PERGAMON

Omega 29 (2001) 561–576

omega
The International Journal of Management Science

www.elsevier.com/locate/dsw

Variable precision rough set theory and data discretisation: an application to corporate failure prediction

Malcolm J. Beynon *, Michael J. Peel

Cardiff Business School, Aberconway Building, Colum Drive, Cardiff, CF10 3EU, UK

Received 18 February 2000; accepted 6 June 2001

Abstract

Since the seminal work of Pawlak (International Journal of Information and Computer Science, 11 (1982) 341–356) rough set theory (RST) has evolved into a rule-based decision-making technique. To date, however, relatively little empirical research has been conducted on the efficacy of the rough set approach in the context of business and finance applications. This paper extends previous research by employing a development of RST, namely the variable precision rough sets (VPRS) model, in an experiment to predict between failed and non-failed UK companies. It also utilizes the FUSINTER discretisation method which neglates the influence of an 'expert' opinion. The results of the VPRS analysis are compared to those generated by the classical logit and multivariate discriminant analysis, together with more closely related non-parametric decision tree methods. It is concluded that VPRS is a promising addition to existing methods in that it is a practical tool, which generates explicit probabilistic rules from a given information system, with the rules offering the decision maker informative insights into classification problems. © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: Data mining; Failure prediction; FUSINTER data discretisation; Rough set theory; Variable precision rough set theory

1. Introduction

Since the nascence of computerisation, together with the evolution of Artificial Intelligence (AI), there has been an explosion in the application of advanced decision-making techniques to solving business problems [1–5]. Following the pioneering study of Altman [6], who used multivariate discriminant analysis (MDA) to differentiate between failed and non-failed US firms, a large body of research has focused on corporate failure prediction (see [7–10] for literature reviews). The prediction of corporate failure continues to be viewed as a matter of considerable interest to both academics and practitioners (including credit and investment analysts), and has obvious importance for the stakeholders (investors, creditors, employees, managers) of a firm. This is evidenced by the recent application of neural networks (NNs), recursive partitioning algorithm (RPA) and case based reasoning to this issue [11–15]. A key advantage of these contemporary methods over their traditional counterparts (such as MDA and logit analysis) is that they do not require pre-specification of a functional form, nor the adoption of restrictive assumptions concerning the distributions of model variables and errors [12,16,10].

More recently, a further non-parametric technique, rough set theory (RST), which has its foundations in mathematical set theory, has been applied to decision problems [17,18]. RST was originated by Pawlak [19] and has been described as 'a new mathematical tool to deal with vagueness and uncertainty. This approach seems to be of fundamental importance to AI and cognitive sciences, especially in the areas of machine learning, knowledge acquisition, decision analysis, knowledge discovery from databases, expert systems, decision support systems, inductive reasoning and pattern recognition ... One of the advantages of RST is that it does

* Corresponding author. Tel./fax: +44-29-2087-5747.
E-mail address: beynonmj@cardiff.ac.uk (M.J. Beynon).

PII: S0305-0483(01)00045-7

not need preliminary or additional information about data, such as probability distributions in statistics, basic probability assignment in the Dempster Shafer theory of evidence, or grade of membership of the value of possibility in fuzzy set theory' [20, p. 89].

RST incorporates the use of indiscernibility (equivalence) relations to approximate sets of objects by upper and lower set approximations and, as noted by Slowinski and Zopounidis [21, p. 79], 'it is a formal framework for discovering deterministic and non-deterministic rules from a given representation of knowledge ... [it] ... assumes knowledge representation in a decision table form which is a special case of an information system'. Initial RST applications focused on medical diagnosis, drug research and process control [22,23], but more recently it has been extended to cover credit fraud detection, stock market rule-generation, market research, climate change and the development of expert systems for the NASA space centre [24,25,20].

Slowinski and Zopounidis [21] also investigated the use of RST to assess the risk of a Greek bank's clients (firms) in terms of granting finance. Although they did not examine the predictive accuracy of the RST rules, they did conclude [21, p. 39] that (based on financial ratios and other firm-specific variables), RST 'is a useful tool for discovering a preferential attitude of the decision maker in multi-attribute sorting problems'. More recently, Dimitras et al. [26, p. 278] reported that (on the basis of financial ratios) a rough set approach to predicting between failed and non-failed Greek firms 'was generally better than those obtained by classical discriminant and logit models'. A limitation of these studies is that the continuous data used to derive the rough set rules, have been discretised (a requirement of RST) with the aid of a selected 'expert'. Clearly different experts may proffer different views and the operational costs and complexities of using RST (and related techniques) will increase when there is over-reliance on an expert. In this context An et al. [27, p. 647] have stated that 'It has to be emphasised ... that the question of how to optimally discretise the attribute (variable) values, is unsolved, and a subject of on-going research'. This paper therefore employs a new (and more objective) discretisation method, namely the FUSINTER technique. However, the motivation for data discretisation extends beyond the requirements of RST, to include discretising data of an imprecise quality ('noisy' data). The ability to formulate rules from interval data (via discretisation) may also facilitate a more informed understanding of the interaction of the characteristics of objects. In this context, it is of interest to note that, even with regard to traditional statistical estimators (logit/discriminant analysis), it has recently been advocated that continuous variables (financial ratios) should be rank-transformed to improve their distributional properties in a failure prediction setting [28].

A further RST innovation has been the development by Ziarko [29] of a variable precision rough sets (VPRS) model, which incorporates probabilistic decision rules. This is an important extension, since as noted by Kattan and Cooper [30, p. 468], when discussing computer based decision techniques in a corporate failure setting, 'In real world decision making, the patterns of classes often overlap, suggesting that predictor information may be incomplete... This lack of information results in probabilistic decision making, where perfect prediction accuracy is not expected'.

An et al. [27] applied VPRS (which they termed 'Enhanced RST') to generating probabilistic rules to predict the demand for water. Relative to the traditional rough set approach, VPRS has the additional desirable property of allowing for partial classification compared to the complete classification required by RST. More specifically, when an object is classified using RST it is assumed that there is complete certainty that it is a correct classification. In contrast, VPRS facilitates a degree of confidence in classification, invoking a more informed analysis of the data, which is achieved through the use of a *majority inclusion* relation [29].

This paper extends previous work by providing an empirical exposition of VPRS, where we present the results of an experiment which applies VPRS rules to the corporate failure decision. In addition, we mitigate the impact of using the subjective views of an expert (as employed in previous studies) to discretise the data, by utilising the sophisticated FUSINTER discretisation technique which is applied to a selection of attributes (variables) relating to companies' financial and non-financial characteristics. The discretised data, in conjunction with other nominal attributes, are then used in this new VPRS framework to identify rules to classify companies in a failure setting.

To facilitate a comparison of our experimental VPRS results with those of existing techniques, we present the predictive ability of classical statistical methods—logit analysis and MDA—together with two more closely related non-parametric decision-tree methods, RPA and the Elysee method, which utilises ordinal discriminant analysis (see [15,31], for an exposition of these methods). However in the spirit of previous experimental research—and more particularly the previous failure prediction study of Frydman et al. [15, p. 239], who concluded that 'we feel that the attributes of new techniques like RPA can be presented and evaluated in a rigorous framework without the necessity of proving its absolute superiority over existing procedures'—the comparative classification results are not meant to be definitive, but rather to illustrate the potential of VPRS. In this context, research on the criteria to select the most efficacious and parsimonious set of VPRS rules (for predictive purposes) is still in its infancy [27].

The remainder of the paper is organised as follows: The next section gives a brief exposition of the VPRS method and a discussion of the FUSINTER discretisation method. The results of the empirical experiments are then reported, including a discussion of the predictive ability of VPRS relative to other existing parametric and non-parametric methods.