Applied Soft Computing 11 (2011) 3671–3679

Contents lists available at ScienceDirect

Applied Soft Computing

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

# Compactness rate as a rule selection index based on Rough Set Theory to improve data analysis for personal investment portfolios

Jhieh-Yu Shyng[a,*], How-Ming Shieh[b], Gwo-Hshiung Tzeng[c,d]

[a] Department of Information Management, Lan-Yang Institute of Technology, No. 79, Fu-Shin Rd, To-Chen, I-Lan 621, Taiwan
[b] Department of Business Administration, National Central University, No. 300, Chung-da Rd., Chung-Li City 320, Taiwan
[c] Department of Business and Entrepreneurial Management, Kainan University, No. 1, Kainan Rd., Luchu, Taoyuan 338, Taiwan
[d] Institute of Management of Technology, National Chiao Tung University, 1001 Ta-Hsueh Road, Hsinchu 300, Taiwan

ARTICLE INFO

*Article history:*
Received 10 November 2008
Received in revised form 11 August 2010
Accepted 30 January 2011

*Keywords:*
Rough Set Theory (RST)
Compactness rate
Strength rate
Support
Investment portfolio

ABSTRACT

This study proposes a selection index technique, namely a compactness rate based on Rough Set Theory (RST), for improving data analysis, eliminating data amount and reducing the number of decision rule. This study uses an empirical real-case involving a personal investment portfolio to demonstrate the proposed method. The presented case includes 75 rules generated by the RST. The rules are vague and fragmentary, making it very difficult to interpret the information. Many rules have the same strength and number of support objects and condition parts. These are creating a critical problem for decision making. The new method proposed in this study not only enables the selection of interesting rules, but it also reduces the data amount, and offers alternative strategies that can help decision-makers analyze data.


## 1. Introduction

Real-world data can suffer incompleteness and inconsistency. Data preprocessing techniques can improve data quality as well as the accuracy and efficiency of subsequent mining. Data preprocessing is an important step in knowledge delivery, since quality decisions require quality data. Early detection of data anomalies and reducing the amount of data requiring analysis can improve decision making. A database may contain data objects that do not comply with the general data behaviour or model. These objects are outliers.

Data mining can help business managers find and reach more suitable customers, and gain critical business insights that can help increase market share and profits. Decision rules, generated from data mining, can provide business managers with information on market competition.

Recently, research on attitudes towards personal wealth has increased and can be found in various places, including The Wall Street Journal [22], Dalal Street Investment Journal [23], and finance reports [4,17]. A well-designed financial plan can help optimize asset allocation and meet customer needs. Asset management is closely linked to personal experience and behaviour. Researchers are increasingly interested in customer retention and relationship marketing, as well as how firms can create profitable relationships with clients. Such relationships are crucial to the success of financial institutions, and recognise the ongoing nature of relationships between firms and their clients and the longevity of many financial products.

Specific areas that have attracted research interest include portfolio method [9], the behaviour of financial services consumers [5], management of personal finances [17], and retirement plans [4], and the assessment of the impacts of customer satisfaction and relationship quality on customer retention [6]. The main influences on investor decision-making regarding their personal asset allocation are the risk level and revenue of investment products which relate the timings of the purchase and sale of portfolio components.

Knowledge is usually acquired from observed data especially business data, which was a valuable resource for researchers and decision-makers. A number of personal portfolio studies have focused on quantification of the problem such as streamlining the parameters and statistically analyzing the data. However, any study of personal portfolios should consider the personal backgrounds and perspectives of investors. The application of the personal background, personal perspective and personal asset allocation decisions involves the following challenges: quality problems, ambiguous information and non-numerical data. These challenges make it difficult to use standard methods of applying statistical tools for knowledge discovery and rule induction.

* Corresponding author. Tel.: +886 2 27492556.
*E-mail addresses:* shyng@mail.fit.edu.tw (J.-Y. Shyng), hmshieh@mgt.ncu.edu.tw (H.-M. Shieh), ghtzeng@cc.nctu.edu.tw, ghtzeng@mail.knu.edu.tw (G.-H. Tzeng).

1568-4946/$ – see front matter. 
doi:10.1016/j.asoc.2011.01.038

Discovered hidden information from real financial data to make intelligent business decisions has been an important issue in recent research. Formalism in knowledge representation is important in helping users understanding the meaning of presented knowledge. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Furthermore, the function of decision rules can be used in classification and prediction.

Several notable methods exist for rule explanation and induction, such as Rough Set Theory (RST), the Inductive Dichotomizer 3 (ID3) and the Neural Network method (NN). The neural network method provides the best fit to numeric data, while ID3 and rough sets perform best with non-numerical data. The neural network method needs more training time compared with RST and ID3 [24,25,28]. The rule presentations for RST and ID3 were explainable and interpretable compared with the Neural Network method. Quinlan developed ID3 [26]. The main idea of ID3 in data classification was based on recursive portioning of the data set into categories. RST processed the relationship between attributes and objects which ID3 did not support. Furthermore, Grzymala-Busse [25] found that RST had better predictive capability compared with ID3 when applied to refine imperfect data. Therefore, the study was based on Rough Set Theory (RST).

Many theories, techniques and algorithms have been developed for analyzing imprecise data. One of the most successful of these is fuzzy set theory. Meanwhile, RST is a new mathematical tool introduced by Pawlak in the early 1980s capable of handling uncertainty and vagueness [27]. Comparison of RST with Fuzzy set theory revealed that RST did not need the membership function, but focused on equivalent relations or indiscernibility, and lower and upper approximation sets. Walczak and Massart [21] proposed a more detailed comparison between fuzzy set and rough set theories. Recently, RST had increasingly been applied in many fields to generate rules, provide reasoning and identify relationships in qualitative, incomplete, or imprecise data. The rules obtained based on rough set analysis can be applied to predict new cases. Such predictions are quite useful, especially in business analysis, because of the large volumes of incomplete and imprecise data involved in financial fields.

Three criteria exist for evaluating rule quality: the first criterion is rule accuracy, which means a rule fitting a specific class should not cover objects belonging to other classes. The second criterion is rule support, which means a good rule fitting a class should be supported by most of the objects belonging to the same class. The third criterion is rule compact according to which rule quality increases with decreasing number of attributes used.

Each decision rule can be characterised by its strength, namely the number of objects covered by the rule and the decision rule belonging to the specified decision class. A strong rule may have shorter and less specialised condition parts, and thus is typically a general rule. Strong rules are rough but not precise. However, as already stated, RST generates many rules, some of which have the same strength rate, number of support objects and condition parts. These factors make it difficult for decision makers to select suitable rules. Li and Chen [8] used the condition attribute activity of a decision rule under the criterion of compact for rule evaluation. This study also agrees that the best rules have the fewest attributes.

This study proposed a compactness rate based on the value domain of the condition attribute as an additional selection index for identifying the interesting rules (important rules) among the decision rules and also for supplying a pruning process based on the compactness rate. The compactness rate can be seen as the denseness of the value domain for each condition attribute. An interesting rule should have a high compactness rate, due to it containing a popular value domain. The compactness rate performs a pruning process and thus functions as a user-specified threshold to eliminate the data amount. Rules with compactness rates below the user-specified threshold are considered uninteresting (unimportant). Alternatively, objects with compactness rates below the user-specified threshold are considered outliers.

Relatively few studies have investigated the use of RST for personal investment analysis. This study used a well-designed questionnaire to survey some real Taiwanese investors about their personal investment styles. The questionnaire considered the influences on decision-making, including sex, age, and number of family members; monthly income [5,13]; and participant basic data, which may provide a basis for understanding participant needs. This study divided the proposed asset allocation model into three categories (types of personal asset allocation portfolio): conservative portfolio, moderate portfolio, and aggressive portfolio. Appendix C presents further details on the personal investment portfolio.

The proposed method successfully distinguishes the interesting rules from decision rules with the same strength, number of support objects and number of condition parts. The result of proposed method also identifies the outlier in the preprocessing data to reduce the data amount. Furthermore, the proposed method can also reduce the number of decision rules by assessing their threshold based on the compactness rate of decision rules.

The remainder of this paper is organised as follows. Section 2 describes the methodology of RST. Section 3 will present the proposed method—compactness rate usage in this study. In Section 4, a real case of personal investment is presented to show the process of the effects of the compactness rate on rules. Finally, in Section 5 presented the conclusions.

## 2. Concepts of RST

In this section, gives a brief summary of RST and its use in decision making. Section 2.1 gives an overview of the history of RST and Section 2.2 presents algorithms of the theory for decision-making are presented.

### 2.1. The history of RST

In 1982, Pawlak designed RST as a tool to describe the dependencies between attributes, evaluate the indiscernibility relation, and deal with inconsistent data [10–12]. Rough Set Theory also can handle data uncertainty and derive knowledge from ambiguous information. The theory has been applied to the management of a number of the issues, including medical diagnosis [8], engineering reliability [19], intelligent decision support systems [14], business failure prediction [1,2], the empirical study of insurance data [15], predicting stock prices [29], and data mining [7,16]. Another theory discusses the preference order of the attribute criteria needed to extend the original RST, such as sorting, choice and ranking problems [3], and using in spatial data methods and vague regions [18]. The Rough Set method is useful for exploring data patterns through a multi-dimensional data space and it determines the relative importance of each attribute with respect to its output.

RST assumes that the indiscernibility relation and data pattern comparison is based on the concept of an information system with indiscernible data, where the data is uncertain or inconsistent. An information system consists of objects in the universe. Those objects characterised by the same amount of information are similar to or indiscernible from one another. These objects can be grouped into classes called elementary sets. Feature/attribute selection is crucial in any data processing that consists of relevant (or maybe irrelevant) data patterns, but it may be redundant in data pattern recognition. Each elementary set is independent of the others [21]. From each elementary set can extract knowledge used in the real world.