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, &