Journal of Retailing and Consumer Services 27 (2015) 11–23

Contents lists available at ScienceDirect

Journal of Retailing and Consumer Services

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

A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring

Fatemeh Nemati Koutanaei [a], Hedieh Sajedi [b,*], Mohammad Khanbabaei [c]

[a] Department of Industrial Engineering, Science and Research Branch, Islamic Azad University, Saveh, Iran
[b] Department of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, Tehran, Iran
[c] Department of Information Technology Management, Science and Research Branch, Islamic Azad University, Tehran, Iran

ARTICLE INFO

Article history:
Received 15 December 2014
Received in revised form
30 June 2015
Accepted 4 July 2015
Available online 16 July 2015

Keywords:
Credit scoring
Classification
Feature selection
Ensemble learning
Data mining

ABSTRACT

Data mining techniques have numerous applications in credit scoring of customers in the banking field. One of the most popular data mining techniques is the classification method. Previous researches have demonstrated that using the feature selection (FS) algorithms and ensemble classifiers can improve the banks' performance in credit scoring problems. In this domain, the main issue is the simultaneous and the hybrid utilization of several FS and ensemble learning classification algorithms with respect to their parameters setting, in order to achieve a higher performance in the proposed model. As a result, the present paper has developed a hybrid data mining model of feature selection and ensemble learning classification algorithms on the basis of three stages. The first stage, as expected, deals with the data gathering and pre-processing. In the second stage, four FS algorithms are employed, including principal component analysis (PCA), genetic algorithm (GA), information gain ratio, and relief attribute evaluation function. In here, parameters setting of FS methods is based on the classification accuracy resulted from the implementation of the support vector machine (SVM) classification algorithm. After choosing the appropriate model for each selected feature, they are applied to the base and ensemble classification algorithms. In this stage, the best FS algorithm with its parameters setting is indicated for the modeling stage of the proposed model. In the third stage, the classification algorithms are employed for the dataset prepared from each FS algorithm. The results exhibited that in the second stage, PCA algorithm is the best FS algorithm. In the third stage, the classification results showed that the artificial neural network (ANN) adaptive boosting (AdaBoost) method has higher classification accuracy. Ultimately, the paper verified and proposed the hybrid model as an operative and strong model for performing credit scoring.

1. Introduction

Recently, banks and financial institutions have extensively started to consider the credit risk of their customers. In order to differentiate customers for offering credit services to them and managing their risks, banks have needed to apply credit scoring systems in their procedures (Gray and Fan, 2008). Lately, non-parametric approaches and data mining practices have been used in the area of customer credit scoring. The statistical methods, non-parametric methods, and artificial intelligence (AI) approaches have been suggested in order to provision the credit scoring developments. In addition, ensemble credit scoring methods have been used in many studies. It should be mentioned that a noticeable number of researches have shown that ensemble learning classification approaches in credit scoring have a better performance in comparison with single classifiers. With respect to the review of these studies, there are nine main approaches in credit scoring researches as provided in the following:

1. Single-classifier credit scoring models.
2. Multiple-classifier credit scoring models.
3. Credit scoring models based on statistical methods.
4. Credit scoring models based on AI methods.
5. Linear and non-linear credit scoring models.
6. Parametric credit scoring models, including linear probability model, discriminant analysis model, probit and logit models, etc.
7. Non-parametric (data mining) credit scoring models, including decision tree, K nearest-neighbor (KNN) model, expert system, ANN, fuzzy logic, GA, etc.
8. Ensemble learning credit scoring models.
9. Hybrid credit scoring models.

* Corresponding author.
E-mail address: hhsajedi@ut.ac.ir (H. Sajedi).

http://dx.doi.org/10.1016/j.jretconser.2015.07.003
0969-6989/

Many researchers have employed the above-mentioned approaches in their investigations. Hu and Ansell (2007) utilized some algorithms, including Naïve Bayes, logistic regression (LR), recursive partitioning, ANN, and sequential minimal optimization (SMO) in their study. In a study by Min and Lee (2008), they applied the credit scoring model based on data envelopment analysis (DEA). In another study, link analysis ranking method with the SVM was used for credit scoring (Xu et al., 2009). Setiono et al. (2009) used GA to optimize the KNN classification algorithm in credit scoring. Moreover, Yeh and Lien (2009) compared the data mining techniques, including KNN, LR, discriminant analysis, Naïve Bayes, ANN, and decision trees. Zhou et al. (2009) used direct search for parameters selection in the SVM classification algorithm. In a study by Ping and Yongheng (2011), neighborhood rough set and the SVM-based classifier were used for credit scoring. In another study (Kao et al., 2012), Bayesian latent variable model with classification regression tree was employed. Vukovic et al. (2012) used the preference theory functions in the case-based reasoning (CBR) model for credit scoring model. Danenas and Garsva (2015) applied particle swarm optimization (PSO) for the optimal linear SVM classifier selection in the domain of credit risk.

As cited above, recently, the ensemble credit scoring models have been used in a number of researches. Tsai and Wu (2008) applied multilayer perceptron (MLP) neural network ensembles for the credit scoring problem. In an investigation by Nanni and Lumini (2009), an ensemble of classifiers, including bootstrap aggregating (Bagging), Random Subspace, Class Switching, and Random Forest, was involved in the credit scoring. In addition, the ensemble of classifiers, including ANN, decision tree, Naïve Bayes, KNN, and logistic discriminant analysis was applied by Twala (2010). Hsieh and Hung (2010) utilized bagging ensemble classifier, including ANN, SVM, and Bayesian network. In another study by Paleologo et al. (2010), the subagging ensemble classifier, including kernel SVM, KNN, decision trees, AdaBoost, and subagged classifiers, was used in the credit scoring. Wang and Ma (2012) proposed a hybrid ensemble learning approach using SVM as a base learner for enterprise credit risk assessment.

Several studies have deployed the FS approach in their credit scoring models. Wang and Huang (2009) applied evolutionary-based FS approaches in a case study of credit approval data. Tsai (2009) compared five famous FS methods used in bankruptcy prediction, which were *t*-test, correlation matrix, stepwise regression, PCA, and factor analysis, in order to examine their performance by using MLP neural networks. Chen and Li (2010) proposed a combined strategy of FS approaches, including Linear Discriminant Analysis (LDA), rough set theory, decision tree, F-score and SVM classification model in credit scoring. In a research by Wang et al. (2012), the rough set and scatter search meta heuristic in FS were used for credit scoring. Chen (2012) developed an integrated FS and a cumulative probability distribution approach based on rough sets in credit rating classification. Hajek and Michalak (2013) suggested an approach to combine the mixed and individual FS methods with well-known machine learning models, such as MLP, radial basis function (RBF), SVM, Naive Bayes, random forest, LDA, and nearest mean classifier in corporate credit rating prediction. Oreski and Oreski (2014) presented a new hybrid GA with ANN to identify an optimum feature subset in order to increase the classification accuracy and scalability in credit risk assessments. Liang et al. (2015) deployed three filters including LDA, *t*-test, and linear regression, and two wrappers including GA and PSO based FS methods, combined with six different prediction models, namely linear SVM, RBF SVM, KNN, Naive Bayes, classification and regression tree (CART), and MLP under some experiments in bankruptcy and credit scoring datasets.

The above-mentioned studies have been stated from three main viewpoints as follows: (1) General credit scoring studies (2) Ensemble credit scoring studies (3) FS based credit scoring studies. This article is differentiated from the rest of the papers due to the simultaneous consideration of these three viewpoints. It is worth mentioning that past studies have considered only one or two of these viewpoints. In addition, it should be stated that the main aim of this article is to propose a proper FS algorithm and an appropriate base and ensemble classifier via three types of evaluation approaches, i.e., the SVM classification accuracy (only for FS), classification accuracy, and the area under the receiver operating characteristic curve (AUC) for classifiers and parameters setting (for both) in the context of hybrid credit scoring model. Moreover, many studies have not examined the effect of several FS methods and classifier parameter setting on the credit scoring problem. As another distinguished aspect, on the basis of the aforementioned attentions, in the present paper, nine approaches are combined in order to build a new hybrid FS and ensemble learning credit scoring model. The proposed model is a combination of FS techniques and several base (single) classifiers and ensemble classifiers in the parametric (owing to the Naïve Bayes algorithm) and non-parametric approaches of credit scoring. The parameters setting of four FS algorithms and two types of classification algorithms (base and ensemble) are used. For each FS algorithm, the performance is examined in terms of the SVM classification accuracy measure. The SVM is an influential learning method for classification problems. As cited by Brown and Mues (2012), "it is based on construction of maximum-margin separating hyper plane in some transformed feature space". It should be indicated that SVM is one of the most popular techniques used in the literature. Then, SVM is utilized to evaluate the performance FS algorithms. Moreover, the classification algorithms are compared according to the classification accuracy and AUC measures. For experimental results, the dataset of the 'Export Development Bank of Iran' is used. In the hybrid model, four FS algorithms are used as follows: (1) PCA; (2) GA; (3) Information gain ratio; and (4) Relief algorithm. Furthermore, two types of classification algorithms prevalent in the previous studies are as follows: (1) Base classification algorithms: Naïve Bayes, CART decision tree, SVM, and ANN; (2) Ensemble classification algorithms: bagging, AdaBoost, random forest, and staking. The results can confirm that the hybrid model of credit scoring has a robust functioning in comparison with the other classification algorithms presented in this paper.

As an abstract representation, the main contributions of this study reflected in the proposed model are as follows:

1. Providing a comprehensive study by comparing different FS methods and classifiers, with respect to the credit scoring problem.
2. Hybrid simultaneous use of three general, ensemble, and FS based credit scoring approaches.
3. Using FS algorithms and comparing their performance with the aid of the accuracy measure of the SVM classification algorithm and also the accuracy and AUC measures of the base and ensemble classifiers.
4. Employing the parameters setting procedures for FS and classification algorithms in order to improve the credit scoring performance with an iterative manner.
5. Simultaneous use and comparison of the base and ensemble learning classification algorithms in the proposed credit scoring model.