Expert Systems with Applications 39 (2012) 10244–10250

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Exploring the behaviour of base classifiers in credit scoring ensembles

A.I. Marqués [a], V. García [b], J.S. Sánchez [b,*]

[a] Department of Business Administration and Marketing, Universitat Jaume I, Av. Sos Baynat s/n, 12071 Castelló de la Plana, Spain
[b] Department of Computer Languages and Systems, Universitat Jaume I, Av. Sos Baynat s/n, 12071 Castelló de la Plana, Spain

ARTICLE INFO

Keywords:
Finance
Credit scoring
Classifier ensemble

ABSTRACT

Many techniques have been proposed for credit risk assessment, from statistical models to artificial intelligence methods. During the last few years, different approaches to classifier ensembles have successfully been applied to credit scoring problems, demonstrating to be more accurate than single prediction models. However, it is still a question what base classifiers should be employed in each ensemble in order to achieve the highest performance. Accordingly, the present paper evaluates the performance of seven individual prediction techniques when used as members of five different ensemble methods. The ultimate aim of this study is to suggest appropriate classifiers for each ensemble approach in the context of credit scoring. The experimental results and statistical tests show that the C4.5 decision tree constitutes the best solution for most ensemble methods, closely followed by the multilayer perceptron neural network and logistic regression, whereas the nearest neighbour and the naive Bayes classifiers appear to be significantly the worst.

1. Introduction

The recent world financial crisis has aroused increasing attention of banks and financial institutions on credit risk. The main problem comes from the difficulty to distinguish the creditworthy applicants from those who will probably default on repayments. The decision to grant credit to an applicant was traditionally based upon subjective judgments made by human experts, using past experiences and some guiding principles. Common practice was to consider the classic five C's of credit: character, capacity, capital, collateral and conditions (Abrahams & Zhang, 2008). This method suffers, however, from high training costs, frequent incorrect decisions, and inconsistent decisions made by different experts for the same application.

These shortcomings have led to a rise in more formal and accurate methods to assess the risk of default. In this context, credit scoring and behavioural scoring have become primary tools for financial institutions to evaluate credit risk, improve cash flow, reduce possible risks and make managerial decisions (Thomas, Edelman, & Crook, 2002). Difference between credit scoring and behavioural scoring is that the former focuses on decisions regarding new applicants for credit, whereas the latter refers to monitoring and predicting the repayment behaviour of existing borrowers.

Credit scoring is essentially a set of techniques that help lenders decide whether or not to grant credit to applicants. The aim of credit scoring models is to discriminate between "good" and "bad" loans, depending on how likely applicants are to default with their repayments. Compared with the traditional subjective methods, credit scoring models present some advantages (Rosenberg & Gleit, 1994; Thomas et al., 2002): (i) they are cheaper to purchase and operate; (ii) they make faster credit decisions; (iii) they provide consistent recommendations based on objective information, thus eliminating human biases and prejudices; (iv) changes in policy and/or economy can easily be incorporated into the system; and (v) the performance of the credit scoring model can be monitored, tracked, and adjusted at any time.

From the seminal reference to credit scoring in the introductory paper by Altman (1968), many other developments have been subsequently proposed in the literature. The most classical approaches to credit scoring are based on statistical and mathematical programming models, such as linear and quadratic discriminant analysis, linear and logistic regression, multivariate adaptive regression splines, Markov chain models, and linear and quadratic programming. However, after the Basel II recommendations issued by the Basel Committee on Banking Supervision in 2004, financial institutions have required to use more complex credit scoring models in order for enhancing the efficiency of capital allocation.

In recent years, many studies have demonstrated that artificial intelligence techniques (decision trees, artificial neural networks, support vector machines, evolutionary computing) can be successfully used for credit evaluation (Chi & Hsu, 2012; Huang, Chen, Hsu, Chen, & Wu, 2004; Huang, Chen, & Wang, 2007; Ince & Aktan, 2009; Martens et al., 2010; Ong, Huang, & Tzeng, 2005). In contrast to the statistical models, the artificial intelligence methods do not assume any specific prior knowledge, but automatically extract information from past observations.

* Corresponding author. Tel.: +34 964 728350.
E-mail address: sanchez@uji.es (J.S. Sánchez).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.02.092

Although previous studies conclude that artificial intelligence techniques are superior to traditional statistical models, there is no overall best method for dealing with credit scoring problems. This is one of the main reasons why there exists an increasing interest in the use of classifier ensembles. Most of these works have demonstrated that the ensemble approach performs better than single classifiers when applied to credit scoring (Doumpos & Zopounidis, 2007; Hung & Chen, 2009; Twala, 2010; Wang, Hao, Ma, & Jiang, 2011; West, Dellana, & Qian, 2005). Despite the progress in this research field, several questions still remain open and should be addressed in order to fully understand the conditions under which the classifier ensembles can improve the model performance.

Taking these considerations into account, the present paper examines the use of seven well-known classifiers with five effective ensemble methods. The aim of this study is to find out what individual models are suitable for each ensemble strategy in the area of credit scoring. To this end, several experiments on six real credit data sets are carried out and the results are analysed for statistically significant differences by means of Friedman and Bonferroni-Dunn post hoc tests.

Hereafter, the paper is organized as follows. Section 2 gives a brief overview of the classifier ensemble approaches used in this study. Section 3 describes the set-up of the experiments carried out. Section 4 discusses the experimental results. Finally, Section 5 remarks the main findings and discusses future research directions.

2. Classifier ensembles

A classifier ensemble (also referred to as committee of learners, mixture of experts, multiple classifier system) consists of a set of individually trained classifiers (base classifiers) whose decisions are combined in some way, typically by weighted or unweighted voting, when classifying new examples (Kittler, 1998; Kuncheva, 2004). It has been found that in most cases the ensembles produce more accurate predictions than the base classifiers that make them up (Dietterich, 1997). Nonetheless, for an ensemble to achieve much better generalization capability than its members, it is critical that the ensemble consists of highly accurate base classifiers whose decisions are as diverse as possible (Bian & Wang, 2007; Kuncheva & Whitaker, 2003).

In statistical pattern recognition, a large number of methods have been developed for the construction of ensembles that can be applied to any base classifier. In the following sections, the ensemble approaches relevant for this study are briefly described.

2.1. Bagging

In its standard form, the bagging (Bootstrap Aggregating) algorithm (Breiman, 1996) creates M bootstrap samples $T_1, T_2, \ldots, T_M$ randomly drawn (with replacement) from the original training set T of size n. Each bootstrap sample T_i of size n is then used to train a base classifier C_i. Predictions on new observations are made by taking the majority vote of the ensemble C^* built from $C_1, C_2, \ldots, C_M$. As bagging resamples the training set with replacement, some instances may be represented multiple times while others may be left out.

Since each ensemble member is not exposed to the same set of instances, they are different from each other. By voting the predictions of each of these classifiers, bagging seeks to reduce the error due to variance of the base classifier.

2.2. Boosting

Similar to bagging, boosting also creates an ensemble of classifiers by resampling the original data set, which are then combined by majority voting. However, in boosting, resampling is directed to provide the most informative training data for each consecutive classifier.

The AdaBoost (Adaptive Boosting) algorithm proposed by Freund and Schapire (1996) constitutes the best known member in boosting family. It generates a sequence of base classifiers $C_1, C_2, \ldots, C_M$ by using successive bootstrap samples $T_1, T_2, \ldots, T_M$ that are obtained by weighting the training instances in M iterations. AdaBoost initially assigns equal weights to all training instances and in each iteration, it adjusts these weights based on the misclassifications made by the resulting base classifier. Thus, instances misclassified by model C_{i-1} are more likely to appear in the next bootstrap sample T_i. The final decision is then obtained through a weighted vote of the base classifiers (the weight w_i of each classifier C_i is computed according to its performance on the weighted sample T_i it was trained on).

2.3. Random subspace

The random subspace method is an ensemble construction technique proposed by Ho (1998), in which the base classifiers $C_1, C_2, \ldots, C_M$ are trained on data sets $T_1, T_2, \ldots, T_M$ constructed with a given proportion K of attributes picked randomly from the original set of features F. The outputs of the models are then combined, usually by a simple majority voting scheme.

This method may benefit from using random subspaces for both constructing and aggregating the classifiers. When the data set has many redundant attributes, one may obtain better classifiers in random subspaces than in the original feature space. The combined decision of such classifiers may be superior to a single classifier constructed on the original training data set in the complete feature space. On the other hand, when the number of training cases is relatively small compared with the data dimensionality, by constructing classifiers in random subspaces one may solve the small sample size problem.

2.4. DECORATE

Melville and Mooney (2005) introduced a new ensemble approach called DECORATE (Diverse Ensemble Creation by Oppositional Relabelling of Artificial Training Examples), which uses an existing learner to build an effective diverse committee in an iterative manner.

At each iteration, some artificial instances are randomly generated and combined with the original training data T in order to build a new ensemble member C_i. The labels for these artificially generated training instances are chosen so as to differ maximally from the current ensemble predictions, thereby increasing diversity when a new classifier is trained on the augmented data and added to the ensemble. While forcing diversity, it is still possible to maintain training accuracy by rejecting a new classifier if incorporating it into the existing ensemble decreases its performance.

2.5. Rotation forest

Rotation forest (Rodriguez, Kuncheva, & Alonso, 2006) refers to a technique to generate an ensemble of classifiers, in which each base classifier is trained with a different set of extracted attributes.

The main heuristic is to apply feature extraction and to subsequently reconstruct a full attribute set for each classifier in the ensemble. To this end, the feature set F is randomly split into K subsets, principal component analysis (PCA) is run separately on each subset, and a new set of linear extracted attributes is constructed by pooling all principal components. The data is transformed linearly into the new feature space. Classifier C_i is trained with this data set. Different splits of the feature set will lead to