Expert Systems with Applications 41 (2014) 2052–2064

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Genetic algorithm-based heuristic for feature selection in credit risk assessment

Stjepan Oreski *, Goran Oreski

Bank of Karlovac, I.G.Kovacica 1, 47000 Karlovac, Croatia

ARTICLE INFO

Keywords:
Artificial intelligence
Genetic algorithms
Classification
Credit risk assessment
Incremental feature selection
Neural network

ABSTRACT

In this paper, an advanced novel heuristic algorithm is presented, the hybrid genetic algorithm with neural networks (HGA-NN), which is used to identify an optimum feature subset and to increase the classification accuracy and scalability in credit risk assessment. This algorithm is based on the following basic hypothesis: the high-dimensional input feature space can be preliminarily restricted to only the important features. In this preliminary restriction, fast algorithms for feature ranking and earlier experience are used. Additionally, enhancements are made in the creation of the initial population, as well as by introducing an incremental stage in the genetic algorithm. The performances of the proposed HGA-NN classifier are evaluated using a real-world credit dataset that is collected at a Croatian bank, and the findings are further validated on another real-world credit dataset that is selected in a UCI database. The classification accuracy is compared with that presented in the literature. Experimental results that were achieved using the proposed novel HGA-NN classifier are promising for feature selection and classification in retail credit risk assessment and indicate that the HGA-NN classifier is a promising addition to existing data mining techniques.

1. Introduction

Credit risk is one of the most important issues in the banking industry; therefore, credit risk assessment has gained increasing attention in recent years (Akkoç, 2012; Danenas et al., 2011; Finlay, 2011; Tsai, Lin, Cheng, & Lin, 2009). Until a few years ago, the body of research on consumer credit risk assessment was quite sparse. Quantitative consumer credit risk assessment models were developed much later than those for business credit, mainly due to the problem of data availability. Data were limited to the databases of financial institutions. Currently, some data are publicly available in several countries, and financial institutions and researchers have developed many different quantitative credit scoring techniques (Šušteršic, Mramor, & Zupan, 2009). Still, there is no standard set of attributes or indicators that would exist in all credit institutions and on the basis of which the classification of retail customers in terms of credit worthiness could be conducted. Therefore, it is necessary to use all of the available data and information, methods and algorithms for feature selection and the precise classification of clients.

Forced by a crisis, banks are exposed to challenges in finding new ways of doing business that must be less risky and entirely efficient and profitable. They are forced by the crisis because the crisis has revealed the level of risks that are embedded in the banking business. In retail, risks were very often taken with no exact estimate of their degree and possible consequences. The large number of decisions that are involved in the consumer lending business make it necessary to rely on models and algorithms rather than on human discretion and to base such algorithmic decisions on "hard" information (Khandani, Kim, & Lo, 2010). In terms of the high growth of the economy, banks achieved high rates of profit for their owners and were not exposed to the challenges of finding substantially new ways of doing business.

The crisis has significantly reduced the profit margin; many banks have run into difficulties, and some have gone into bankruptcy. Investors have become increasingly cautious and not willing to invest their capital in such troubled banks, and regulators require banks to strengthen their capital (BIS, 2011), raising their resistance to such emergencies. State governments and international organisations have become involved in rescuing the situation and preventing even larger consequences of the crisis.

The crisis was deep because one of the fundamental causes of the crisis was the way that banks functioned. Few exact methods have been used in assessing the retail risk, and taking collateral for borrowed funds has been used as a major surrogate. When it became obvious that this collateral was not as valuable as its prior assessed value and that the credit risk of the clients was not appropriately assessed, the losses became inevitable. We should work on

* Corresponding author. Tel.: +385 98 246 328; fax: +385 47 614 306.
E-mail addresses: stjepan.oreski@kaba.hr (S. Oreski), goran.oreski@gmail.com (G. Oreski).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.09.004

all of the causes of the crisis and even on the way that banks do business, primarily in terms of taking a risk when lending funds. Operations should be faster, less risky, more exact and based on data. Banks should use the capital at their disposal in a better way. This capital is not only in terms of money but is also in terms of the data collected in their databases. The capital in the form of customer data should be managed better by the banks. It should be transformed into knowledge and ultimately money.

The data in the databases can be used for credit risk assessments, but these data are commonly high dimensional. Irrelevant features in a training dataset could produce less accurate results in the classification analysis. Feature selection is required to select the most significant features to increase the predictive accuracy, speed and scalability. In difficult learning problems, such as in credit risk assessment, using the appropriate set of features is critical for the success of the learning process and therefore, by itself, is an important issue. Hence, we investigate the possibilities for feature selection methods that provide increased accuracy and scalability of the algorithms and that enable incremental feature selection. A novel and efficient hybrid classifier is designed here.

The present research is focused on the genetic algorithm (GA) and its capabilities for enhancement. The enhancement of the genetic algorithm involves the prevention of spending time in exploring irrelevant regions of the search space. Therefore, the theme of this paper is the advanced heuristic algorithm creation by the hybridisation of the genetic algorithm with some of the filter techniques. The novel classifier, called HGA-NN, is composed of fast filter techniques, a hybrid genetic algorithm (HGA) and an artificial neural network. Research was conducted on solving the problems of feature selection and classification when assessing retail credit risks.

The remaining sections of this paper are organised as follows. Section 2 describes the problem of feature selection to be studied in the paper and reviews the previous literature related to the problem. A brief overview of the genetic algorithm design is given in the third section. Section 4 describes the experimental design and model development. Section 5 discusses the experimental results with performance evaluation and comparison. Section 6 concludes this paper and provides guidelines for future work.

2. Problem statement and literature review

Feature selection is a pre-processing technique that is commonly used on high-dimensional data, and its purposes include reducing the dimensionality, removing irrelevant and redundant features, facilitating data understanding, reducing the amount of data needed for learning, improving the predictive accuracy of the algorithms, and increasing the interpretability of the models (Oreski, Oreski, & Oreski, 2012). Feature selection is the problem of choosing a small subset of features that ideally is necessary and sufficient for describing the target concept (Kira & Rendell, 1992). When the feature selection is poorly performed, it could lead to problems that are associated with incomplete information, noisy or irrelevant features, and not the best set of features, among other problems. The learning algorithm that is used is slowed down unnecessarily due to the large number of dimensions of the feature space, while also experiencing lower classification accuracies due to learning irrelevant information.

The problem of the selection of m features from a set of n features can be solved with different algorithms. From the perspective of the processor time necessary for solving the problem, the computational complexity of this problem is $\binom{n}{m}$ and belongs to the class of NP problems. For larger dimensions, these problems cannot be solved by means of exhaustive search or simple heuristics. In recent years, various feature selection algorithms (techniques) have been proposed. Some of them will be mentioned below.

Aha and Bankert (1996) report positive empirical results with forward and backward sequential feature selection algorithms. They show that feature selection improves the performance of classifiers and provides evidence that wrapper models outperform filter models. Danenas et al. (2011) applied feature selection for datasets by using a correlation-based feature subset selection algorithm with Tabu search for search in attribute subsets. Jin et al. (2012) proposed the attribute importance measure and selection method based on attribute ranking. In the proposed attribute selection method, input output correlation is applied for calculating the attribute importance, and then the attributes are sorted in descending order. The hybrid of Back Propagation Neural Network (BPNN) and Particle Swarm Optimisation (PSO) algorithms is also proposed. PSO is used to optimise weights and thresholds of BPNN for overcoming the inherent shortcoming of BPNN. Their experimental results show that the proposed attribute selection method is an effective preprocessing technique.

Piramuthu (2006) considers decision support tools for credit-risk evaluation from a machine learning perspective. He discusses a few means of improving the performance of these tools through data preprocessing, specifically through feature selection and construction. He stated simply that one must take data and/or problem characteristics as well as the suitability of a given algorithm to obtain better performance. Performance, in this context, depends on at least two different entities: the algorithm and the dataset.

All of these feature selection techniques can be divided into three groups: filter, wrapper and hybrid techniques. The filter techniques rely on the general characteristics of the data to evaluate and select attribute subsets without involving a classification algorithm. One advantage of filter techniques is that because they do not use the classification algorithm, they are usually fast and therefore suitable for use with large datasets. Additionally, they are easily applicable to various classification algorithms. The wrapper techniques first implement an optimising algorithm that adds or removes attributes to produce various subset attributes and then employ a classification algorithm to evaluate this subset of attributes. The wrapper techniques are known to be more accurate compared to the filter techniques, and they are computationally more expensive. Because the classification algorithm is called repeatedly, wrapper techniques are slower than filter techniques and do not scale up well to large, high-dimensional datasets. The hybrid techniques attempt to take advantage of the filter and wrapper techniques by exploiting their complementary strengths (Jin & et al., 2012).

Hybrid techniques are usually a combination of filter and wrapper techniques and are designed to trade the accuracy with the computational speed by applying a wrapper technique to only those subsets that are preselected by the filter technique (Jin et al., 2012). The strategies used for searching the feature space in hybrid techniques are very different. Because of the time complexity of the problem, meta-heuristics are often used. One of the meta-heuristics is GAs. The advantage of GAs compared with other search algorithms is that more strategies can be adopted together to find good individuals to add to the mating pool in a GA framework, in both the initial population phase and the dynamic generation phase (Pezzella, Morganti, & Ciaschetti, 2007). Recently, various variants of GAs have been proposed.

Yang, Li, and Zhu (2011)) describe an improved genetic algorithm for optimal feature subset selection from a multi-character feature set (MCFS). They divide the chromosome into several segments according to the number of feature groups in MCFS for local management. A segmented crossover operator and a segmented mutation operator are employed to operate on these segments to avoid invalid chromosomes. The probability of crossover and