Expert Systems with Applications 39 (2012) 8389–8395

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

A case-based reasoning model that uses preference theory functions for credit scoring

Sanja Vukovic *, Boris Delibasic, Ana Uzelac, Milija Suknovic

University of Belgrade, Faculty of Organizational Sciences, Jove Ilica 154, Belgrade, Serbia

ARTICLE INFO

Keywords:
Case-based reasoning
Preference functions
Genetic algorithm
Credit scoring
Classification

ABSTRACT

We propose a case-based reasoning (CBR) model that uses preference theory functions for similarity measurements between cases. As it is hard to select the right preference function for every feature and set the appropriate parameters, a genetic algorithm is used for choosing the right preference functions, or more precisely, for setting the parameters of each preference function, as to set attribute weights. The proposed model is compared to the well-known *k*-nearest neighbour (*k*-NN) model based on the Euclidean distance measure. It has been evaluated on three different benchmark datasets, while its accuracy has been measured with 10-fold cross-validation test. The experimental results show that the proposed approach can, in some cases, outperform the traditional *k*-NN classifier.

1. Introduction

The number of risks banks face every day is constantly growing. Banks are financial organizations which turn risk into profits. Most of their revenues are generated by credits, i.e. lending operations. Credit lending is therefore one of the most important profit generators for a bank. Naturally, the main risk when granting a loan is that the clients will not be able to fulfil their obligations towards the bank and that the bank will lose its funds.

During the last couple of decades, a rapid growth has been noticed in both availability and use of credits. Until recently, the decision to grant a credit was based on human judgement to assess the risk of default (Thomas, 2000). The growth in the demand for credit, however, has led to a larger interest in the use of more formal and objective methods (generally known as credit scoring) to help credit providers decide whether or not to grant credit to an applicant (Akhavein, Frame, & White, 2005; Chye, Chin, & Peng, 2004).

Credit scoring is a classification problem. That is why credit scoring models help to decide whether to grant a credit to new applicants, considering the customer's characteristics such as age, income and marital status (Chen & Huang, 2003). Credit granting is a very important part of a bank activity, as it may yield big profits, but there is also a significant risk involved in making decisions in this area and the mistakes may be very costly for financial institutions (Zakrzewska, 2007).

For all the above reasons, the decision-making related to credit granting is one of the crucial elements in the policy of each bank. The key problem is to distinguish between good (that surely repay) and bad (that likely default) credit applicants. This means that credit risk evaluation consists of building classification rules that properly define bank customers as good or bad payers (Zakrzewska, 2007).

For many years, the decision whether to grant a loan has been done by credit analysts. The analysts usually had to write down the rules they used to assess a loan applicant's credibility in repaying the loan. Credit decisions were made using these rules.

Credit scoring methodology can be used for different purposes, such as credit cards, personal loan applications, home loans, small business loans, as well as insurance applications and renewals. Furthermore, it can be used to increase the response rate to advertising campaigns, etc. (Thomas, 2000). Therefore, it is essential to find a way to build an effective customer classification model that can predict the customer's behaviour more accurately.

There are a number of models which can be used for credit evaluation in the banking industry. Some of these methods are statistical, while some of them rely on artificial intelligence (AI) approaches. The statistical methods often used for credit scoring are multiple regression (e.g. Meyer & Pifer, 1970), discriminant analysis (e.g. Altman, 1968; Banasik, Crook, & Thomas, 2003), and logistic regression (e.g. Desai, Crook, & Overstreet, 1996; Dimitras, Zanakis, & Zopounidis, 1996; Elliott & Filinkov, 2008; Lee, Chiu, Lu, & Chen, 2002; Martin, 1977), while the AI methods include inductive learning (e.g. Han, Chandler, & Liang, 1996; Shaw & Gentry, 1998), artificial neural networks (e.g. Boritz & Kennedy, 1995; Coakley & Brown, 2000; Jo & Han, 1996; Lee & Chen, 2005; West, 2000; Zhang, Hu, Patuwo, & Indro, 1999), genetic algorithms (e.g.

* Corresponding author.
E-mail addresses: sanja.vukovic@gmail.com (S. Vukovic), boris.delibasic@fon.bg.ac.rs (B. Delibasic), ana.uzelac@sf.bg.ac.rs (A. Uzelac), milija.suknovic@fon.bg.ac.rs (M. Suknovic).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.01.181

Desai, Convay, Crook, & Overstreet, 1997; Huang, Chen, & Wang, 2007; Huang, Tzeng, & Ong, 2006; Yobas, Crook, & Ross, 2000), and artificial immune system (e.g. Leung, Cheong, & Cheong, 2007).

Since the seminal work of Schank and Abelson (1977), case-based reasoning (CBR) has been successfully applied in many areas including credit scoring (e.g. Bryant, 1997; Buta, 1994; Shin & Han, 2001; Wheeler & Aitken, 2000).

CBR is one of the methods which can be successfully applied to financial problems such as credit scoring. It can also be used in many other areas, such as customer segmentation (e.g. Changchien & Lin, 2005; Chiu, 2002; Chun & Park, 2006), medical and manufacturing industry (e.g. Hsu, Chiu, & Hsu, 2004; Im & Park, 2007; Tseng, Chang, & Chang, 2005), etc.

Despite its many advantages, there are some problems that must be solved in order to design an effective CBR system (Ahn & Kim, 2008):

- How to select the appropriate similarity function to generate classification from stored cases?
- How to select the appropriate features, known as feature selection?
- How to determine the weight of each feature, which is known as feature weighting?
- How to determine the optimal k parameter if k-nearest neighbour (k-NN) algorithm is used?
- How to compute similarity for categorical variables which could, among the numerical variables, also describe cases?

There have been many studies attempting to resolve these problems. The selection of the appropriate similarity measures, and the choice of feature subsets and their weights in the case of the retrieval step have been the most popular research issues (e.g. Ahn, Kim, & Han, 2007; Chiu, Chang, & Chiu, 2003; Kim & Han, 2001; Liao, Zhang, & Mount, 1998; Shin & Han, 1999; Wang & Ishii, 1997).

Similarity measurement is an important part of every CBR model. Therefore, it is usually used the Euclidean metric. Preference theory can also be used for measuring similarity between cases, especially as it provides more opportunities in expressing the decision-makers' preferences. Li, Sun, and Sun (2009) and Li and Sun (2010) proposed combining CBR with preference functions (outranking relations) for financial distress and business failure prediction, respectively.

Li et al. (2009) used the outranking relations based on the preference function in Electre III, while Li and Sun (2010) constructed a hybrid CBR forecasting system which used all the available preference functions in outranking approaches, such as Electre, Promethee, and Oreste. In this paper, we have used the preference functions proposed in method Promethee for measuring similarity between cases.

Li and Sun (2010) used a trial-and-error iterative process to identify the optimal hybrid CBR module with corresponding preference function and parameters. In this paper, genetic algorithm has been used for such purposes.

It has been interesting to consider whether the domain knowledge, which can be expressed through preference functions, could be better exploited in such a way to improve the predictive performance of a CBR system.

The main difference between CBR with preference functions and the traditional CBR (k-NN) is the mechanism of similarity computation. We hypothesize that the use of preference theory functions in CBR can show better results than the traditional k-NN, based on the Euclidean distance measure in loan granting.

We have also analysed the number of neighbours that are taken into account for classification, as well as the influence of attribute (feature) weights on the accuracy.

The remainder of this paper is organized as follows: Section 2 introduces the basic concepts of the methods used in this paper. Section 3 proposes a new hybrid method that combines CBR and preference functions, with support of GA. Experimental evaluation of the model on benchmark credit scoring data is also presented in this section. Finally, some concluding remarks and the ideas for future research are discussed in Section 4.

2. Basic concepts

In this study, we evaluate the usefulness of CBR model with combining preference functions and a genetic algorithm (GA). The first two parts of this section present a review of basic concepts of CBR, with the emphasis on case retrieval, as this is the main topic of consideration in this investigation. The third part describes the main types of preference functions. The final, fourth part, describes the GA setup that has been used in this paper.

2.1. Case-based reasoning

Credit scoring methodology requires experience-based expertise. When solving a new problem, the experts rely on the past scenarios. They need to know which credits have been successful and which have failed. They also need to know how to modify an old case to fit the new situation. CBR is a general paradigm for experience-based reasoning. It assumes a memory model for representing, indexing, and organizing the past cases and a process model for retrieving and modifying the old cases and assimilating the new ones (Slade, 1991).

CBR solves new problems by relating some previously solved problems or experiences to the new problems thus forming analogical inferences for problem solving (Kolodner & Mark, 1992). Facing a new problem, CBR retrieves similar cases stored in a case base and adapts them to the new problem. The key factors affecting the performance of a CBR retrieval mechanism are case representation, case indexing and similarity metric (Buta, 1994).

CBR is generally quite simple to implement and can often handle complex and unstructured decisions very effectively (Ahn et al., 2007).

The retrieval of relevant previous cases is crucial to the success of a CBR system. The aim of case-based retrieval is to retrieve the most useful previous cases towards the optimal resolution of a new case and to ignore those previous cases that are irrelevant (Montazemi & Gupta, 1997).

A good retrieving function should take into account the features of a case that are more important. The case that matches the important features but not the less important ones will almost certainly be a better match than the one that matches less important features but does not match the important ones. For this reason, the integration of domain knowledge into the case matching and retrieving function is highly recommended in modelling a successful CBR system (Park & Han, 2002).

Solving a problem by CBR involves obtaining a problem description, measuring the similarity of the current problem to the previous problems stored in a case base (or memory) with their known solutions, retrieving one or more similar cases and attempting to reuse the solution of one of the retrieved cases, possibly after adapting it to account for differences in problem descriptions. The solution proposed by the system is then evaluated (e.g., by being applied to the initial problem or assessed by a domain expert). Then, if the proposed solution is adequate the problem description and its solution can be retained as a new case, and the system has learned to solve a new problem (Lopez de Mantaras et al., 2005).

According to Kolodner (1993), CBR comprises four major steps: (1) case representation, (2) case indexing, (3) case retrieval, and (4)