Expert Systems with Applications 37 (2010) 534–545

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A data driven ensemble classifier for credit scoring analysis

Nan-Chen Hsieh [a,*], Lun-Ping Hung [b]

[a] Department of Information Management, National Taipei College of Nursing, No. 365, Min-Ten Road, 11257 Taipei, Taiwan, ROC
[b] Department of Information Management, Technology and Science Institute of Northern Taiwan, No. 2, Xueyuan Road, Peitou, 112 Taipei, Taiwan, ROC

ARTICLE INFO

Keywords:
Clustering
Ensemble classifier
Neural network
Bayesian network
Class-wise classification
Credit scoring system

ABSTRACT

This study focuses on predicting whether a credit applicant can be categorized as good, bad or borderline from information initially supplied. This is essentially a classification task for credit scoring. Given its importance, many researchers have recently worked on an ensemble of classifiers. However, to the best of our knowledge, unrepresentative samples drastically reduce the accuracy of the deployment classifier. Few have attempted to preprocess the input samples into more homogeneous cluster groups and then fit the ensemble classifier accordingly. For this reason, we introduce the concept of class-wise classification as a preprocessing step in order to obtain an efficient ensemble classifier. This strategy would work better than a direct ensemble of classifiers without the preprocessing step. The proposed ensemble classifier is constructed by incorporating several data mining techniques, mainly involving optimal associate binning to discretize continuous values; neural network, support vector machine, and Bayesian network are used to augment the ensemble classifier. In particular, the Markov blanket concept of Bayesian network allows for a natural form of feature selection, which provides a basis for mining association rules. The learned knowledge is represented in multiple forms, including causal diagram and constrained association rules. The data driven nature of the proposed system distinguishes it from existing hybrid/ensemble credit scoring systems.

1. Introduction

In response to the recent growth of the credit industry and in the management of large loan portfolios, various kinds of credit scoring systems (Thomas, 2000) have been developed and applied successfully to support credit approval decisions. The main objective of credit scoring systems is to classify samples into affinity groups. Generally, credit scoring problems are related to classification by statistical methods. Investigating more sophisticated classifiers to match the characteristics of the samples is crucial in providing results that meet the needs of particular credit scoring applications.

Techniques for developing classifiers have evolved from simple parametric to nonparametric statistical methods. Altman (1968) applied simple parametric discriminant analysis and multiple discriminant analysis (MDA) to the corporate credit granting problem, then compared model's performance using linear discriminant analysis and neural networks (Altman, Marco, & Varetto, 1994). Lawrence and Arshadi (1995) used a logistic model to analyze the loan management problem using a series of borrower and bank variables. Charitou, Neophytou, and Charalambous (2004) used logit and neural network models to predict failed and non-failed UK public industrial firms over the period 1988–1997. These studies concluded that there certainly should be further studies and tests using statistical and artificial intelligence techniques and suggested a combined approach for predictive reinforcement.

Other efforts are leading to the investigation of nonparametric statistical methods for credit scoring applications. For example, McKee (1998) used rough sets for bankruptcy prediction and compared the performance of rough sets versus auditor signaling rates (McKee, 2003). A distinguished classifier, Bayesian networks (BN), has been proposed as a probabilistic white-box classifier which permits higher order relationships between the variables of the problem under study. Sarkar and Sriram (2001) utilized Bayesian networks for early warning of bank failures. They found the naïve Bayesian network and the composite attribute Bayesian network classifiers to have superior performance compared to the induced decision tree algorithm. Baesens et al. (2004) compared several Bayesian network classifiers with statistical and other artificial intelligence techniques for classifying customers, and concluded that Bayesian network classifiers offer an interesting and viable alternative for customer lifecycle slope estimation. Sun and Shenoy (2007) provided detailed operational guidance for building naïve Bayesian network classifiers for bankruptcy prediction.

Neural networks are the most common classifiers for credit scoring applications. Swales and Yoon (1992) used neural networks to

* Corresponding author.
E-mail addresses: nchsieh@ntcn.edu.tw (N.-C. Hsieh), robin@mail.tsint.edu.tw (L.-P. Hung).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.05.059

differentiate stocks, and found that neural networks outperformed significantly the linear multiple discriminant models. Tam and Kiang (1992) compared the neural network approach with linear classifier, logistic regression, *k*-NN, and ID3 to predict bank failures. They demonstrated that neural networks are more accurate, adaptive, and robust than other methods. Desai, Crook, and Overstreet (1996) found that the performance of discriminant analysis is comparable to the performance of back-propagation neural networks in classifying loan applicants into good and bad credit. They pointed out that more customized architectures might be necessary for building effective models to classify consumer loan applications in a credit union environment. West (2000) investigated the credit scoring accuracy of five neural network models, and reported that the nonparametric and hybrid design architectures are very useful in developing effective credit scoring systems.

In brief, the idea of conventional classifiers is generally based on a single classifier or a simple combination of these classifiers, showing moderate performance. However, even exquisitely designed classifiers still have deficiencies which cannot appropriately distinguish the samples. One way to alleviate the classifier and data match problem is to use an ensemble of classifiers. A variety of classifiers, either different types of classifiers or different instantiations of the same classifier, are combined before a final classification decision is made. Thus, the ensemble classifier allows different characteristics of the samples to be handled by classifiers suited to their particular needs, and provides an extra degree of bias/variance tradeoff.

Recently, the ensemble classifier has been demonstrated to be outperformed by a single classifier in having greater accuracies and smaller prediction errors when applied to the credit scoring data sets (Tsai & Wu, 2008; West, Dellana, & Qian, 2005). However, the potential to reduce the generalization error of a single classifier varies among different training data sets. This study investigates the concept of preprocessing the data set into more homogeneous cluster groups first and then fitting the ensemble classifier for stably predicting the categories of applicants. For this reason, we introduce the concept of class-wise classification as a preprocessing step in order to obtain an efficient ensemble classifier. This strategy would work better than a direct ensemble of classifiers without class-wise classification.

The meaning of effective ensemble classifier is twofold, relating to accuracy and easy interpretation of classified results. Prior work has already introduced ensemble classifiers to credit scoring applications. However, there is still improvement possible in proper guidance of the reclassified samples into homogeneous cluster groups, the optimal associate binning for the discretization of continuous values, the ensemble of various classifiers, and mining constrained association rules by means of classifiers. To make ensemble classifiers useful as a credit approval decision aids, this study proposes a novel ensemble classifier architecture. The overall method of the decision making process is shown in Fig. 1.

The overall method contains four phases. The control flow through different phases of the proposed method is: First, we start with a raw classification data set, then preprocess data set by applying binning and selection procedures (optimal associate binning, Pearson chi-square). Second, refine the quality of samples by class-wise classification. Third, increase the diversity and accuracy of individual classifiers in an ensemble classifier by class-wise bagging ensemble strategy. Fourth, provide user credit approval decision making with abundant information (ensemble classifier, causal diagram among features, association rules).

This study addresses the following research questions. First, a class-wise classification method is proposed to guide the reclassified samples into more homogeneous cluster groups that can be used to develop a well-performing ensemble classifier for credit scoring applications. Second, the proposed architecture focuses

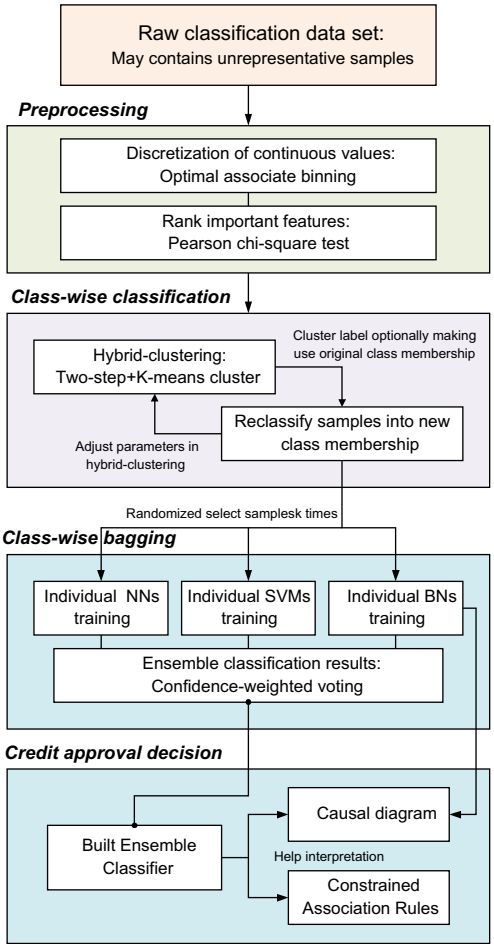

Fig. 1. Method overview. The overall method contains four phases. The control flow through different phases of the proposed method is: First, we start with a raw classification data set, then preprocess data set by applying binning and selection procedures (optimal associate binning, Pearson chi-square). Second, refine the quality of samples by class-wise classification. Third, increase the diversity and accuracy of individual classifiers in an ensemble classifier by class-wise bagging ensemble strategy. Fourth, provide user credit approval decision making with abundant information (ensemble classifier, causal diagram among features, association rules).

on fusing three types of classifiers – Neural network (NN), Bayesian network (BN), and Support vector machine (SVM) – which are simple to implement and have been shown to perform well in credit prediction (Huang, Chen, & Wang, 2007; Sarkar & Sriram, 2001). The rationale of employing these classifiers is that NNs are generally superior to the conventional classifiers, BNs can easily model complex relationships among variables, and SVMs can be used as the benchmark technique. Third, NNs and SVMs work well for continuous and discrete valued features, but BNs generally use only discrete valued features. Proper discretization of continuous values is critical for the building of a BN classifier. This study employs a heuristic method based on the assumption of linear dependence, as measured by correlations between variables to the target features, so an optimal associate binning technique for discretization of continuous values was gained. Through discretization, continuous values are converted into discrete values with several states. Fourth, the use of the same training data set for all individual classifiers may reduce the diversity among individual classifiers, while different training sets for individual classifiers may decrease the accuracy of individual classifiers. It is important to construct appropriate training data sets that maintain a good balance