Chapter 11
An Exposition of Feature Selection and Variable Precision Rough Set Analysis: Application to Financial Data

Malcolm J. Beynon
Cardiff University, UK

Benjamin Griffiths
Cardiff University, UK

ABSTRACT

This chapter considers, and elucidates, the general methodology of rough set theory (RST), a nascent approach to rule based classification associated with soft computing. There are two parts of the elucidation undertaken in this chapter, firstly the levels of possible pre-processing necessary when undertaking an RST based analysis, and secondly the presentation of an analysis using variable precision rough sets (VPRS), a development on the original RST that allows for misclassification to exist in the constructed “if ... then ...” decision rules. Throughout the chapter, bespoke software underpins the pre-processing and VPRS analysis undertaken, including screenshots of its output. The problem of US bank credit ratings allows the pertinent demonstration of the soft computing approaches described throughout.

INTRODUCTION

There are a number of types of information industries, providing information on a wide range of areas such as, scientific, technical, medical, media, and relevant to this chapter, business and financial information (Fayyad et al., 1996). In his 1984 book Megatrends, Naisbitt (pp. 24) wrote, “We are drowning in information, but starved for knowledge...”, this is a sentiment which still holds true over 20 years later.

Utilising modern technology (computers), the process by which knowledge is extracted from databases of information is commonly known as data mining, and is seen as a major step of the broader discipline of Knowledge Discovery in Databases (KDD). Where knowledge management, in a business context, is the process by which companies organise, collect and assimilate this knowledge into their systems (Zorn and Taylor, 2003). Due to
the volume of data available, and facilitated by the advances made in modern computers, new techniques are being developed, both in industry and academia, to exploit this increasing abundance of information. Tay et al. (2003, pp. 1) notes that:

*A new generation of techniques and tools is emerging to intelligently assist humans in analyzing mountains of data, finding useful knowledge and in some cases performing analysis automatically...*

This chapter considers one of the more nascent data mining methods, namely Variable Precision Rough Sets (VPRS) (Ziarko, 1993a), an extension of Rough Set Theory (RST) (Pawlak, 1982), within the field of quantitative financial analysis. The financial analysis in this case is with respect to the classification and prediction of banks to Fitch’s Individual Bank Strength Ratings (Fitch, 2007).

Mitra et al. (2002) in a review of the impact of soft computing in data mining, state (pp. 3):

*Soft computing methodologies (involving fuzzy sets, neural networks, genetic algorithms, and rough sets) are most widely applied in the data mining step of the overall KDD process.*

Their review specifically suggests RST has emerged as a major mathematical tool for managing uncertainty, which arises from granularity in the domain of discourse, and has proved to be useful in a variety of KDD processes. Further, it (RST) offers mathematical tools to discover hidden patterns in data and therefore its importance, as far as data mining is concerned, can in no way be overlooked.

In this chapter, bespoke, software is described, which incorporates a suite of facilities capable of tackling some of the most relevant issues within the field KDD and data mining, and pertinent applies VPRS, to produce sets of decision rules associated with the data being analysed. Due to their contemporary nature, there are no strict definitions of KDD and data mining. Frawley et al. (1992, pp. 58) described KDD as the, “non-trivial extraction of implicit, unknown, and potentially useful information from data”. Although the terms, KDD and data mining, are often used synonymously (data mining is considered to be the more popular term, Piatetsky-Shapiro, 2000), a clear distinction can be drawn, as stated by Fayyad et al. (1996, pp. 39):

*KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from data...*

Amongst the facilities considered is the pre-processing of the data (as a prerequisite for the later data mining stage), which includes methods for, handling missing data, tackling imbalanced data, and discretisation of continuous valued data into discrete data (a requirement of some data mining methods, including RST). Often, data sets contain variables (also described as features or attributes) that are redundant or irrelevant, hence, a feature selection facility may be required prior to data mining, considered here also.

These facilities associated with KDD, could be described as supportive processes for the main stage (facility), that is, data mining, here undertaken through the employment of VPRS (an extension of RST). Developed by Pawlak (1982, 1991), RST is a set theoretical approach for dealing with imprecise or vague concepts within knowledge. With regards to data mining, it offered a new methodology for, construction and application of decision tables, based on inconsistent data sets (Ziarko, 2003). VPRS is an extension of RST, which relaxes an assumption that the given classifications within the data set are totally correct (Mi et al., 2004). Hence, VPRS allows for a level of uncertainty, which may be inherent within some data sets (Ziarko, 1993a). This was an important development, since it enabled the construction of probabilistic decision rules when using VPRS,