obtained as $(\hat{s}_{A,j})^2 = \hat{R}^2_{A,j}$, which provides a measure of how well outcomes of the most accurate decision tree are replicated by successive combined trees, as the proportion of predicted classes of C_A explained by C_i.

Finally, an adjusted square-correlation (R^2) weighted vote is computed for the jth class on point $t \in T$ as,

$$CADFw_j = \sum_{i=1}^{M} p_i[1-(\hat{s}_{A,i})^2]\cdot f(t)_i,\quad f(t)_i = \begin{cases} 1 & \text{if } c_i(t) = j \\ 0 & \text{if } c_i(t) \neq j \end{cases}.$$

The final output is obtained by majority rule. As observed, CADF merges classes predicted by individual DT, instead of individual scores. This strategy introduces a noise component to avoid overfitting, which is consistent with prevalent decision forests strategies (Breiman, 2001; Friedman, 1999a; Friedman, 1999b). Since the adjusted weighted vote scheme overweighs accurate and independent classifiers, it is expected to outperform vote methods while maintaining high transparency of results.

4. Empirical set up

4.1. Data set characteristics

In evaluating the performance of CADF, a real-life dataset on a major German financial institution is used.[4] The German dataset includes 1000 examples of good payers (700) and bad payers (300) of consumer loans, characterised by a set of 20 risk drivers. The German dataset has been widely used to compare performance of alternative credit scoring models, since it is imbalanced and contains many categorical variables together heterokedastic, non-normal, high-correlated continuous features.

4.2. Parameter tuning and input selection

Departing on previous dataset, we analysed the accuracy and comprehensibility of the CADF proposal compared to eleven individual classifiers and two alternative DF. Before techniques were run, dummy variables were created for the categorical attributes. The LDA and LR do not require tuning; preliminary feature selection was undertaken using step-wise (forward) regression based on Wilks's partial lambda ($p = 0.05$). For k-nn algorithm, $k = 7$ was selected by experimental procedure, with k included in the $[2, 2\cdot\sqrt{N}]$ interval as a rule of thumb. Both lineal SVM and 2-degree polynomial SVM were run; parameter setting included a tolerance for accuracy of 0.001 (minimum relative improvement at each step), and a C value of 1 that control margin failures. NN architecture included one hidden output with logistic activation function and 2 hidden neurons; the best performing number of neurons was established based on 10-fold cross validation samples to prevent overfitting.

Different DT were considered: ChAID including Bonferroni adjustment ($p = 0.00001$ for decision nodes, $p = 0.05$ for leaves); Assistant with a 'gain criterion' splitting rule, including a pre-pruning strategy; (iii) C4.5 with a 'gain ratio criterion' splitting rule, pruning strategy of 25% pessimistic error rate; (iv) CART-univariate with a 'Gini index' splitting criterion, pruning based on re-sampling; and (v) CART-oblique with similar parameters than the univariate variant.

About decision forests, key parameters include forest size (number of trees), the level of randomness, and the maximum size of individual trees (number of nodes). In random forests, the best performing number of trees was automatically selected based on the out-of-sample error rate from bagging samples; different settings were tested for defining the number of random attributes per tree. About gradient boosting, a maximum number of 400 trees and 6-nodes per tree were established; the best number of trees was obtained using 10-fold cross validation samples. CADF was built on univariate DT (ChAID, Assistant, C4.5, CART univariate) and mixed DT (univariate DT plus CART oblique). Different combination functions were used, including the adjusted-R^2 weighted vote, but also simple majority vote (SVDF) and weighted majority vote (WVDF) decision forests, for benchmarking purposes.

4.3. Accuracy measures

Predictive accuracy is tested through both the accuracy rate (AR), and the area under the receiver operation characteristic curve (AUC) (Baesens et al., 2003; Crook et al., 2007; Hand, 2009). AR is a discriminatory measure that informs on the percent of correctly classified observations, considering a cutoff-value (S_c) used to partition the score range into those to be labelled good payers (non-default) and those to be labelled bad payers (default),

$$\mathrm{AR} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}},$$

where TP is the number of true positives (good predicted as good), TN is true negatives (bad predicted as bad), FP is false positives (bad predicted as good), FN is false negatives (good predicted as bad).

Departing on AR, the error rate is defined as (1-AR). A weakness of AR is that it ignores the cost of different error types (bad classified as good, or vice versa), which usually are much different in credit

[4] Publicly available at the UCI Machine Learning repository (http://archive.ics.uci.edu/ml/).

scoring applications. This is the reason why it becomes especially interesting to measure the error on each individual class (Wang & Ma, 2012; Wang et al., 2012, 2014; 2014): type I error (bad classified as good) = $\frac{FP}{TN+FP}$; and type II error (good classified as bad) = $\frac{FN}{TP+FN}$.

The AUC is an alternative discrimination power statistics that does not depend of class distribution or misclassification costs. The receiver operating characteristic curve (ROC) is a plot of the proportion of the true positive rate (sensitivity) against the false positive rate (1-specificity) at all values of S_c, where sensitivity = $\frac{TP}{TP+FN}$; and specificity = $\frac{TN}{TN+FN}$.

The AUC statistic represents the area under the ROC curve, moving between 0 (no discriminant power) and 1 (perfect fit). While this measure does not depend on cuttoff selection, it represents an average statistics that considers cutoffs that the lender could be, for operational reasons, not interested in.

Since each measure has its merits and limitations, a combination of them is preferred rather than a single one (Wang et al., 2012, 2014; 2014). In this study we use AR with a 0.5 cutoff-point (Nanni & Lumini, 2009; Wang & Ma, 2012), type I error, type II error, and AUC. In order to estimate the out-of-sample error rate, a .632E bootstrapping is used, with $B = 50$ replications representing a good trade-off between statistical robustness and practical estimation purposes (Efron & Tibshirani, 1993).

4.4. Interpretability measures

Whereas the measures of accuracy are well-known, interpretability measures are difficult to define since depends on multiple factors as model structure, type of features, or the own subjectivity of the concept of comprehension. Focusing on ensemble models, researches have tried to overcome the lack of interpretability by developing some variable importance measures **(Breiman, 2001; De Bock & Van den Poel, 2012; Kruppa et al., 2013). However, these indexes are focused on feature selection instead of internal knowledge, which restricts its validity for model comprehensibility.

Traditional measures of interpretability are very simple, considering aspects as the mean number of rules or rule conditions. More advanced interpretability measures have been developed for fuzzy rule-based systems. Gacto et al. (2011) present an overview of the proposed interpretability measures and propose a taxonomy for linguistic rules with four quadrants (complexity-based or semantic-based interpretability at the rule-base or variable partition level). Authors state that there is not a single global measure to quantify the interpretability of decision models, and propose to combine measures from each quadrant, such as: number of rules; number of features; number of rules fired at the same moment; and some measures of distinguishability on variable partitions. In this paper, we follow their suggestions to assess DF interpretability using both complexity-based and semantic-based measures.

4.5. Statistical comparison of classifiers

Several statistical tests are used to compare the accuracy of classifiers (Demšar, 2006). First of all, the McNemar non-parametric test for two related samples is conducted taking the 2×2 accuracy contingency table for the full dataset (Yu et al., 2010). A non-parametric Wilcoxon signed rank test is then applied to test if the AR/type I error/type II error of CADF was the same that AR/type I error/type II error of alternative models (Huang et al., 2007), using out-of-sample bootstrapping vectors (A_b).

DeLong test is used to compare the AUC of CADF against alternative classifiers **(DeLong, DeLong, & Clarke-Pearson, 1988), considering the full dataset. Besides, Friedman test (Friedman, 1940) is used to compare out-of-sample AUC on bootstrapping test vectors ($D = 50$) for alternative models (Brown & Mues, 2012; Marques et al., 2012b), as follows:

$$\chi_F^2 = \frac{12D}{K(K+1)}\left[\sum_{k=1}^{'K} AvR_j^2 - \frac{K(K+1)^2}{4}\right],$$

where $AvR_j = \frac{1}{D}\sum_{i=1}^{D} r_i^j$, D is the number of data sets, K is the number of classifiers, and r_i^j is the rank of the jth classifier on the ith dataset; if D and K are big enough the statistics is distributed according to χ_F^2 with $K-1$ degrees of freedom ($D > 10$, $K > 5$ as a rule of thumb). If the null-hypothesis of the Friedman test (all models are equivalent) is rejected, the post hoc Nemenyi test is applied to find the particular pairwise comparison that produce significant differences in out-of-sample bootstrapping vectors. Nemenyi test states that the performances of two classifiers are significantly different if their average ranks differ by at lest a critical difference (CD), given by

$$CD = q_{\alpha,\infty,K}\sqrt{\frac{K(K+1)}{12D}},$$

where the value $q_{\alpha,\infty,K}$ is based on the studentised range statistics (Nemenyi, 1963). To compare alternative models with a particular control classifier (CADF), CD with Bonferroni–Dunn (B-D) adjustment is preferred (Demšar, 2006), where CD is computed as previously but compared to alternative critical values.

5. Results and discussion

Table 2 reports AUC, AR, type I error, and type II error of classifiers for the full German dataset. Generalisation (test) accuracy is also provided, in order to assess robustness of credit models to fit external, independent observations; both.632E error rates and $AUC_{.632}$ are computed, including 95% confidence intervals. As an initial pre-condition, we tested the stability and overfitting reluctance of classifiers (Li et al., 2012). Comparisons between training and test performance suggest that knn and Assistant are largely overfitted, while LDA, LR, ChAID and SVM are the less biased classifiers (Fig. 3). Any decision forest is fairly robust, particularly in terms of AUC.

In Table 2, both the individual model and decision forest achieving the highest out-of-sample AUC and AR (respectively lowest type I and type II errors) are underlined. Based on the above results, initial findings are drawn in the following,

(a) For the individual models, it is obvious that CART oblique, a transition model between univariate DT and decision forests, achieves the highest AUC and AR (0.732, 75.3%), with an acceptable balance between type I and type II errors; however, it is a poorly interpretable model since some splits linearly combine up to five attributes. Closely following CART oblique is CART univariate with an AUC of 0.730, AR of 74.5%. CART univariate generates 10 easily interpretable rules from six variables, which highlight the significance of good checking accounts, and the different impact of credit amount in short-term and long term credit risk decisions. An accuracy AUC and AR rank of the rest of DT includes Assistant (0.725, 74.1%), C4.5 (0.714, 73.9%), and ChAID (0.709, 71.6%). About interpretability, ChAID obtains five rules from three variables, which are simple to understand: Clients are preferred if checking account is high and credit amount is reduced, and could also have accounts or credit in other banks. C4.5 generates 11 rules from 6 variables, which positively highlight high checking and saving accounts, short-term credits, and employed applicants. Assistant generates 15 rules from 8 features, being able to reduce the most type I error. Checking accounts, short-term credits, reduced credit amounts, and

Table 2
Empirical results. Accuracy measures.

	Number of rules/trees	Full sample				Bootstrapping test vectors (E_0)				Out-of-sample generalisation ($\hat{E}_{.632E}$) (confidence intervals, $\alpha = 0.05$)			
		AR	Type I error	Type II error	AUC	AR (1-E_0)	Type I error E_0	Type II error E_0	AUC E_0	$AR_{.632E}$ (1-$\hat{E}_{.632E}$)	Type I error	Type II error	$AUC_{.632}$
LDA	–	0.750	0.583	0.107	0.762	0.719	0.681	0.109	0.691	0.730 (0.710, 0.751)	0.645 (0.551, 0.740)	0.108 (0.074, 0.142)	0.717
LR	–	0.746	0.597	0.107	0.769	0.718	0.689	0.107	0.685	0.729 (0.707, 0.750)	0.655 (0.562, 0.748)	0.107 (0.074, 0.140)	0.716
k-nn	–	0.794	0.463	0.096	0.899	0.642	0.643	0.235	0.594	0.698 (0.630, 0.766)	0.577 (0.479, 0.674)	0.184 (0.116, 0.252)	0.706
SVM (linear)	–	0.743	0.610	0.106	0.765	0.712	0.856	0.044	0.681	0.723 (0.702, 0.745)	0.766 (0.596, 0.935)	0.067 (0.012, 0.121)	0.712
SVM (2-polyn.)	–	0.752	0.677	0.064	0.761	0.719	0.750	0.079	0.687	0.731 (0.711, 0.752)	0.723 (0.616, 0.829)	0.074 (0.036, 0.111)	0.714
NN	–	0.776	0.464	0.120	0.784	0.706	0.599	0.164	0.683	0.732 (0.696, 0.767)	0.550 (0.430, 0.671)	0.148 (0.081, 0.215)	0.720
ChAID	5 rules	0.739	0.570	0.129	0.773	0.702	0.713	0.122	0.672	0.716 (0.667, 0.736)	0.660 (0.457, 0.842)	0.125 (0.047, 0.211)	0.709
Assistant	15 rules	0.793	0.360	0.141	0.800	0.711	0.539	0.183	0.681	0.741 (0.705, 0.776)	0.473 (0.390, 0.534	0.168 (0.128, 0.209)	0.725
C4.5	11 rules	0.776	0.533	0.091	0.767	0.717	0.622	0.138	0.683	0.739 (0.713, 0.760)	0.589 (0.501, 0.678)	0.121 (0.076, 0.167)	0.714
CART-univariate	10 rules	0.787	0.470	0.103	0.802	0.721	0.524	0.175	0.688	0.745 (0.713, 0.769)	0.504 (0.343, 0.596)	0.149 (0.099, 0.213)	0.730
CART-oblique	8 rules	0.805	0.513	0.059	0.793	0.723	0.517	0.175	0.696	0.753 (0.720, 0.775)	0.516 (0.359, 0.612)	0.132 (0.082, 0.197)	0.732
Gradient boosting	113 trees	0.796	0.433	0.106	0.843	0.741[a]	0.578	0.122	0.772	0.763 (0.705, 0.787)	0.525 (0.418, 0.642)	0.114 (0.083, 0.197)	0.798
Random forests	68 trees	0.793	0.623	0.029	0.852	0.732[b]	0.718	0.080	0.738	0.754 (0.709, 0.777)	0.683 (0.567, 0.789)	0.061 (0.026, 0.109)	0.780
SVDF univariate	5 + 15 + 11 + 10 trees	0.782	0.537	0.081	0.850	0.728	0.720	0.077	0.742	0.748 (0.706, 0.776)	0.653 (0.478, 0.672)	0.078 (0.065, 0.154)	0.782
WVDF univariate	5 + 15 + 11 + 10 trees	0.794	0.460	0.097	0.851	0.734	0.594	0.123	0.745	0.756 (0.710, 0.780)	0.545 (0.450, 0.643)	0.113 (0.071, 0.160)	0.784
CADF univariate	5 + 15 + 11 + 10 trees	0.797	0.430	0.106	0.852	0.736	0.585	0.124	0.752	0.758 (0.711, 0.781)	0.528 (0.439, 0.632)	0.117 (0.074, 0.164)	0.789
SVDF mixed	5 + 10 + 11 + 10 + 8 trees	0.805	0.440	0.090	0.850	0.734	0.596	0.122	0.752	0.760 (0.717, 0.788)	0.539 (0.446, 0.626)	0.110 (0.072, 0.165)	0.788
WVDF mixed	5 + 10 + 11 + 10 + 8 trees	0.805	0.440	0.090	0.851	0.734	0.596	0.122	0.753	0.760 (0.717, 0.788)	0.539 (0.446, 0.626)	0.110 (0.072, 0.165)	0.789
CADF mixed	5 + 10 + 11 + 10 + 8 trees	0.823	0.486	0.044	0.850	0.744	0.554	0.124	0.755	0.774 (0.735, 0.798)	0.529 (0.443, 0.592)	0.095 (0.060, 0.138)	0.790

[a] 51 decision trees in average (bootstrapping samples).
[b] 40 decision trees in average (bootstrapping samples).

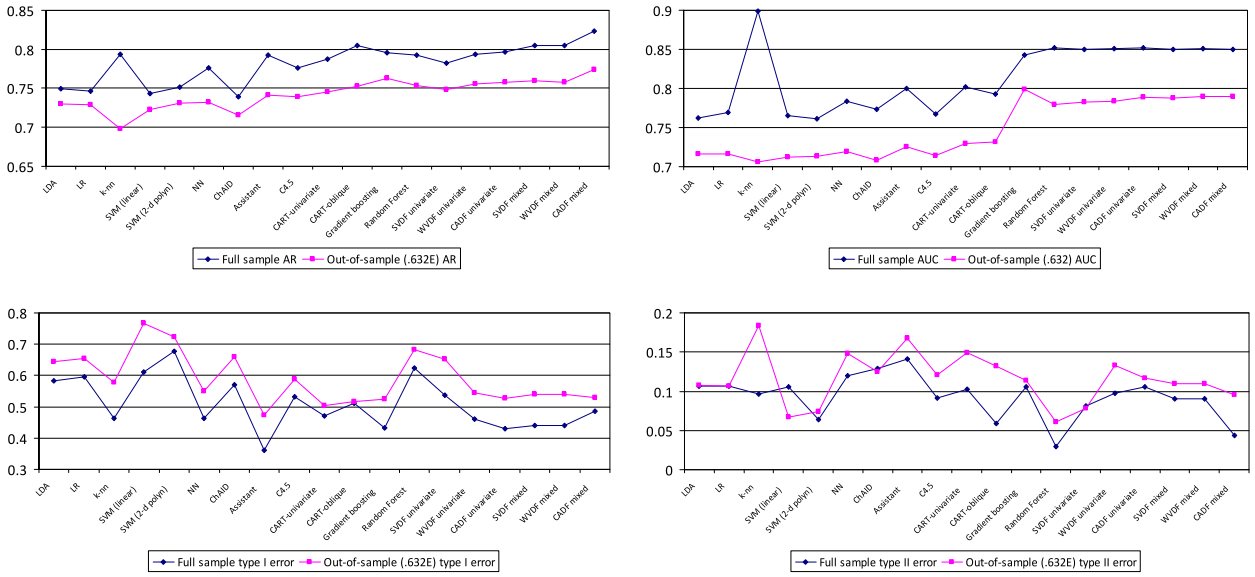


Fig. 3. Credit scoring models: accuracy comparison.

Table 3
Empirical results. Interpretability of decision forests.

	N. of rules	N. of rules fired each time	N. of features	Distinguishability of variable partitions	
				Univariate rules vs. oblique rules	Unweighted vs. weighted variables (into rules)
Gradient boosting	[min, max][a] [226, 678]	113	20	100% vs. 0%	100% vs. 0%
Random forests	[min, max][a] [136, 408]	68	20	100% vs. 0%	100% vs. 0%
SVDF/WVDF/CADF univariate	41	4	9	100% vs. 0%	100% vs. 0%
SVDF/WVDF/CADF mixed	49	5	14	83.66% vs. 16.33%	70.09% vs. 29.91%

[a] Number of rules is not available for gradient boosting and random forests. Maximum and minimum values are provided in brackets (considering a minimum of 2 nodes per tree, and a maximum of 6 nodes per tree).

longstanding employs are preferred in this model; in presence of long-term credits, a good savings account position is required. Also, the purchase of a new car is preferred to other credit purposes, while the existence of critical accounts in other banks is also evaluated.

(b) The most commonly used credit scoring techniques LDA (0.717, 73.0%) and LR (0.716, 72.9%) are fairly close to DT results, but a more imbalanced error distribution is obtained. Besides, they are quite competitive with the more complex ML techniques as NN (0.720, 73.2%), SVM 2-degree polynomial (0.714, 0.731), and SVM linear (0.712, 72.3%); last techniques also obtain very imbalanced errors, which suggest models being specialised in type II error minimization. These results are in line with Baesens et al. (2003) and Brown and Mues (2012) findings, who observe just slight accuracy increases in advanced ML techniques with respect to statistical models. The most simple knn approach obtains the worst accuracy results of any classifier (0.706, 69.8%), since overfitting dramatically increases type II error.

(c) About DF, any of them improve the prediction accuracy of a single classifier, in line with previous literature (Brown & Mues, 2012; Li et al., 2012; Marques et al., 2012a; Marques et al., 2012b). The technique achieving the highest AUC is gradient boosting (0.798), being closely followed by CADF mixed (0.790) and CADF univariate (0.789). AUC of random forests is lower than alternative DF even if it includes two sources of diversity. Results slightly differ in terms of AR, where CADF mixed provides the best accuracy across the eight decision forests (77.4%), followed by gradient boosting (76.3%) and CADF univariate (75.8%). Moreover, the adjusted-R^2 weighted vote outperforms simple and weighted voting DF, particularly in terms of AR and type I error. The reason that DF get higher average accuracy could be to reduce type I error or type II error: As shown in Fig. 3, DF get higher accuracy than CART univariate and CART oblique due to reductions in type II error, even if type I errors are similar and could be even increased. Such results are consistent with Wang et al. (2012) and Marques et al. (2012a), and suggest a highest decrease in the error associated with majority class which should be deeply analysed. Opposite, CADF are visibly the forests with a more balanced error distribution.

(d) As shown in Table 3, interpretability measures largely differ between decision forests. DF based on multiple-data partition generate a huge number of rules (over 100), acting as black-box models that can not be directly interpreted. Gradient boosting and random forests combine 113 and 68 trees respectively,[5] which are fired for each example, including up to 20 features each. CADF mixed generates a lower number of rules (49) departing on 14 different features, but just 5 of them are fired each time. Since this model introduces oblique rules based on linear combination of weighted variables, distinguishability of variable partitions is greatly

[5] The lower number of rules of random forests compared to gradient boosting could explain its minor relative accuracy: the dual source of diversity has produced a slightly higher error but a significant less complex ensemble model than gradient boosting.

Table 4
Statistical model comparisons.

		Full dataset		Out-of-sample boostrapping vectors			
		McNemar paired test (p value)	DeLong test (p value)	Wilcoxon test			Friedman test = 216.95 (0.000)*** Nemenyi B-D test CD (B-D) (α = 0.01) = 3.692 CD (B-D) (α = 0.5) = 3.262 CD (B-D) (α = 0.1) = 3.036
Control model	Comparison model	2 × 2 contingency matrix	AUC_{E0}	$AR_{(1\text{-}E0)}$	Type I_{E0}	Type II_{E0}	AUC_{E0}
CADF univariate	LDA	0.000***	38.93 (0.000)***	3.818 (0.000)***	−4.349 (0.000)***	1.771 (0.076)$^{+}$	3.870**
	LR	0.000***	38.40 (0.000)***	3.837 (0.000)***	−4.590 (0.000)***	1.988 (0.046)*	3.920**
	Knn	0.887	10.30 (0.001)**	5.567 (0.000)***	−3.026 (0.003)**	−5.990 (0.000)***	9.580**
	SVM lineal	0.000***	37.18 (0.000)***	4.460 (0.000)***	−5.826 (0.000)***	5.449 (0.000)***	5.550**
	SVM 2-pol	0.001**	37.84 (0.000)***	3.905 (0.000)***	−5.604 (0.000)***	4.749 (0.000)***	3.800**
	NN	0.113	27.68 (0.000)***	4.416 (0.000)**	−0.734 (0.463)	−2.650 (0.008)**	5.330**
	ChAID	0.000***	41.58 (0.000)***	4.638 (0.000)***	−4.514 (0.000)***	0.217 (0.828)	6.300**
	Assistant	0.688	29.92 (0.000)***	3.253 (0.001)**	3.186 (0.001)**	−5.551 (0.000)***	5.180**
	C4.5	0.078^{+}	46.27 (0.000)***	3.142 (0.002)**	−2.014 (0.044)*	−1.958 (0.050)$^{+}$	3.780*
	CART-uni	0.568	29.19 (0.000)***	3.935 (0.000)***	3.213 (0.001)**	−5.371 (0.000)***	4.050**
	CART-obli	0.386	2.65 (0.103)	2.471 (0.014)**	2.828 (0.005)**	−4.947 (0.000)***	2.960
	Gradient boosting	0.999	0.55 (0.460)	−1.627 (0.104)	0.275 (0.783)	1.390 (0.165)	−0.410
	Random forests	0.814	0.01 (0.914)	1.299 (0.194)	−4.909 (0.000)***	5.029 (0.000)***	1.290
	SVDF univariate	0.044*	0.51 (0.475)	3.293 (0.001)**	−6.088 (0.000)***	6.053 (0.000)***	1.970
	WVDF univariate	0.648	0.13 (0.717)	2.019 (0.044)*	−1.578 (0.115)	−0.076 (0.939)	0.750
Control model	Comparison model	2 × 2 contingency matrix	AUC_{E0}	$AR_{(1\text{-}E0)}$	Type I_{E0}	Type II_{E0}	Friedman test = 292.80 (0.000)*** Nemenyi B–D test CD (B–D) (α = 0.01) = 4.465 CD (B–D) (α = 0.5) = 3.958 CD (B–D) (α = 0.1) = 3.704
CADF mixed	LDA	0.000***	44.69 (0.000)***	4.865 (0.000)***	−5.285 (0.000)***	1.945 (0.052)$^{+}$	6.880**
	LR	0.000***	44.21 (0.000)***	4.730 (0.000)***	−5.420 (0.000)***	2.244 (0.025)*	6.890**
	Knn	0.035*	10.45 (0.001)**	5.797 (0.000)***	−4.470 (0.000)***	−6.000 (0.000)***	12.920**
	SVM lineal	0.000***	42.62 (0.000)***	5.053 (0.000)***	−6.009 (0.000)***	5.662 (0.000)***	8.730**
	SVM 2-pol	0.000***	43.83 (0.000)***	4.745 (0.000)***	−5.980 (0.000)***	5.015 (0.000)***	6.920**
	NN	0.000***	30.66 (0.000)***	5.073 (0.000)***	−1.810 (0.070)$^{+}$	−2.824 (0.005)**	8.250**
	ChAID	0.000***	50.14 (0.000)***	5.638 (0.000)***	−4.865 (0.000)***	0.246 (0.806)	9.690**
	Assistant	0.024*	33.37 (0.000)***	5.232 (0.000)***	1.835 (0.066)$^{+}$	−5.961 (0.000)***	8.250**
	C4.5	0.000***	51.39 (0.000)***	5.054 (0.000)***	−3.832 (0.000)***	−2.051 (0.040)*	6.950**
	CART-uni	0.003**	32.14 (0.000)***	4.707 (0.000)***	0.710 (0.478)	−5.734 (0.000)***	6.850**
	CART-obli	0.606	2.50 (0.114)	3.919 (0.000)***	1.787 (0.074)$^{+}$	−5.362 (0.000)***	5.800**
	Gradient boosting	0.015*	0.65 (0.419)	−0.130 (0.896)	−1.497 (0.134)	1.530 (0.126)	1.420
	Random forests	0.010*	0.02 (0.882)	3.066 (0.022)*	−5.488 (0.000)***	5.087 (0.000)***	3.790^{+}
	SVDF univariate	0.000***	0.12 (0.733)	4.656 (0.000)***	−6.149 (0.000)***	5.749 (0.000)***	4.620*
	WVDF unvariate	0.007**	0.00 (0.997)	4.254 (0.000)***	−3.952 (0.000)***	0.403 (0.687)	3.210^{+}
	CADF univariate	0.021*	0.00 (0.944)	3.782 (0.000)***	−3.679 (0.000)***	−0.324 (0.746)	2.330
	SVDF mixed	0.054^{+}	0.56 (0.454)	4.378 (0.000)***	−4.262 (0.000)***	0.537 (0.592)	2.970
	WVDF mixed	0.054^{+}	0.25 (0.615)	4.378 (0.000)***	−4.262 (0.000)***	0.537 (0.592)	2.970

*** $p < 0.001$.
** $p < 0.01$.
* $p < 0.05$.
$^{+}$ $p < 0.1$.

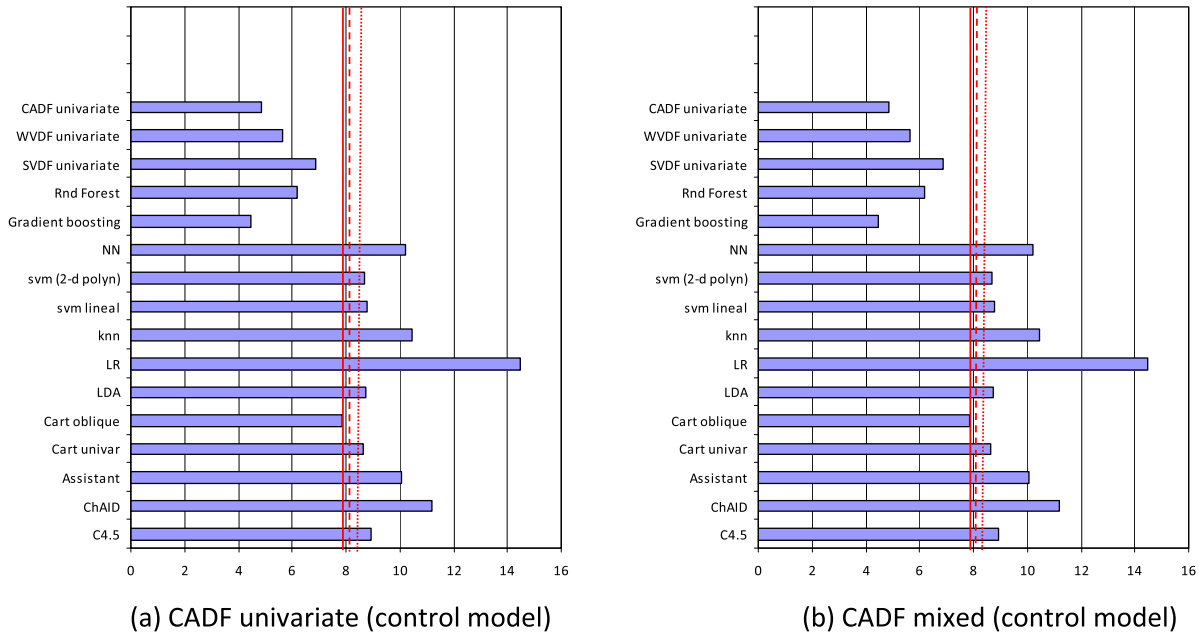


Fig. 4. Significance diagram for the Nemenyi Bonferroni–Dunn test ($\alpha = 0.01$, $\alpha = 0.05$, $\alpha = 0.1$).

reduced. CADF univariate is by large the most interpretable DF, including 41 intuitive rules (5 of them being fired each time); it also has significantly lesser number of selected features than any other DF. To illustrate interpretability, an example of CADF univariate and CART mixed rule firing is included in Appendix B.

To further evaluate the significance of differences between CADF and alternative classifiers, Table 4 includes non-parametric statistic tests for model comparison. As preliminary results based on the full dataset, the McNemar paired test finds significant AR differences between CADF mixed and any other model but CART oblique ($p < 0.05$). However, AR differences are not so significant whether CADF univariate is considered as control model. DeLong test allows to discover whether there exist significant differences in the AUC results for the full dataset. Results confirm CADF mixed and CADF univariate to be statistically better than any individual classifiers but CART oblique ($p < 0.001$). However, no significant AUC differences are found between DF models for the full dataset. Previous results have been necessarily completed and extended with out-of-sample statistical tests based on bootstrapping partitions as follows,

(a) Based on the nonparametric Wilcoxon signed rank test, CADF mixed has a significantly higher AR performance than any other model ($p < 0.001$), excluding gradient boosting where no statistical differences are reported. Respectively, CADF univariate statistically outperforms any classifier but CADF mixed ($p < 0.001$), while differences with gradient boosting and random forests are not statistically significant. Moreover, the performance of the adjusted-R^2 weighted vote strategy is found to be superior to that of simple vote and weighted vote in univariate and oblique CADF.

(b) About type I and type II errors, Wilcoxon test reports some statistical differences between CADF and alternative classifiers, due to distinct balances in class predictions. In terms of type I error, CADF mixed performs significantly better than any other classifier ($p < 0.05$), with except of NN, Assistant, CART univariate, CART oblique and gradient boosting where no significant differences are found. Opposite, it presents a higher type II error than logistic regression, SVM, random forest and SVDF univariate proposals. In the same line, CADF univariate gets a significantly minor type I error than any model but NN, Assistant, CART univariate, CART oblique and gradient boosting; however, type II error is superior than such from logistic regression, SVM, and random forests.

(c) The Friedman test confirmed significant differences in the AUC of different classifiers ($p < 0.001$), so the post hoc Nemenyi test with Bonferroni–Dunn (B–D) adjustment was applied to report any significant differences with respect to CADF (univariate and mixed). Fig. 4 displays the AUC mean ranks of the classifiers, along with the critical values at $\alpha = 0.01$, $\alpha = 0.05$, and $\alpha = 0.10$ (successive vertical lines from left to right). All algorithms above these cut lines perform significantly worse than CADF. From the Nemenyi B-D test, one can observe that CADF mixed statistically outperforms any not-ensemble techniques ($\alpha < 0.01$), random forests ($\alpha < 0.1$), SVDF univariate ($\alpha < 0.05$) and WVDF ($\alpha < 0.1$). Besides, not substantial differences are found with gradient boosting (the best performing AUC model) or CADF univariate. Similarly, the CADF univariate performs better than any individual model but CART oblique where no differences are reported. It is interesting to remark that AUC differences between CADF univariate and multiple-data partition DF (gradient boosting, random forests) are not statistically significant.

In summary, it can be concluded that the CADF univariate yielded a very good accuracy, being better than any individual model and as good as the best complex DF, with 5% level of confidence. In terms of model's interpretability, CADF univariate was the best of ensemble strategies, providing a small number of fired rules based on a reduced feature subset, revealing that it could be used as a competitive solution to credit risk evaluation. Previous results are in line with Zhou et al. (2010) that induce simple ensemble models built on a dual diversity strategy, without reporting significant accuracy losses in comparison to more complex models. Oppositely, CADF mixed obtains a better AR than univariate proposal but its interpretability is much lower. Finally, multiple data

partition DF as gradient boosting and random forests do not perform significantly better than CADF but produce an ensemble structure with very limited interpretability. It is a major concern that could lead on reluctance to use credit scoring models.

6. Conclusions and future work

Due to the rising number of defaults in last years, credit risk assessment has become a prevalent research interest in the post-crisis scenario. Since an increase in the discrimination ability between good and payers of even a fraction of a percent may give to significant savings for financial institutions, continuous proposals have been done to improve models' accuracy. In this line, ensemble strategies have been proposed as some of the most powerful Machine Learning paradigms to assess credit default. From them, decision forests based on merging multiple decision trees, are one of the most accepted ensemble strategies. However, ensemble methods exhibit a major limitation: the lack of interpretability of the results. However, on real world lending environments, the understanding and the explicability of the scoring process encapsulated within the classifiers is key for their use and even mandatory in many countries.

In this study, a novel ensemble strategy, the correlated-adjusted decision forest (CADF) is proposed to balance accuracy and interpretability of credit models. Based on a three-stage structure, CADF explicitly introduces multiple sources of diversity and a pseudo-correlation penalty combination function, with the aim of obtaining a good accuracy while maintaining a high interpretability through a limited number of logical, human understanding rules. Different variants of CADF were presented and compared with thirteen well-known classifiers, including alternative decision forests (gradient boosting, random forests). For verification a publicly available credit dataset has been used to the effectiveness of the proposed ensemble approach, using a .632 bootstrapping approach. The classification power of these techniques was assessed based on multiple accuracy measures (accuracy rate, type I error, type II error, AUC). McNemar test and Wilcoxon test were applied to determine statistical differences between accuracy results and errors. Respectively, DeLong, Friedman and Nemenyi Bonferroni–Dunn tests were applied to analyse if the differences between the average ranked AUC performance of CADF and other techniques were statistically significant.

Empirical results revealed that ensemble methods consistently outperform to any individual model in terms of accuracy rate and AUC measures. Among them, the CADF univariate provided the best interpretably results in terms of number of rules, number of features, number of rules fired each time, and distinguishability of variable partition. Besides, CADF univariate outperformed any individual classifier in terms of out-of-sample accuracy, leading to similar classification rates than more complex ensemble proposals. Even if differences could be considered as numerically small, in credit risk context very small improvements in classification accuracy could results in millions of dollars of additional profits. The CADF mixed alternative, based on merging simple and oblique decision tree rules, obtained the best accuracy rate of any other comparable model; however model's complexity reduces its interpretability. Results suggest that the proposed CADF univariate model provides a promising solution to the demanded increase of interpretability of ensemble methods, while maintaining a very high accuracy rate for credit scoring predictions.

In comparison to alternative decision forests, CADF has potential advantages: It reduces final model dimensionality's, exhibiting a white-box internal structure that can be directly interpreted; besides, it provides a good balance between type I and type II errors that suggest its suitability for dealing with imbalance datasets (Kim et al., 2015). As a result, we believe our proposal has practical managerial implications on credit risk assessments. It provides a potential solution for the practice-oriented need of interpretability without losing accuracy, which facilitates its adoption as a managerial tool by industrial organisations and users (Chen & Cheng, 2013). Model's interpretability also allows the integration of the quantitative model with expertise information; if experts decide that some rules are not fully adequate, they could refine it according to their managerial experience (Tomczak & Zieba, 2015). In the same way, the easy ensemble structure of CADF facilitates its update with time going on, a key issue almost neglected by literature (Sun et al., 2014).

Different theoretical contributions can be highlighted. Our proposal uses complementary sources of diversity as mechanisms to balance accuracy and interpretability of ensemble models. Up to our knowledge it is the first attempt to do so in front of the accuracy focus of previous diversity-based researches (Kuncheva & Whitaker, 2003; Sun & Li, 2012). To get it, we depart on a multiple learning classifier strategy, which has been infrequently explored in credit scoring literature; our results add evidence to Sun et al. (2014), who report multiple classifier ensembles as the strongest diversity strategy for financial default problems. Besides, the use of a pseudo-correlation penalty function replies to the research calling for searching combination mechanisms better than voting in credit risk domains (Sun et al., 2014). In parallel, CADF provides decision rules that fully represent the decision process of the ensemble classifier, in front of rule-extraction approaches that partially reveal the internal classification system (Baesens et al., 2003; Chen & Cheng, 2013; Derelioglu & Gürgen, 2011; Martens et al., 2007; Setiono et al., 2011; Wu & Hu, 2012). Interpretability facilitates comparisons between algorithmic results and research hypotheses on default causes, in order to develop an overall theoretical framework on financial distress and credit risk.

Our proposal also exhibit several limitations and directions for future research. The main shortcoming is that base decision trees have a high linear association that reduces the true classifier diversity in spite of the correlation-adjusted penalty strategy; to face it, alternative comprehensible algorithms should be explored as imprecise trees (Abellan & Mantas, 2014), oblique random forest (Menze, Kelm, Splitthoff, Koethe, & Hamprecht, 2011) or disjunctive normal random forests (Seyedhosseini & Tasdizen, 2015). As a result, merged decision rules are redundant, so rule-filtering algorithms should be applied to simplify simultaneously fired rules (Chen & Cheng, 2013; Hsieh & Hung, 2010). Since our results are based on a well-known but only credit risk dataset, a large number of datasets from other domains should be examined to assess the quality of the model. Further comparisons of CADF are also needed regarding alternative ensemble approaches (p.e. stacking models), combination functions (p.e. Kuncheva & Whitaker, 2003 Yule's Q statistics), and quality measures.

Acknowledgements

This work has been partially supported by the Spanish Andalusian government under grants SEJ-1933, SEJ-5061 and SEJ-111.

Appendix A

See Appendix A and B.

Appendix A

Ensemble strategies and their applications in credit scoring.

	Ensemble strategy			Datasets	Model comparison	Findings	
	Diversity sources	Base algorithms	Combination function			Accuracy gains	Interpretability
Abellan and Masegosa (2010)	Bagging	Credal DT	Majority vote (100-500 base classifiers)	25 credit risk and non-credit risk public datasets	Bagging ensemble of DT (C4.5)	Slightly better classification accuracy in presence of noise, bias and variance	Relative number of rules of decision trees (36 to 416 per tree).
Hsieh and Hung (2010)	Bagging + multi-classifier system	NN, Bayesian network (BN), SVM	Confidence-weighted vote	German credit risk dataset	No	High accuracy but no model comparison	Decision rules based on individual classifier (BN)
Paleologo et al. (2010)	Bagging variant (subagging)	Single (DT, Linear SVM, polynomial SVM, RBF SVM, NN, knn)	Majority vote (100 base classifiers)	Company customer dataset (proprietary)	Individual classifiers: DT, Linear SVM, polynomial SVM, RBF SVM, NN, knn	Subagging strategies improve AUC of individual classifiers	Average tree size in subgagging ensemble (3-25 nodes), variable importance scores.
Twala (2010)	Single: bagging, boosting, stacking	LR, DT, NN, SVM.	Majority vote	German, Australian, loan dataset, Texas bank datasets	Individual classifiers: LR, k-nn,DT, Naïve Bayes, NN	Ensemble proposals improve the performance of alternative classifiers	No
Yu et al. (2010)	Bagging + parameter diversity	SVM	Majority vote (100-5000 base classifiers), weighted majority vote, adaptive linear NN combination function	British credit card dataset	Individual classifiers: LDA, QDA, LR, NN, SVM	Ensemble model with adaptive linear NN combination function outperforms alternative approaches	No
Zhou et al. (2010)	Simple data partition + parameter diversity	SVM (10 base classifiers)	Correlation-based selective functions, majority vote, weighted vote, reliability based combination	German and UK credit risk datasets	Individual classifiers: LDA, QDA, LR, probit, DT, NN, probabilistic NN, Bayesian classifier, knn, boosting (SVM)	High accuracy in any datasets (performance at least as good as the best)	No
Finlay (2011)	Single: Multi-classifier system, bagging, boosting, boosting variant (error trimmed), local accuracy dynamic	Single (LR, LDA, DT, NN, knn)	Majority vote (any ensemble). LR combination and NN combination (multi- classifier system)	Two UK retail datasets	Individual classifiers: LDA, LR, NN, knn, DT.	Error trimmed boosting and multi-classifier system (logistic and NN combination) outperform alternative models	No
Wang et al. (2011)	Single: bagging, boosting, multi-classifier system	LR, DT, NN, SVM.	Majority vote (100 base classifiers)	Australian, German, and Chinese credit risk datasets	Individual classifiers: SVM, NN, DT, LR	Stacking and bagging DT as the most accurate methods	No
Brown and Mues (2012)	Gradient boosting, random forests (bagging + random subspace)	DT	Majority vote (10-1,000 base classifiers)	German, Australia and three Benelux credit risk datasets	Individual classifiers: LR, LDA, QDA, knn, SVM	Random forests obtains the highest accuracy	No
De Bock and Van den Poel (2012)	Bagging + random subspace	Generalized additive models	Majority vote	Six real-life customer churn datasets	Individual classifiers: LR and generalised additive models. Bagging, random subspace and boosting.	Higher accuracy than alternative classifiers but random forests	Variable importance scores and predictive tendencies
Li et al (2012)	Boosting variant (soft margin boosting) + different parameters of learning algorithms	Relevance vector machines (RVM)	Majority vote (up to 200 base classifiers)	Australian and Japanese credit risk datasets	Individual classifiers: LR, SVM, RVM. Boosting of previous models	Higher accuracy than alternative classifiers	No
Marques et al (2012a)	Single: Bagging, boosting, random subspace, Decorate or rotation forest	Single (DT, MLP NN, radial basis function NN, knn, Naïve Bayes, LR,SVM)	Majority vote	Australian, German, Japanese, Iranian, Polish, UCSD credit datasets	Individual classifiers (DT, MLP NN, radial basis function NN, knn, Naïve Bayes, LR, SVM)	Boosting DT obtains the highest accuracy in most of datasets	No
Marques et al (2012b)	Single data partition (boosting, bagging) + single feature subsets (random subspace, rotation forest)	DT (C4.5)	Majority vote	Australian, German, Japanese, Iranian, Polish, UCSD credit datasets	Individual classifiers: NN, LR, knn, SVM, DT. Bagging, boosting, rotation subspace and rotation forest of previous classifiers	Dual-diversity ensembles outperform alternative models. Bagging+rotation forest as the most accurate method	No
Sun and Li (2012)	Single feature subsets (PCA, stepwise LDA, stepwise LR) + parameter diversity	SVM	Weighted majority vote (2-15 base classifiers)	Chinese stock exchange bankruptcy dataset	Individual SVM	SVM ensemble with at least 9 classifiers outperforms individual models	No
Tomczak and Zieba (2012)	No	RVM	No	Australian, German, Kraggle, Polish loans datasets.	Individual classifiers (MLP, LR, SVM, DT). Boosting, bagging, random forests (DT)	High accuracy of RVM and scoring tables, not always optimal	Scoring tables
Wang and Ma (2012)	Bagging + random subspace	SVM	Majority vote	Chinese enterprise credit risk dataset (proprietary)	Individual classifiers: LR, DT, NN, SVM (linear), SVM (polynomial). Bagging and boosting or previous models	Bagging + random subspace gets the highest performance (accuracy rate, type I, type II errors)	No
Wang et al. (2012)	Bagging + random subspace Random subspace + bagging	DT	Majority vote	Australian and German datasets	Individual classifiers: LR, LDA, DT, MLP NN, radial basis function NN. Bagging and boosting.	Higher performance than individual classifiers and alternative ensembles.	No
Fedorova et al. (2013)	First stage: Single feature selection (LDA, logistic regression, DT)	Second stage: NN	Third stage: Boosting, logistic regression	Russian manufacturing companies bankruptcy	Individual classifiers: MDA, LR, DT, NN	Boosting of five NN produce a higher accuracy than individual classifiers	No (just important variables are identified)
Kruppa et al. (2013)	Random Forests (bagging + random subspace)	Probability estimation trees (PET)	Majority probability (500 base classifiers)	Company customer dataset (proprietary)	Knn, bagged knn, optimized logistic regression	Higher accuracy than comparison approaches (Brier score)	Variable importance scores.
Abellan and Mantas (2014)	Bagging	Credal DT	Majority vote	Japanese, Australian and German credit datasets	Bagging and random subspace ensembles of decision trees (C4.5) and NN	Bagging credal DT obtain the highest AUC	No
Tsai et al. (2014)	Single: Boosting, bagging	Single (NN, SVM, DT)	Majority vote (10-1000 base classifiers)	Japanese, Australian, German credit datasets. Taiwan bankruptcy dataset	Individual classifiers: NN, SVM, DT	Superior performance of boosting-100 DT over alternative models	No (just computational cost)
Wang et al. (2014)	Boosting + random subspace variant (feature selection boosting)	DT	Majority vote	Two real bankruptcy datasets	Individual classifiers: LR, DT, Naïve Bayes, SVM, NN. Bagging and boosting of previous classifiers	Boosting + random subspace gets the highest performance (accuracy rate, type I, type II errors)	No
Geng et al (2015)	Multi-classifier system	NN, DT, SVM	Majority vote	China bankruptcy dataset	12 individual classifiers (statistical and ML models)	Similar performance of ensemble model and NN	No
Kim et al (2015)	Boosting variant (geometric mean based boosting)	SVM	Majority vote (up to 25 base classifiers)	Korean commercial bank bankruptcy dataset	Boosting, cost-sensitive boosting	Higher accuracy than alternative models	No

Appendix B

Example of decision rule firing.

CADF univariate (logical rules[a])		Score (good)	Score (bad)
ChAID	If VAR1 = {2} and VAR5 < 12296.50 then good		
Assistant	If VAR1 = {1,2} and VAR2 $\geqslant$ 47.50 and VAR6 $\neq$ {4, 5} then bad	SVDF: 0.250	SVDF: 0.750
C4.5	If VAR1 = {2} and VAR2 $\geqslant$ 22.50 and VAR6 = {1, 2} then bad	WVDF: 0.239	WVDF:0.761
CART univariate	If VAR1 $\neq$ {3, 4} and VAR2 > 47.50 and VAR6 $\neq$ {4, 5} then bad	CADF:0.157	CADF: 0.843
CAF mixed (logical rules[a]). Previous rules plus			
CART oblique	If VAR1 $\neq$ {3, 4} and -1(VAR2) + 0.002(VAR5) $\leqslant$ 0.064 and	SVDF: 0.200	SVDF: 0.800
	0.064(VAR2)$-$0.001(VAR5)$-$0.109(VAR8)$-$0.274(VAR11)$-$0.956(VAR16) $\leqslant -2.267$ and	WVDF: 0.189	WVDF: 0.811
	$-$0.012(VAR2)$-$0.009(VAR5)$-$0.576(VAR8)$-$0.817(VAR18) $\leqslant$ 5.412 then bad	CADF: 0.154	CADF: 0.846

[a] VAR1 = Checking account {1: <0; 2: $\geqslant$0 and <200DM; 3: $\geqslant$200DM; 4: no checking account}. VAR2 = Credit duration (months). VAR5 = Credit amount; VAR6 = Saving accounts/bonds {1: <100DM; 2: $\geqslant$100 DM and <500DM; 3: $\geqslant$500DM and <1000DM; 4: $\geqslant$1000 DM; 5: unknown/no savings account}. VAR 8 = Installment rate in percentage of disposable income. VAR11 = Present residence since; VAR16 = Number of existing credits at this bank. VAR18 = Number of people liable to provide maintenance for.

References

Abellan, J., & Mantas, C. (2014). Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications, 41*, 3825–3830.

Abellan, J., & Masegosa, A. R. (2012). Bagging schemes on the presence of class noise in classification. *Expert Systems with Applications, 39*, 6827–6837.

Altman, E. I. (1968). Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy. *The Journal of Finance, 23*, 589–609.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society, 54*(6), 627–635.

Basel Committee on Banking Supervision (BCBS). (2011, June). Basel III: A global regulatory framework for more resilient banks and banking systems. Basel: Bank for International Settlements.

Breiman, L., Freidman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont: Wadsworth.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123–140.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications, 39*, 3446–3453.

Cestnik, B., Kononenko, I., & Bratko, I. (1987). ASSISTANT 86: A knowledge elicitation tool for sophisticated users. In I. Bratko & N. Lavrac (Eds.), *Progress in machine learning*. Wilmslow: Sigma Press.

Chen, Y.-S., & Cheng, C.-H. (2013). Hybrid models based on rough set classifiers for setting credit risk rating decision rules in the global banking industry. *Knowledge-Based Systems, 39*, 224–239.

Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research, 183*, 1447–1465.

Danenas, P., & Garsva, G. (2015). Selection of support vector machines based classifiers for credit risk domain. *Expert Systems with Applications, 42*, 3194–3204.

Daubie, M., Levecq, P., & Meskens, N. (2002). A comparison of the rough sets and recursive partitioning induction approaches: An application to commercial loans. *International Transactions in Operational Research, 9*, 681–694.

De Bock, K. W., & Van den Poel, D. (2012). Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models. *Expert Systems with Applications, 39*, 6816–6826.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics, 44*, 837–845.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, 7*, 1–30.

Derelioglu, G., & Gürgen, F. (2011). Knowledge discovery using neural approach for SME's credit risk analysis problem in Turkey. *Expert Systems with Applications, 38*, 9313–9318.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In J. Kittler & F. Roli (Eds.), *First international workshop on multiple classifier systems* (pp. 1–15). New York: Springer Verlag.

Efron, B., & Tibshirani, R.J. (1995). Cross-validation and the bootstrap: Estimating the error rate of a prediction rule. Technical Report 176, Stanford University.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman and Hall.

Fedorova, E., Gilenko, E., & Dovzhenko, S. (2013). Bankruptcy prediction for Russian companies: Application of combined classifiers. *Expert Systems with Applications, 40*, 7285–7293.

Feldman, D., & Gross, D. (2005). Mortgage default: Classification trees analysis. *The Journal of Real State Finance and Economics, 30*(4), 369–396.

Finlay, S. (2011). Multiple classifier architectures and their applications to credit risk assessment. *European Journal of Operational Research, 210*, 368–378.

Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics, 11*, 86–92.

Friedman, J.H. (1999a, February). Greedy function approximation: A gradient boosting machine. Technical Document, Stanford University.

Friedman, J.H. (1999b, March). Stochastic gradient boosting. Technical Document, Stanford University.

Gacto, M. J., Alcala, R., & Herrera, F. (2011). Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures. *Information Sciences, 181*, 4340–4360.

Geng, R., Bose, I., & Chen, X. (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research, 241*, 236–247.

Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science, 21*, 1–15.

Hand, D. J. (2009). Measuring classifier performance. A coherent alternative to the area under the ROC curve. *Machine Learning, 77*, 103–123.

Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M., & Dougherty, E. R. (2010). Small-sample precision of ROC-related estimates. *Bioinformatics, 26*, 822–830.

Härdle, W., Moro, R., & Schafer, D. (2005). Predicting bankruptcy with support vector machines. In P. Cizek, W. Härdle, & R. Weron (Eds.), *Statistical tools for finance and insurance* (pp. 225–248). Berlin: Springer.

Harris, T. (2015). Credit scoring using the clustered support vector machine. *Expert Systems with Applications, 42*(2), 741–750.

Henley, W. E., & Hand, D. J. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society, Series A, 160*(3), 523–541.

Ho, T.K. (1998). C4.5 decision forest. In *Proceedings of the 14th international conference on pattern recognition* (pp. 545–549), Brisbane, Australia.

Hsieh, N.-C., & Hung, L.-P. (2010). A data driven ensemble classifier for credit scoring analysis. *Expert Systems with Applications, 37*, 534–545.

Huang, C. L., Chen, M. C., & Wang, C. J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications, 33*, 847–856.

Hung, C., & Chen, J. H. (2009). A selective ensemble based on expected probabilities for bankruptcy prediction. *Expert Systems with Applications, 36*, 5297–5303.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics, 29*(2), 119–127.

Kestens, K., Van Cauwenberge, P., & Vauwedhe, H. V. (2012). Trade credit and company performance during the 2008 financial crisis. *Accounting and Finance, 52*, 1125–1151.

Kim, J. W. (1993). Expert systems for bond rating: A comparative analysis of statistical, rule-based and neural network systems. *Expert Systems, 10*, 167–171.

Kim, M.-J., Kang, D.-K., & Kim, H. B. (2015). Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. *Expert Systems with Applications, 42*, 1074–1082.

Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning, 51*(2), 181–207.

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications, 40*, 5125–5131.

Lee, T.-S., & Chen, I. F. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications, 28*(4), 743–752.

Li, S., Tsang, I. W., & Chaudhari, N. S. (2012). Relevance vector machine based infinite decision agent ensemble learning for credit risk analysis. *Expert Systems with Applications, 39*, 4947–4953.

Marques, A. I., Garcia, V., & Sanchez, J. S. (2012a). Exploring the behaviour of base classifiers in credit scoring ensembles. *Expert Systems with Applications, 39*, 10244–10250.

Marques, A. I., Garcia, V., & Sanchez, J. S. (2012b). Two-level classifier ensembles for credit risk assessment. *Expert Systems with Applications, 39*, 10916–10922.

Martens, D., Baesens, B., Van Gestel, T., & Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research, 183*(3), 1466–1476.

Menze, B. H., Kelm, B. M., Splitthoff, D. N., Koethe, U., & Hamprecht, F. a. (2011). On oblique random forests. In D. Gunopulos, T. Hofmann, D. Malerba, & M. Vazirgiannis (Eds.), *Machine learning and knowledge discovery in databases* (pp. 453–469). Berlin: Springer.

Mues, C., Baesens, B., Files, C. M., & Vanthienen, J. (2004). Decision diagrams in machine learning: an empirical study on real-life credit-risk data. *Expert Systems with Applications, 27*, 257–264.

Nanni, L., & Lumini, A. (2009). An experimental comparison of ensemble classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications, 36*, 3028–3033.

Nemenyi, P.B. (1963). *Distribution-free multiple comparisons* [Ph.D. thesis]. Princeton University.

Ong, C.-S., Huang, J.-J., & Tzeng, G.-H. (2005). Building credit scoring models using genetic programming. *Expert Systems with Applications, 29*(1), 41–47.

Paleologo, G., Eliseeff, A., & Antonini, G. (2010). Subagging for credit scoring models. *European Journal of Operational Research, 201*, 490–499.

Quinlan, J. R. (1979). Discovering rules by induction from large collections of examples. In D. Michie (Ed.), *Expert systems in the micro electronic age*. Edinburgh: Edinburg University Press.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Los Altos, CA: Morgan Kaufmann.

Rodriguez, J. J., Kuncheva, L. I., & Alonso, C. J. (2006). Rotation forests: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(10), 1619–1630.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning, 5*, 197–227.

Setiono, R., Baesens, B., & Mues, C. (2011). Rule extraction from minimal neural networks for credit card screening. *International Journal of Neural Systems, 21*(4), 265–276.

Seyedhosseini, M., & Tasdizen, T. (2015). Disjunctive normal random forests. *Pattern Recognition, 48*, 976–983.

Sun, J., Li, H., Huang, Q.-H., & He, K. Y. (2014). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge-Based Systems, 57*, 41–56.

Sun, J., & Li, H. (2012). Financial distress prediction using support vector machines: Ensemble vs. individual. *Applied Soft Computing, 12*, 2254–2265.

Tsai, C.-F., Hsu, Y.-F., & Yen, D. C. (2014). A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing, 24*, 977–984.

Tomczak, J. M., & Zieba, M. (2015). Classification restricted Boltzmann machine for comprehensible credit scoring model. *Expert Systems with Applications, 42*, 1789–1796.

Twala, B. (2010). Multiple classifier application to credit risk assessment. *Expert Systems with Applications, 37*, 3326–3336.
Wang, G., & Ma, J. (2011). Study of corporate credit risk prediction based on integrating boosting and random subspace. *Expert Systems with Applications, 38*, 13871–13878.
Wang, G., & Ma, J. (2012). A hybrid ensemble approach for enterprise credit risk assessment based on support vector machine. *Expert Systems with Applications, 39*, 5325–5331.
Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications, 38*, 223–230.
Wang, G., Ma, J., Huang, L., & Xu, K. (2012). Two credit scoring models based on dual strategy ensemble trees. *Knowledge-Based Systems, 26*, 61–68.
Wang, G., Ma, J., & Yang, S. (2014). An improved boosting based on feature selection for corporate bankruptcy prediction. *Expert Systems with Applications, 41*, 2353–2361.
West, D. (2000). Neural networks credit scoring models. *Computers & Operations Research, 22*, 1131–1152.
Wolpert, D. (1992). Stacked generalization. *Neural Networks, 5*, 241–259.
Wu, T.-H., & Hu, M.-F. (2012). Credit risk assessment and decision making by a fusion approach. *Knowledge-Based Systems, 35*, 102–110.
Yu, L., Yue, W., Wang, S., & Lai, K. K. (2010). Support vector machine based multiagent ensemble learning for credit risk evaluation. *Expert Systems with Applications, 37*, 1351–1360.
Yu, L., Wang, S., & Lai, K. K. (2008). Credit risk assessment with a multistage neural network ensemble learning approach. *Expert Systems with Applications, 34*, 1434–1444.
Zhou, L., Lai, K. K., & Yu, L. (2010). Least squares support vector machines ensemble models for credit scoring. *Expert Systems with Applications, 37*, 127–133.