Expert Systems with Applications 38 (2011) 9313–9318

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Knowledge discovery using neural approach for SME's credit risk analysis problem in Turkey

Gülnur Derelioğlu [a,b,*], Fikret Gürgen [b]

[a] Yapı ve Kredi Bankası A.Ş., Information Technology Management, Güzeltepe Mevki, 41480, Şekerpınar, Kocaeli, Turkey
[b] Computer Eng. Dept., Bogazici University, Bebek 34342, İstanbul, Turkey

ARTICLE INFO

Keywords:
Credit risk analysis (CRA)
Small and medium enterprises (SMEs)
Multilayer perceptron (MLP)
Continuous/Discrete Rule Extractor via Decision Tree Induction (CRED)
Support vector machines (SVM)
k-Nearest Neighbor (*k*-NN)

ABSTRACT

This study proposes a knowledge discovery method that uses multilayer perceptron (MLP) based neural rule extraction (NRE) approach for credit risk analysis (CRA) of real-life small and medium enterprises (SMEs) in Turkey. A feature selection and extraction stage is followed by neural classification that produces accurate rule sets. In the first stage, the feature selection is achieved by decision tree (DT), recursive feature extraction with support vector machines (RFE-SVM) methods and the feature extraction is performed by factor analysis (FA), principal component analysis (PCA) methods. It is observed that the RFE-SVM approach gave the best result in terms of classification accuracy and minimal input dimension. Among various classifiers *k*-NN, MLP and SVM are compared in classification experiments. Then, the Continuous/Discrete Rule Extractor via Decision Tree Induction (CRED) algorithm is used to extract rules from the hidden units of a MLP for knowledge discovery. Here, the MLP makes a decision for customers as being "good" or "bad" and reveals the rules obtained at the final decision. In the experiments, Turkish SME database has 512 samples. The proposed approach validates the claim that is a viable alternative to other methods for knowledge discovery.

1. Introduction

Credit risk analysis (CRA) recently attracts more attention since credit volume in real market has shown great increase and economical fluctuations has become more often. Credit risk (CR) is a general term which implies to future losses. CRA aims to decrease future losses by estimating the potential risk and eliminating the new credit proposal if the risk is higher than a defined tolerance value. This is also called as CR classification which labels a customer as "good" if he could pay the loan back, otherwise as "bad".

In Turkey, 95% of real enterprises are accepted as small and medium enterprises (SME) that reveals the importance of them in national economy (Fantazzini & Figini, 2009). Not only in Turkey, but also in many developing countries in the world, especially in recent years, SME credits have been gaining much more importance according to their high growth in financial world. In contrast with its increasing growth rate in the world-wide financial sector, there is a limited research for SMEs' CRA.

In our study, we highlight the potential of neural networks as tools of knowledge discovery in SMEs CRA problem in Turkey. A trained multilayer perceptron (MLP) makes a credit decision as being "good" or "bad" of customers and our neural rule extraction algorithm discovers the knowledge embedded in the MLP. First, various input attribute selection methods are used to select a minimal input dimension. Although, small subset of the original SME portfolio is used as an input, when the real, huge portfolio size is taken into consideration; the dimension reduction becomes indispensable phase. Decision Tree (DT) and Recursive Feature Elimination with Support Vector Machine (SVM-RFE) are applied for feature selection, and then Factor Analysis (FA) and Principal Component Analysis (PCA) are applied for feature extraction. In the experiments, six permutations of this data set were obtained and they were labeled as Dataset1, Dataset2, . . . , Dataset6. Then, the purpose of the study is to show that for the six datasets, it is possible to achieve good accuracies. More importantly, this can be achieved by neural network with one hidden layer and having enough connections which makes it possible to extract comprehensible rules using our rule extraction algorithm, Continuous/Discrete Rule Extractor via Decision Tree Induction (CRED). We first present the results from applying MLP, SVM and *k*-NN classifiers on the six datasets. After the good accuracy rates were achieved, we address the issue of knowledge discovery from the neural networks. For this purpose, we obtain rules from the CRED algorithm using trained MLP with both continuous and discrete features. The work flow of our proposed method is given in Fig. 1.

* Corresponding author. Address: Ornek M G:46/2(Tahir)S No3/7, Atasehir,34704, İstanbul, Turkey. Tel.: +90 5052757605.

E-mail addresses: gulnurderelioglu@gmail.com (G. Derelioğlu), gurgen@boun.edu.tr (F. Gürgen).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.01.012

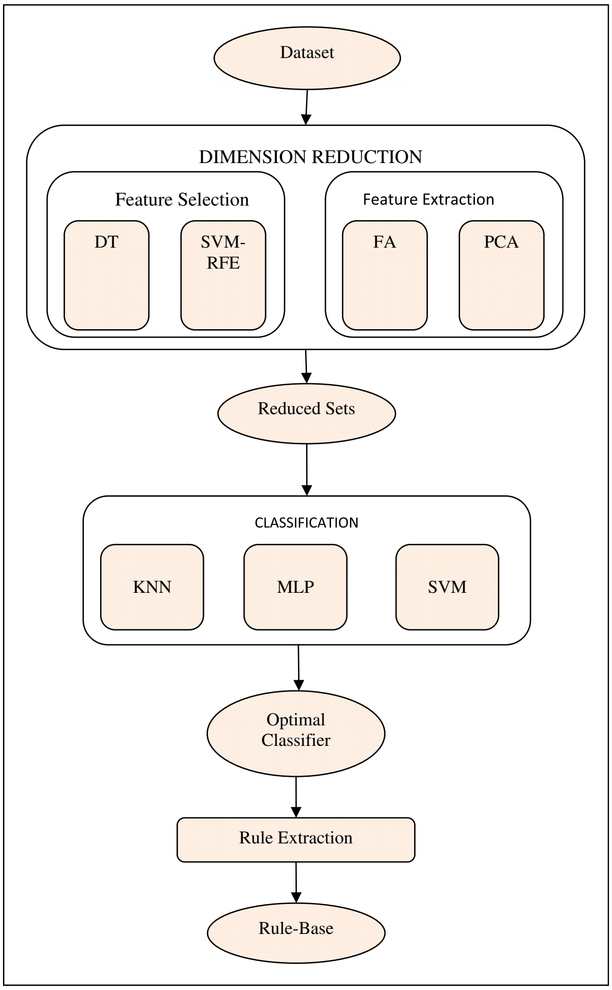

Fig. 1. Rule base obtained from trained MLP on Dataset2.

1.1. Previous work

CRA is an appealing topic where a 1% improvement in accuracy, which seems insignificant, will reduce losses in a large loan portfolio and save billions of euro. Thus, there have been many techniques proposed for CRA: Neural Networks (NN), SVM, logistic regression (LR) and *k*-NN algorithms are widely used on different dataset and provided good prediction accuracy. Although many researches (Gaganis, Pasiouras, Spathis, & Zopounidis, 2007; Galindo & Tamayo, 2000; Kaya, Gürgen, & Okay, 2008; Wei, Li, & Chen, 2007; Yang & Duan, 2008; Zhou & Bai, 2008) have been done on CRA for large corporates, personal credits and credit cards, there are only a few works for SME CRA. SMEs show different behavior than corporate and individuals. Altman and Sabato remind that analysis on German and French SMEs show that SMEs are riskier with lower assets correlation with each other than large corporate. Thus, developed models for corporate would not be suitable for SMEs. Altman and Sabato developed a default prediction model based on LR using US SMEs' data and indicated that their proposed method outperformed Multivariate Discriminant Analysis. Fantazzini and Figini (2009) also proposed Random Survival Forest Model which gave slightly better performance than classical LR model on the real-life dataset obtained from Creditreform.

1.2. Dataset

The real-life dataset is provided by Yapı ve Kredi Bankası A.Ş. which consists of SMEs information collected from credit portfolio until 2007. Only a small subset of the original portfolio is taken randomly, without any sampling methodology, thus we do not affirm that the dataset reveals all portfolios behavior perfectly.

Dataset has 512 samples with 27 features and a binary class variable either good or bad. Class distribution is not homogeneous as 144 customers (28%) were good and 368 customers were bad. Six of the features are categorical and the others are continuous variables. These features mainly cover four different types of information: demographical, financial, risk and delinquency information.

The remainder of this paper is organized as follows: Firstly, in Section 2 input attribute selection and extraction techniques are outlined, credit risk classification methodologies are described in Section 3, rule-base extraction is provided in Section 4. Lastly, experimental results are discussed in Section 5.

2. Input feature selection and extraction

In data mining applications, the time and space complexity of any classifier or regressor directly depends on the input data size (Alpaydin, 2004). Dimensionality reduction techniques can be applied to the input data to obtain a reduced representation of the dataset without losing the integrity of the original data (Han & Kamber, 2001). These techniques can generally be divided into two groups: feature selection aims to obtain a subset of the original dataset with a minimum loss of information which is also called as subset selection. Feature extraction aims to find a new set of features that are the combinations of the original variables.

2.1. Feature selection

We employ DT and SVM-RFE techniques for feature selection.

2.1.1. Decision tree

DT learning is based on divide-and-conquer and generates hierarchical rules. It starts at the root node by splitting the input space into two subspaces and continues recursively with the corresponding subset until a leaf node is obtained. Learning ends when the best split is reached. The measure of the good split is impurity which is determined as if all instances of the branch are labeled as the same class. The final univariate tree consists of the most relevant features and discards irrelevant ones. In our work, J48, which is a C4.5 tree (Quinlan, 1993) implemented in WEKA, is used as the first feature selection method.

2.1.2. Recursive feature elimination with support vector machine

RFE-SVM utilizes the generalization capability embedded in SVM. RFE keeps the independent features containing the original dataset information while eliminating weak and redundant features (Chen & Jeong, 2007). However, the subset produced by SVM-RFE is not necessarily the ones that are individually most relevant. Only taken together the features of a produced subset are optimal informative (Guyon, Weston, Barnhill, & Vapnik, 2002). The working methodology of SVM-RFE is based on backward selection where algorithm starts with whole features and iteratively eliminates the worst one until the predefined size of the final subset is reached at each iteration, and the remaining features must be ranked again (Thang, Zhang, & Huang, 2007). We apply Weka SVM-RFE algorithm with square of weight as ranking criterion and at each step the feature which causes minimum variation in the SVM cost function is removed from feature space.