Information Sciences 236 (2013) 174–185

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

Application of hybrid case-based reasoning for enhanced performance in bankruptcy prediction

Chun-Ling Chuang *

Department of Information Management, Kai-Nan University, No. 1, Kainan Road, Luzhu, Taoyuan 338, Taiwan, ROC

ARTICLE INFO

Article history:
Received 10 November 2010
Received in revised form 29 January 2013
Accepted 10 February 2013
Available online 26 February 2013

Keywords:
Decision analysis
Data mining
Rough set theory
Business failure prediction
Grey relation analysis
Classification and regression tree

ABSTRACT

Predicting business failure is an important and challenging issue that has served as an impetus for many academic studies over the past three decades. This study aims at developing CBR-based hybrid models of predicting business failure. The need to supplement CBR (Case-Based Reasoning) with other classification and diagnosis techniques is triggered by the fact that accuracy and effectiveness tend to get reduced when CBR alone is applied to handle too many attributes. To enhance the accuracy of bankruptcy prediction, the hybrid models developed by this study include: RST–CBR (combining Rough Set Theory with CBR), RST–GRA–CBR (integrating RST, Grey Relational Analysis, and CBR), and CART–CBR (combining Classification and Regression Tree with CBR). In order to verify the ability of the proposed models to achieve optimal accuracy rate, this study further compares the predictive ability of CBR with those of other comparative models.

1. Introduction

Bankruptcy prediction is one of the most important issues influencing financial and investment decision-making [41]. For example, banks were facing higher financial risks in 2008 [65]. Business failures not only cripple the operation of corporations or organizations but may further undermine social functions and agitate economical structures [56]. Therefore, the evaluation of business failure has emerged as a professional area prompting academics and professionals to develop effective prediction models based on their specific interests for practical applications in the business community.

Methods developed, such as univariate statistical models, Multiple Discriminant analysis, Linear Probability models, Logistic Regression (LR), and Probit analysis [7,39,40], have been proven to be very effective [49]. However, those methods require quite a few assumptions and are usually constrained by their demand for data linearity; it is therefore fairly difficult for statistical methods to deal with massive and complicated data [4,42]. To get rid of the cumbersome requirements of statistical methods, such as a large number of samples, normal distribution of independent variables, and linear relationship between all variables, scholars in related fields have come up with alternatives for predicting the risk of business failure, notably Artificial Neural Network (ANN), Case-Based Reasoning (CBR), and Data Mining techniques [11,24,44,47,53,59].

ANN has been reported to perform fairly well in terms of predictive accuracy. However, ANN classifies every object by its computational characteristic and is therefore often criticized as a "black box" approach for its lack of transparency [28]. In contrast to the "black box" scenario, CBR, developed by Schank and Abelson in the 1970s [50], uses "similar" cases stored in a case base to build a solution to a new case based on the adaptation of solutions to past similar cases [26]. Literature indicates that CBR has been extensively applied to the diagnosis domain to improve on the deficiencies of ANN and statistical models [6,43]. However, it should be noted that the success of a CBR model depends on its ability to index cases and retrieve the

* Tel.: +886 3 3412500x6107; fax: +886 3 3412173.
E-mail address: clchuang@mail.knu.edu.tw

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.02.015

most relevant ones in support of the solution to a new case. Case retrieval consists of subtasks, referred to as identifying features, initial matching, searching, and selecting [1]. Many attributes will suffer from weak discriminating power [61–63]. Identifying important attributes has been studied in different fields [64,66]. Feature (attribute) selection, critical to CBR performance, is to facilitate the efficient organization and accurate retrieval by removing irrelevant or redundant features to keep the retrieved cases relevant to the target. As a result, the presence of too many attributes can reduce both the effectiveness and accuracy of CBR.

To solve the problem of unnecessary or excessive attributes, RST can be adopted to find the significant attributes. Attributes, once identified, may need to be further prioritized by taking their weights into consideration, and this task can be accomplished by performing Grey Relational Analysis (GRA) that compares quantitative analysis to the development between every variable in the grey system dynamically.

Another potential reinforcement to CBR is Classification and Regression Tree (CART) that has been successfully applied to solve classification problems. CART has been adopted as a popular alternative to classical methods like discriminant analysis, linear regression, and logistic regression [18]. Not only does it indicate that a particular object belongs to a specific class when the built rules are satisfied, but also it points out which variables are important in classifying objects [32]. All the three techniques – RST, CART, and GRA – can be adopted to reinforce the performance of CBR in predicting business failure.

Three CBR-based hybrid models are accordingly proposed. The first model (RST–GRA–CBR) proposed by Lin et al. [35] incorporates three stages starting with applying RST to extract key attributes, proceeding to employ GRA to obtain the weights of the key attributes, and culminating in feeding the retrieved key attributes and weights into CBR. The second model combines RST and CBR with equal weights (RST–CBR). The third applies CART as attribute selector and data pre-processor for CBR (CART–CBR) to strive for greater accuracy. In order to verify the ability of the proposed models to achieve optimal accuracy rate, this study compares the predictive ability of CBR with those of the hybrid RST–CBR, RST–GRA–CBR, and CART–CBR models. Moreover, prediction by LR (Logistic Regression) that is selected as a representative of classical methods has also been performed and the results are compared.

Following the present section as an introduction, four major sections are incorporated in the paper. Section 2 outlines the backgrounds to RST, GRA, CART, and CBR. Section 3 explains the research models. Section 4 centers on experiment results and model evaluations while Section 5 prespents our conclusions and offers remarks and suggestions for future researches.

2. Research background

2.1. Rough Set Theory (RST)

Studies have shown that RST has played an important role in dealing with classification problems [13,14,53]. As a powerful mathematical tool introduced by Pawlak [45], RST can discover facts from imperfect data [57] and make significant contribution to knowledge reduction [68]. RST has been successfully applied to real-world classification tasks having the following advantages [53]: (a) identifying potentially important information contained in a large number of cases and reducing the information into a model including a generalized description of the knowledge; (b) requiring no interpretation as the model is a set of easily understandable decision rules; and (c) demanding no additional information, such as probability in statistics.

Many researchers [13,21,37,38] have used RST to construct a prediction model or combined it with other methods as a data pre-processor. In particular, when RST is further integrated with other methods, such as the proposed hybrid models, the results tend to be more accurate than those obtained by using RST alone [37]. The following sections present some basic concepts of RST.

A finite set of objects U and a finite set of attributes A can be considered as a relational system $S = (U, A)$. Each attribute a belongs to the considered set of attributes $A(a \in A)$; $a(x)$ represents the value of attribute a of object x. For every set of attributes $B \in A$, an indiscernibility relation $Ind(B)$ is defined in the following way:

$$Ind(B) = \{(x_i, x_j) \in U;\ b(x_i) = b(x_j);\ b \in B\} \tag{1}$$

The equivalence class of x_i in relation to $Ind(B)$ is represented as $[x_i]_{Ind(B)}$. Let X denote a subset of elements of U, the upper and lower approximations can be represented as

$$\overline{B}(X) = \{x_i \in U | [x_i]_{Ind(B)} \cap X \neq \phi\} \tag{2}$$

$$\underline{B}(X) = \{x_i \in U | [x_i]_{Ind(B)} \subseteq X\} \tag{3}$$

The boundary of X in U is defined as $Bnd(X) = \overline{B}(X) - \underline{B}(X)$. If $Bnd(X)$ is an empty set, then the set X is definable with respect to B; otherwise, X will be referred to as a rough set. "Reduct" and "core" are two fundamental RST concepts. Assume that the set of attributes $R \subseteq A$ depends on the set of attributes $B \subseteq A$ in S (denotation $B \to R$) iff $Ind(B) \subseteq Ind(R)$. The minimal subset is $R \subseteq B \subseteq A$ such that $\mu_B F = \mu_R F$[1] is called F-reduct of B, which means a reduct is the minimal subset of attributes to represent the whole set of attributes. In addition, S might have more than one F-reduct. Intersection of all F-reducts is called F-core of B, i.e. $CORE_B(F) = \cap RED_B(F)$. The core is a collection of the most significant attributes in the system. In other words, the attributes in B

[1] $\mu_B F$ and $\mu_R F$ are called the quality of approximation of partition F by set of attributes B and R.