Expert Systems with Applications 38 (2011) 12939–12945

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A genetic algorithm-based approach to cost-sensitive bankruptcy prediction

Ning Chen [a,*], Bernardete Ribeiro [b], Armando S. Vieira [a], João Duarte [a], João C. Neves [c]

[a] GECAD, Instituto Superior de Engenharia do Porto, Instituto Politecnico do Porto, Rua Dr. Antonio Bernardino de Almeida, 431, 4200-072 Porto, Portugal
[b] CISUC, Department of Informatics Engineering, University of Coimbra, Rua Silvio Lima-Polo II, Coimbra 3030-790, Portugal
[c] ISEG-School of Economics, Technical University of Lisbon, Rua do Quelhas 6, Lisbon 1200-781, Portugal

ARTICLE INFO

Keywords:
Neural network
Learning vector quantization
Classification
Cost-sensitive learning
Feature selection
Genetic algorithm

ABSTRACT

The prediction of bankruptcy is of significant importance with the present-day increase of bankrupt companies. In the practical applications, the cost of misclassification is worthy of consideration in the modeling in order to make accurate and desirable decisions. An effective prediction system requires the integration of the cost preference into the construction and optimization of prediction models. This paper presents an evolutionary approach for optimizing simultaneously the complexity and the weights of learning vector quantization network under the symmetric cost preference. Experimental evidences on a real-world data set demonstrate the proposed algorithm leads to significant reduction of features without the degradation of prediction capability.

1. Introduction

Bankruptcy prediction is a widely studied research topic in financial analysis due to the increasing tendency of bankrupt enterprises and deepening financial crisis nowadays. For example, in Portugal the bankruptcy of enterprises represented an increase of 49% during the year 2009 compared to the previous year. The prediction systems which can identify the risk of failures correctly are important to bank decision and early warning. In the field of bankruptcy prediction, two errors are mostly concerned, namely type I error and type II error. The former stands for classifying a bankrupt company as an insolvent one, which results in the cost of losing principal and interest. While the latter stands for classifying a solvent company as a bankrupt one, which results in the cost of losing profit. As is well known, the two costs are usually asymmetric and should be considered in the practical application to make a tradeoff between the two errors. Various studies focus on the modification of conventional classification algorithms for the purpose of incorporating the cost preference into the classification. Some researchers argue that by specifying an appropriate cost-relevant objective function, the classification can be regarded as an optimization problem and solved by evolutionary algorithms. Genetic algorithm (GA) gained rapid popularity and proved to be effective in optimizing the linear discriminant analysis model, support vector machine, back-propagation neural networks, etc. However, few studies illustrate the integration of GA and learning vector quantization (LVQ) for the purpose of cost-sensitive bankruptcy prediction.

It has been shown that LVQ with genetically evolved connected weights outperforms a modified LVQ which integrates the cost information into learning methodology (Chen, Ribeiro, Vieira, Duarte, & Neve, 2010). The idea is to enhance the LVQ training through the global search of genetic algorithm with respect to the appropriate fitness function. Since only the connected weights are optimized without taking the feature selection and parameter determination into consideration, the optimal solution may be missed. As pointed out by many evidences, feature selection plays an important role in classification in terms of improving the predictive accuracy and decreasing the complexity of models. Additionally, the resultant predictive model is somewhat dependent on the parameters employed. GA provides the facility to simultaneously optimize the factors that potentially impact on the performance so that it does not require the prior knowledge about the important features and the number of units needed to represent the classes. In this paper, we present a genetic algorithm-based approach to integrate the connected weight optimization, parameter determination and feature selection in an evolutionary procedure. The cost preference is directly incorporated into the fitness function of the genetic algorithm for performance evaluation. The performance of the proposed algorithm is investigated on the real-life data. The obtained results demonstrate that the reduction of features contributes to the improvement of prediction capability.

The rest of this paper is organized as follows. Section 2 reviews the previous work of bankruptcy prediction, cost-sensitive learning

* Corresponding author. Tel.: +351 228340500; fax: +351 228321159.
E-mail addresses: ningchen74@gmail.com (N. Chen), bribeiro@dei.uc.pt (B. Ribeiro), asv@isep.ipp.pt (A.S. Vieira), jmmd@isep.ipp.pt (J. Duarte), jcneves@iseg.utl.pt (J.C. Neves).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.04.090

and feature selection. Section 3 presents the framework of a GAFS-LVQ algorithm. Section 4 describes the experimental design and results. The last section summarizes the paper with contributions and future research issues.

2. Related work

2.1. Bankruptcy prediction

Bankruptcy prediction has profound impact on bank decisions and profitability. The main concern of interest is to construct the prediction model representing the relationship between the bankruptcy and financial ratios and then deploy the model to identify the high risk of failure in the future. In the literature, bankruptcy prediction was solved by statistical methods and machine learning methods. Statistical methods comprise discriminant analysis, logistic model, factor analysis, etc. Machine learning methods include neural network, decision tree, support vector machine, case-based reasoning, fuzzy logic, rough set, hybrid and ensemble approach (Ravi & Ravi, 2007). Among them, neural network is one of the most widely applied tool and its capability has been proved by a large variety of work. Vector quantization (VQ) forms a quantized approximation of input vectors through a finite number of prototypes (connected weights). LVQ is a supervised variant of VQ and useful for complicated non-linear separation problems (Kohonen, 2001). The network is composed of two levels, in which the input level is fully connected with the output level. The modeling technique is based on the neurons representing prototype vectors and the nearest neighbor classification rule. The goal of learning is to determine the weights that best represent the classes. LVQ has been employed to detect the distressed companies with satisfactory performance (Neves Boyacioglu, Kara, & Baykan, 2009; Chen & Vieira, 2009; Neves & Vieira, 2006). In this paper, we use the LVQ model for cost-sensitive bankruptcy prediction.

2.2. Cost-sensitive learning

Cost-sensitive learning addresses the challenging classification problem in which there are different costs associated with different errors. Compared with most existing classification methods which aim to minimize the total number of errors, cost-sensitive learning assigns the cost to the errors and intends to minimize the total cost of errors. Generally, cost-sensitive learning is performed on the data level or the algorithm level. The first approach performs as a preprocessing (or a postprocessing) phase of error-based classifiers for general-purpose, such as stratification which changes the frequency of classes in the training data (Chan & Stolfo, 1998), MetaCost which re-labels the training samples with their estimated minimal-cost classes (Domingos, 1999), and threshold-moving which moves the output towards the expensive class (Pendharkar, 2008). The second approach explicitly incorporates the cost information into the learning methodology. The existing algorithms include decision tree, regularized least square, boosting learning, back-propagation neural network, and mathematical programming (Koh, 1992; Ling, Yang, Wang, & Zhang, 2004; Pendharkar & Nanda, 2006; Sun, Kamela, Wong, & Wang, 2007; Vo & Won, 2007). These methods are implemented by adapting the learning methodology to asymmetric cost preference. Evolutionary algorithms are a promising approach to cost-sensitive learning, which can be represented as an optimization problem. The results reported in Nanda and Pendharkar (2001) show that a genetic-based approach that incorporates the asymmetric cost preference in the linear discriminant analysis model leads to desirable results.

Some efforts have been undertaken to make LVQ cost-sensitive in both data level and algorithm level. In Chen, Vieira, and Duarte (2009), the cost matrix is integrated with basic LVQ algorithm using standard sampling and threshold-moving techniques. In Chen, Vieira, Duarte, Ribeiro, and Neves (2009), a cost-LVQ algorithm is presented based on the modification of a batch LVQ algorithm. The cost information is incorporated into the model when performing the update of map neurons so that the instances of expensive class are harder to be misclassified. A hybrid algorithm (Chen et al., 2010) employs genetic algorithm to optimize the connected weights of a LVQ model. The prototypes are coded as the input to genetic evolution and optimized through the genetic operators. The superiority of this approach is demonstrated compared to the local search strategy.

2.3. Feature selection

In the field of bankruptcy prediction, a large number of indicators are usually involved so that the training data is insufficient to cover the decision space, which is called as the curse of dimensionality. Feature selection addresses the problem by removing irrelevant, redundant and correlated features, improving the accuracy and compactness of classification model, decreasing the computational effort, and facilitating the use of models.

Feature selection is basically an optimization problem which searches through the space of feature subsets to identify the relevant features. The previous studies can be divided into two categories, namely filter approach and wrapper approach. Filter approach selects the features based on desirable properties before model construction. Despite the computational efficiency, filter approach ignores the induction algorithms and is prone to unexpected failures (Fogel, 2000). Wrapper approach embeds the feature selection into the model learning and searches for an optimal solution to the particular classifier employed. Several searching strategies are deployed, including greed backward and forward technique and evolutionary technique. Evolutionary technique produces superior and more reliable results than greedy technique which does not consider the correlation among features. An embedded feature selection paradigm is implemented in the weight training procedure for neural network (Castellani & Marques, 2008). Genetic algorithm is also used to improve the performance of support vector machine (SVM) in both feature subset selection and parameter optimization (Min, Lee, & Han, 2006). Besides, swarm intelligence represented by ant colony optimization and particle swarm optimization was employed for solving the feature selection problem and enhancing the prediction capability of machine learning models (Lin, Ying, Chen, & Lee, 2008; Marinakis, Marinaki, Doumpos, & Zopounidis, 2009). In this paper, we embed a wrapper feature selection and LVQ learning into a genetic algorithm framework.

3. Simultaneous optimization of feature selection and parameter determination using genetic algorithm

The complexity of prediction models is a critical problem relevant to underfitting or overfitting. Regarding LVQ, the complexity comprises the number of weights (number of units) and the size of weights (subset of features). A feasible way is to integrate the complexity optimization with weight optimization in the framework of GA. GA is an evolutionary technique to find the optimal or near optimal solution to optimization problems. The population of solutions is generated randomly and evolved towards optimization under the direction of fitness function. GA has been extensively applied to various combination optimization problems in the conjunction with machine learning methods. In an evolutionary LVQ, GA is used to discover the right number of prototypes needed to represent the classes (Cordella, Stefano, Fontanella, & Marcelli, 2006). In this section, we present a GA-based approach