Expert Systems with Applications 39 (2012) 3446–3453

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

An experimental comparison of classification algorithms for imbalanced credit scoring data sets

Iain Brown *, Christophe Mues

School of Management, University of Southampton, Highfield, Southampton SO17 1BJ, UK

ARTICLE INFO

Keywords:
Credit scoring
Imbalanced datasets
Classification
Benchmarking

ABSTRACT

In this paper, we set out to compare several techniques that can be used in the analysis of imbalanced credit scoring data sets. In a credit scoring context, imbalanced data sets frequently occur as the number of defaulting loans in a portfolio is usually much lower than the number of observations that do not default. As well as using traditional classification techniques such as logistic regression, neural networks and decision trees, this paper will also explore the suitability of gradient boosting, least square support vector machines and random forests for loan default prediction.

Five real-world credit scoring data sets are used to build classifiers and test their performance. In our experiments, we progressively increase class imbalance in each of these data sets by randomly undersampling the minority class of defaulters, so as to identify to what extent the predictive power of the respective techniques is adversely affected. The performance criterion chosen to measure this effect is the area under the receiver operating characteristic curve (AUC); Friedman's statistic and Nemenyi post hoc tests are used to test for significance of AUC differences between techniques.

The results from this empirical study indicate that the random forest and gradient boosting classifiers perform very well in a credit scoring context and are able to cope comparatively well with pronounced class imbalances in these data sets. We also found that, when faced with a large class imbalance, the C4.5 decision tree algorithm, quadratic discriminant analysis and *k*-nearest neighbours perform significantly worse than the best performing classifiers.

1. Introduction

The aim of credit scoring is essentially to classify loan applicants into two classes, i.e., good payers (i.e., those who are likely to keep up with their repayments) and bad payers (i.e., those who are likely to default on their loans). In the current financial climate, and with the recent introduction of the Basel II Accord, financial institutions have even more incentives to select and implement the most appropriate credit scoring techniques for their credit portfolios. It is stated in Henley and Hand (1997) that companies could make significant future savings if an improvement of only a fraction of a percent could be made in the accuracy of the credit scoring techniques implemented. However, in the research literature, portfolios that can be considered as very low risk, or low default portfolios (LDPs), have had relatively little attention paid to them in particular with regards to which techniques are most appropriate for scoring them. The underlying problem with LDPs is that they contain a much smaller number of observations in the class of defaulters than in that of the good payers. A large class imbalance is therefore present which some techniques may not be able to successfully handle. Typical examples of low default portfolios include high-quality corporate borrowers, banks, sovereigns and some categories of specialised lending (Van Der Burgt, 2007) but in some countries even certain retail lending portfolios could turn out to have very low numbers of defaults compared to the majority class. In a recent FSA publication regarding conservative estimation of low default portfolios, regulatory concerns were raised about whether firms can adequately asses the risk of LDPs (Benjamin, Cathcart, & Ryan, 2006).

A wide range of classification techniques have already been proposed in the credit scoring literature, including statistical techniques, such as linear discriminant analysis and logistic regression, and non-parametric models, such as *k*-nearest neighbour and decision trees. But it is currently unclear from the literature which technique is the most appropriate for improving discrimination for LDPs. Table 1 provides a selection of techniques currently applied in a credit scoring context, along with references showing some of their reported applications in the literature.

Hence, the aim of this paper is to conduct a study of various classification techniques based on five real-life credit scoring data sets. These data sets will then have the size of their minority class

* Corresponding author. Address: 44 Holters Mill, The Spires, Canterbury, Kent CT2 8SP, UK. Tel.: +44 (0) 7840057162.

E-mail addresses: i.brown@soton.ac.uk (I. Brown), C.Mues@soton.ac.uk (C. Mues).

0957-4174
doi:10.1016/j.eswa.2011.09.033

Table 1
Credit scoring techniques and their applications.

Classification techniques	Application in a credit scoring context
Logistic regression (LOG)	Arminger, Enache, and Bonne (1997), Baesens et al. (2003), Desai et al. (1996), Steenackers and Goovaerts (1989), West (2000), Wiginton (1980)
Decision trees (C4.5, CART, etc.)	Arminger et al. (1997), Baesens et al. (2003), West (2000), Yobas et al. (2000)
Neural networks (NN)	Altman (1994), Arminger et al. (1997), Baesens et al. (2003), Desai et al. (1996), West (2000), Yobas et al. (2000)
Linear discriminant analysis (LDA)	Altman (1968), Baesens et al. (2003), Desai et al. (1996), West (2000), Yobas et al. (2000)
Quadratic discriminant analysis (QDA)	Altman (1968), Baesens et al. (2003)
k-Nearest neighbours (*k*-NN)	Baesens et al. (2003), Chatterjee and Barcun (1970), West (2000)
Support vector machines (SVM, LS-SVM, etc.)	Baesens et al. (2003), Yang (2007)

of defaulters further reduced by decrements of 5% (from an original 70/30 good/bad split) to see how the performance of the various classification techniques is affected by increasing class imbalance.

The five real-life credit scoring data sets used in this empirical research study include two data sets from Benelux (Belgium, Netherlands and Luxembourg) institutions, the German Credit and Australian Credit data sets which are publicly available at the UCI repository (http://kdd.ics.uci.edu/), and the fifth data set is a behavioural scoring data set, which was also obtained from a Benelux institution.

The techniques that will be applied in this paper are logistic regression (LOG), linear and quadratic discriminant analysis (LDA, QDA), least square support vector machines (LS-SVM), decision trees (C4.5), neural networks (NN), nearest-neighbour classifiers (*k*-NN10, *k*-NN100), a gradient boosting algorithm and random forests. We are especially interested in the power and usefulness of the gradient boosting and random forest classifiers which have yet to be thoroughly investigated in a credit scoring context.

All techniques will be evaluated in terms of their area under the receiver operating characteristic curve (AUC). This is a measure of the discrimination power of a classifier without regard to class distribution or misclassification cost (Baesens et al., 2003).

To make statistical inferences from the observed difference in AUC, we followed the recommendations given in a recent article (Demšar, 2006) that looked at the problem of benchmarking classifiers on multiple data sets. The recommendations given were for a set of simple robust non-parametric tests for the statistical comparison of the classifiers (Demšar, 2006). The AUC measures will therefore be compared using Friedman's average rank test, and Nemenyi's post hoc test will be employed to test the significance of the differences in rank between individual classifiers. Finally, a variant of Demšar's significance diagrams will be plotted to visualise their results.

The organisation of this paper is as follows. Section 2 will begin by providing a literature review of the work that has been conducted on the topic of classification for imbalanced data sets. A brief explanation will then be given for the ten classification techniques to be used in the analysis of the data sets. Secondly, the empirical set up and criteria used for comparing the classification performance will be described. Thirdly, the results of our experiments are presented and discussed. Finally, conclusions will be drawn from the study and recommendations for further research work will be outlined.

2. Literature review

A wide range of different classification techniques for scoring credit data sets has been proposed in the literature, a non-exhaustive list of which was provided earlier in Table 1. In addition, some benchmarking studies have been undertaken to empirically compare the performance of these various techniques (e.g., Baesens et al., 2003), but they did not focus specifically on how these techniques compare on heavily imbalanced samples, or to what extent any such comparison is affected by the issue of class imbalance. For example, in Baesens et al. (2003) seventeen techniques including both well-known techniques such as logistic regression and discriminant analysis and more advanced techniques such as least square support vector machines were compared on eight real-life credit scoring data sets. Although more complicated techniques such as radial basis function least square support vector machines (RBF LS-SVM) and neural networks (NN) yielded good performances in terms of AUC, simpler linear classifiers such as linear discriminant analysis (LDA) and logistic regression (LOG) also gave very good performances. However, there are often conflicting opinions when comparing the conclusions of studies promoting differing techniques. For example, in Yobas, Crook, and Ross (2000), the authors found that linear discriminant analysis (LDA) outperformed neural networks in the prediction of loan default, whereas in Desai, Crook, and Overstreet (1996), neural networks were reported to actually perform significantly better than LDA. Furthermore, many empirical studies only evaluate a small number of classification techniques on a single credit scoring data set. The data sets used in these empirical studies are also often far smaller and less imbalanced than those data sets used in practice. Hence, the issue of which classification technique to use for credit scoring, particularly with a small number of bad observations, remains a challenging problem (Baesens et al., 2003).

The topic of which good/bad distribution is the most appropriate in classifying a data set has been discussed in some detail in the machine learning and data mining literature. In Weiss and Provost (2003) it was found that the naturally occurring class distributions in the 25 data sets looked at, often did not produce the best-performing classifiers. More specifically, based on the AUC measure (which was preferred over the use of the error rate), it was shown that the optimal class distribution should contain between 50% and 90% minority class examples within the training set. Alternatively, a progressive adaptive sampling strategy for selecting the optimal class distribution is proposed in Provost, Jensen, and Oates (1999). Whilst this method of class adjustment can be very effective for large data sets, with adequate observations in the minority class of defaulters, in some low default portfolios there are only a very small number of loan defaults to begin with.

Various kinds of techniques have been compared in the literature to try and ascertain the most effective way of overcoming a large class imbalance. Chawla, Bowyer, Hall, and Kegelmeyer (2002) proposed a synthetic minority over-sampling technique (SMOTE) which was applied to example data sets in fraud, telecommunications management, and detection of oil spills in satellite images. In Japkowicz (2000), over-sampling and downsizing were compared to the author's own method of "learning by recognition" in order to determine the most effective technique. The findings, however, were inconclusive but demonstrated that both