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.

### 2.2. The algorithm of RST

Rough Set Theory is a mathematical approach to managing vague and uncertain data or problems related to information systems, indiscernibility relations and approximation sets, reduct and core attribute sets, decision rules. The remainder of this section discusses the above areas in detail.

#### 2.2.1. Information systems

Given a questionnaire model $QM$ (an information system), $QM = (U, A, V, \rho)$; $U = \{x_1, x_2, \ldots, x_n\}$, where $U$ denotes the universal object sets of $QM$; $A$ represents the model's attribute sets, assumed $A$ as consisting of attributes $c_1, c_2$ and $c_3$; $V = \bigcup_{c \in A} V_c$, is a set of values of the attributes.

Let $\rho : U \times A \to V$ be a description function; and let $\rho x$ be the description of $x$ in $QM$, where $\rho(x, c) \in V_c$ for each $c \in A$ and $x \in U$ [12].

A knowledge representation system containing the set of condition attributes (denoted as $CA$) and the set of decision attributes (denoted as $DA$) were used to construct a decision table.

#### 2.2.2. Indiscernibility relation and approximation set

Any subset $B$ of $A$ determines a binary relation $IND(B)$ on $U$, which be called an indiscernibility relation, and it is defined as $c \in B$, if $\rho_{x_1}(c) = \rho_{x_2}(c)$ for every $c \in A$. The equivalence class of $IND(B)$ is called an elementary set (of atoms) of $QM$. Assume a family $Y = \{X_1, X_2, \ldots, X_m\}$ is a family of non-empty sets (classification) where $X_i \subseteq U$, $X_i \neq \varnothing$, $X_i \cap X_j = \varnothing$ for $i \neq j$, $i, j = 1, 2, \ldots, m$ and $\cup_{i=1}^{m} X_i = U$. Objects grouped in the same class are called elementary sets, and the process is called classification. Thus, any $X_i$ of $U$ can be induced so that the value sets of attributes represented in $B$ are in the same class. The classification that processes $CA$ and $DA$, generates condition and decision classes.

Let any subset $X \subseteq U$, and $R$ be an equivalence relation and $x_i$ express objects $x_1, x_2, \ldots, x_n$, where $i = 1, 2, \ldots, n$, then $\underline{R}X = \{x \in U : [x]_R \subseteq X\}$, is the lower approximation of $X$; $\overline{R}X = \{x \in U : [x]_R \cap X \neq \varnothing\}$ is the upper approximation of $X$; $BndR(X) = \overline{R}X - \underline{R}X$, the boundary region of $X$ where the objects are inconsistent or ambiguous. If a family $Y = \{X_1, X_2, \ldots, X_m\}$ of non-empty sets (classification), then $\underline{R}Y = \{\underline{R}X_1, \underline{R}X_2, \ldots, \underline{R}X_m\}$ and $\overline{R}Y = \{\overline{R}X_1, \overline{R}X_2, \ldots, \overline{R}X_m\}$, are called the $R$-lower approximation and $R$-upper approximation of the family $Y$, respectively.

#### 2.2.3. Reduct and core attribute sets

In an information system, some attributes may be redundant and useless. The superfluous attributes can be removed without affecting the results. The goal of reducts is to improve the precision of decisions so that the reducts process for attributes reduces elementary set numbers. Given an attribute space $A = (CA, DA)$, where $CA \neq \varnothing$, $DA \neq \varnothing$; then $DA \cap CA = \varnothing$, and $DA \cup CA = A$, which are the elements of the decision table. Let $RED(B) \subseteq A$ and $RED(B)$ be the reduct set composed of a set of attributes $B$, that is, a minimal set of attributes that affects the process of decision-making. There may be several reduct attribute sets. The intersection of all reduct attribute sets is the core attribute set, which is the most important attribute in the decision-making process, $COR(C) = \cap RED(B)$, in which $COR(C)$ is the core composed of a set of attributes $C$. The decision rules can be induced and applied the reduct set to the model.

#### 2.2.4. Decision rules

Objects that have the same $IND(DA)$ are grouped together and called decision elementary sets (decision classes). The reduct condition attribute sets maintain the important relationships with the decision classes. Due to the functional dependencies between condition and decision attributes, a decision table may also be seen as a set of decision rules. The syntax can use the "if. . .then. . ." rule to specify as follows:

If $f(x, c_1)$ and $f(x, c_2)$ and. . . and $f(x, c_k)$, then $x$ belongs to $d_1$ or $d_2$ or $d_n$, where $\{c_1, c_2, \ldots, c_k\} \subseteq CA$ are condition attributes and $\{d_1, d_2, \ldots, d_n\} \subseteq DA$ are decision classes.

According to Pawlak (2002), a decision rule in $QM$ is expressed as $\Phi \to \Psi$, where $\Phi$ and $\Psi$ are referred to as the conditions and decisions of the rule, respectively (read as: if $\Phi$ then $\Psi$); $\sigma_{QM}(\Phi, \Psi) = \text{supp}_{QM}(\Phi, \Psi)/card(U)$ is the strength of the decision rule $\Phi \to \Psi$ in $QM$, where the number $\text{supp}_{QM}(\Phi, \Psi) = card(\left\| \Phi \wedge \Psi \right\|_{QM})$ will be called the support of the rule $\Phi \to \Psi$ in $QM$; and $card(U)$ is the cardinality of set $U$. This implies that a stronger rule will cover more objects and the strength rate of each decision rule can be calculated in order to decide on the appropriate rules.

A rule may cover several objects. A stronger rule definitely covers more objects. RST always generates fragment rules as the reduct process can improve the precision of a decision. Therefore, the condition attributes part of the fragment rules (compact rules) consists of the reduct attribute set or core attribute set. Those compact rules are usually general and less specialised. However, as mentioned previously, the theory generates many rules and some of them have the same strength rate and the same number of support objects, thus making it hard for decision makers to choose suitable rules. The method of compactness rate was proposed as the additional selection index to solve the problem and also as a pruning process to eliminate the data scope and decision rules. The next section is about using the compactness rate in decision rules.

## 3. Compactness rate method

In this section, the proposed new measuring method of compactness rate is described for use in decision rules.

**Definition 1.** A general form for object $x_i$ expressed as $x_i = (c^i_{1,1}, \ldots, c^i_{1,q1}; c^i_{2,1}, \ldots, c^i_{1,q2}; \ldots; c^i_{m,1}, \ldots, c^i_{m,qm}; d^i_{1,1}, \ldots, d^i_{1,p})$, where $qm$ is the number of sub-attributes of the attribute $c^i_m$; $p$ is the number of sub-attributes of the decision attribute $d$. Where $c^i_{j,p}$ is 1 if $c^i_j$ (the value of $c^i_p$ of $x_i$) equals $p$; otherwise $c^i_{j,p}$ is 0. For example, the $x_1$ is expressed in binary notation as $x1 = (1, 0; 0, 1, 0; 1, 0, 0; 1, 0)$. Here assumed there was only one decision attribute.

An information table can be seen as a decision table, assuming $QM = (U, CA \cup DA, V, \rho)$; $CA$ represents the condition attribute set, consisting of attributes $\{c_1, c_2, c_3\}$; and $DA$ represents the decision attribute set, and has a decision attribute $\{d\}$. $V_{c_1} = \{1, 2\}$ is a set of values of the attribute $c_1$, the same as $V_{c_2} = \{1, 2, 3\}$, $V_{c_3} = \{1, 2, 3\}$ and $V_d = \{1, 2\}$ as the value set of the attribute $c_2$, $c_3$ and $d$, respectively.

**Definition 2.** Frequency form for $c_{1,1}$ is $Freq(c_{1,1}) = \sum_{i=1}^{n} c^i_{1,1}$, where has $n$'s objects.

The frequency form computes the frequency for each sub-attribute of each attribute. The frequency expresses the popular degree of the value domain for each condition attribute.

**Definition 3.** Compute the weight for sub-attributes by the frequency.

The weight formula for $j$th sub-attribute of $m$th attribute expressed as $Weight(c_{m,j}) = Freq(c_{m,j})/n$, if there are $n$'s objects. The weight (frequency rate) can be seen as the important (or the interesting) rate of the value domain for each attribute.

**Definition 4.** A general form expressed the $i$th object compactness rate as $Comp(x_i) = Ob(Weight(c^i_g))$, where $g = 1, 2, \ldots, m$, has $m$'s attributes.

The function Ob( ) as the aggregation, maximum or other math function computes the object weight. The object weight is the

object's compactness rate, which aggregates the object's importance degree by the value domain of each attribute of the object. Therefore, the compactness rate of an object can be compared with that of other objects so that interesting objects can be identified. The compactness rates of objects can distinguish between the interesting objects and the less interesting objects.

**Definition 5.** A general form expressed the $l$th rule for the decision class $d$ as $R^{d,l}$.

Each decision rule can be characterised by the strength rate to establish whether the rule is stronger or not. The strength of a rule means the number of objects satisfying the condition part of the rule and belonging to the specific decision class.

**Definition 6.** A general form expressed the rule's compactness rate as $Comp(R^{d,l}) = \sum_{i \in U^{d,l}} Comp(x_i)$, where $U^{d,l}$ is the strength objects of the $l$th rule for the decision class $d$ and $U$ was the universe objects set of $QM$.

For example, $Comp(R^{1,1}) = Comp(x_i) + Comp(x_j)$, if $x_i$, $x_j$ were the support objects of $R^{1,1}$. The compactness rates of rules are summarised by the compactness rates of objects that are covered in the rules.

The compactness rate of rules can be established as a threshold to select decision rules. Rules with a compactness rate below the threshold are eliminated to improve the decision making.

Each rule should support several objects. Therefore, the compactness rate of rules can be summarised using the compactness rate of covered objects. The compactness rates of such objects are calculated from the frequency rate of the value domain of the condition attributes belonging to the object. The frequency rate is the weight calculated based on the frequency of the value domain of condition attributes divided by the total number of objects. The frequency rate can explore the degree of popularity or the important value domain for each attribute. The frequency, or the frequency rate of the value domain for each attribute, can identify the aggregative degree of the value domain for each attribute. Each object comprises a different value domain of attributes that produces a different compactness rate and implies a different characteristic for each object. If the compactness rate of a rule summarises that of an object, then higher compactness rates for objects are associated with higher compactness rates for rules. Rules with a high compactness rate are considered interesting (important). The following section details an experiment.

## 4. An empirical case of personal investment portfolio for the compactness rate

In this empirical case, a series of questions was asked to obtain information about personal/consumers' investments and portfolio types in 2006, e.g., monthly salary and portfolio type. The participants' personal data, namely, their gender, age, marital status, professional status, the number of years in the workforce, and educational level were used to classify purchase intentions, based on the definition by Executive Yuan of Taiwan.

### 4.1. Data analysis

The questionnaires were distributed to investors in the North and Northeast districts of Taiwan. Data were collected randomly based on a nominal scale. There were 200 valid questionnaires from a total of 221 received. The percentage of valid questionnaires was 90%. Among the valid respondents, there were 108 females and 92 males. The valid questionnaire data table presented below and participants were randomly collected in the empirical experiment.

| Valid data | |
|---|---|
| Male | Female |
| 92 (45%) | 108 (55%) |
| Total: 200 | |
| Valid rate: 90% (200/221) | |

The questionnaire considered the factors that affect decision-making, such as sex, age, and the number of family members; monthly income [5,13]; the nature of the investment products; and participants' basic data, which may serve as the basis for understanding their needs.

### 4.2. Empirical process

For the given information system $QM$, expert knowledge is used to process attributes for extraction. There are eight attributes: seven condition attributes, namely Age ($c_1$), Gender ($c_2$), Marital Status ($c_3$), Education ($c_4$), Number of Working Years ($c_5$), Professional Status ($c_6$), Monthly Salary ($c_7$); and one decision attribute, namely, portfolio type ($d$) representing the investment characteristics. After a reduct process was applied to the condition attributes which labelled the reduct attribute set as Age ($c_1$), Gender ($c_2$), Marital Status ($c_3$), Education ($c_4$), Number of Working Years ($c_5$), Professional Status ($c_6$), Monthly Salary ($c_7$). The core attribute set was the same as the reduct attribute set. The value set for the decision attribute was $\{1, 2, 3\}$, and there were three decision classes, that is, conservative portfolio, moderate portfolio, and aggressive portfolio. There are more details about the decision classification and condition attributes at Appendix C. The original attribute specification is detailed in Table A1 of Appendix A. The decision table is shown in Table 1.

### 4.3. The results of the empirical evaluation

Table 2 is the frequency and weight for each sub-attribute of the condition attributes.

The frequency of the value domain for the condition attributes can be computed by adding the binary value of each sub-attribute that belongs to the condition attributes. The weight for each sub-attribute was computed by the frequency of the value domain for the sub-attribute that occupied the total sample amount. The compactness rate of an object is summarised by the weight of each condition attribute. Table 3 shows the decision table with the compactness rate of each object for this study. Thus, the compactness rate for each decision rule can be summarised by the compactness rate of objects covered in the rules and then can be used to reveal the rule ranking by the compactness rates of the rules.

This study generates 75 rules by ROSE2 [14], which provides different techniques on rule induction based on the RST. The decision rules numbered from 56 to 75 are approximate rules, which mean that the rule did not belong to a specific decision class and may overlap with more than one decision class. However, too many decision rules impede decision making. For examples, we have randomly chosen several decision rules, some with the same strength rate, strength object number, accuracy rate of 100% and the same condition to compare the rules' compactness rates. Table 4 showed the partial rule examples of decision class 1 and decision class 3. The original rules generated from ROSE2 are detailed in Table B1 of Appendix B. Table 4 converts the original rules into a meaningful explanation and compares them using the compactness rate.

We proposed using the compactness method to find the interesting rule among a selection of rules. Based on the survey results which identified that people with conservative personal investment portfolios tend to be single college graduates under 29 years old, with less than four years work experience and a monthly salary under NT$30,000 (US$900). The profile of a typical aggressive investment portfolio holder is as follows: married, female, college

**Table 1**
Decision table.

| Object# | Attributes | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Condition | | | | | | | Decision |
| | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $c_5$ | $c_6$ | $c_7$ | $d$ |
| 1 | 1 | 2 | 1 | 4 | 1 | 1 | 1 | 3 |
| 2 | 2 | 2 | 2 | 3 | 3 | 1 | 2 | 1 |
| 3 | 2 | 1 | 1 | 3 | 2 | 1 | 2 | 3 |
| 4 | 1 | 1 | 2 | 3 | 2 | 1 | 2 | 3 |
| 5 | 1 | 2 | 1 | 4 | 1 | 2 | 1 | 1 |
| 6 | 2 | 2 | 1 | 5 | 1 | 2 | 2 | 2 |
| 7 | 1 | 1 | 1 | 4 | 1 | 1 | 2 | 1 |
| 8 | 4 | 1 | 2 | 4 | 4 | 1 | 3 | 3 |
| 9 | 4 | 2 | 2 | 5 | 4 | 1 | 2 | 3 |
| 10 | 1 | 1 | 1 | 3 | 1 | 1 | 1 | 2 |
| | | | | ... | | | | |

**Table 2**
Frequency and weight for each sub-attribute of condition attributes.

| Value domain | Frequency | | | | | | | Weight | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $c_5$ | $c_6$ | $c_7$ | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $c_5$ | $c_6$ | $c_7$ |
| 1 | 85 | 108 | 115 | 6 | 99 | 157 | 86 | 0.425 | 0.54 | 0.575 | 0.03 | 0.495 | 0.785 | 0.43 |
| 2 | 62 | 92 | 83 | 12 | 39 | 34 | 89 | 0.31 | 0.46 | 0.415 | 0.06 | 0.195 | 0.17 | 0.445 |
| 3 | 35 | | **2** | 100 | 30 | **9** | 15 | 0.175 | | 0.01 | 0.5 | 0.15 | 0.045 | 0.075 |
| 4 | 15 | | | 68 | 32 | 0 | **3** | 0.075 | | | 0.34 | 0.16 | 0 | 0.015 |
| 5 | **3** | | | 14 | 0 | 0 | **7** | 0.015 | | | 0.07 | 0 | 0 | 0.035 |
| 6 | | | | | 0 | 0 | | | | | | 0 | 0 | |
| 7 | | | | | 0 | | | | | | | 0 | | |
| Sum | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

**Table 3**
Decision table with compactness rate.

| Object# | Attributes | | | | | | | Object compactness rate |
|---|---|---|---|---|---|---|---|---|
| | Condition | | | | | | Decision | |
| | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $c_5$ | $c_6$ | $c_7$ | $d$ | |
| 1 | 1 | 2 | 1 | 4 | 1 | 1 | 1 | 3 | 3.510 |
| 2 | 2 | 2 | 2 | 3 | 3 | 1 | 2 | 1 | 2.605 |
| 3 | 2 | 1 | 1 | 3 | 2 | 1 | 2 | 3 | 3.350 |
| 4 | 1 | 1 | 2 | 3 | 2 | 1 | 2 | 3 | 3.305 |
| 5 | 1 | 2 | 1 | 4 | 1 | 2 | 1 | 1 | 2.895 |
| 6 | 2 | 2 | 1 | 5 | 1 | 2 | 2 | 2 | 2.525 |
| 7 | 1 | 1 | 1 | 4 | 1 | 1 | 2 | 1 | 3.605 |
| 8 | 4 | 1 | 2 | 4 | 4 | 1 | 3 | 3 | 2.390 |
| 9 | 4 | 2 | 2 | 5 | 4 | 1 | 2 | 3 | 2.410 |
| 10 | 1 | 1 | 1 | 3 | 1 | 1 | 1 | 2 | 3.750 |
| | | | | | ... | | | | |

graduate, less than 39 years old, with 15–19 years work experience and a monthly salary under NT$80,000 (US$2424). The moderate personal investment portfolio is typically chosen by single college graduates less than 29 years old, with less than nine years work experience and a monthly salary between NT$30,000 (US$900) and NT$80,000 (US$2,424).

### 4.4. Discussions

This section discusses two main contributions related to rule evaluation and threshold usage with compactness rate. Section 4.4.1 then discusses the evaluation of the rules regarding the compactness rate. Finally, Section 4.4.2 presents the threshold usage with the compactness rate.

#### 4.4.1. The rule evaluation with compactness rate

Objects whose attributes have the same values are classed together. RST uses mathematical methods to perform the classification. The elementary sets are used to extract relations with the same degree, which are then used to induce decision rules. The limitation of RST relates to the algorithm, which generates many rules;

**Table 4**
The rule explanation (partial rule examples).

| Rule# | Decision class | Meaning | Strength (support)objects# | Strength rate (%) | Accuracy rate (%) | Rule compactness rate |
|---|---|---|---|---|---|---|
| 2 | 1 | If $(c_3=1)\&(c_5=2)\wedge(c_7=1)\Rightarrow(d=1)$ | 57, 109 | 2.67 | 100 | 6.305 |
| 4 | 1 | If $(c_1=1)\wedge(c_4=2)\wedge(c_6=1)\Rightarrow(d=1)$ | 40, 154 | 2.67 | 100 | 6.540 |
| 7 | 1 | If $(c_4=1)\wedge(c_5=2)\wedge(c_7=2)\Rightarrow(d=1)$ | 45, 135 | 2.67 | 100 | 4.665 |
| 15 | 1 | If $(c_1=1)\wedge(c_4=4)\wedge(c_5=2)\Rightarrow(d=1)$ | 109, 189 | 2.67 | 100 | 6.515 |
| 38 | 3 | If $(c_5=4)\wedge(c_7=4)\Rightarrow(d=3)$ | 138, 165 | 2.99 | 100 | 4.570 |
| 39 | 3 | If $(c_4=2)\wedge(c_6=2)\Rightarrow(d=3)$ | 63, 146 | 2.99 | 100 | 3.725 |

some of which have the same strength rate, as well as the same number of support objects and condition attributes. This situation makes it difficult for decision-makers to select a reliable rule using these indices.

This study proposed an easy and scientific method, namely compactness rate, which can be employed as the denseness of the value domain for each condition attribute. The compactness rate is an additional selection index used to identify interesting decision rules. Interesting or popular objects should aggregate the higher compactness rate or select the highest or other math function from the value domain from each condition attribute. Therefore, an interesting rule should have higher compactness, because of being supported by several objects that satisfy the condition part of the rule. Rule or objects lower than the user-specified compactness rate are considered uninteresting or outliers.

There are three criteria for evaluating rule quality: the first criteria is the accuracy of a rule, the second criteria is the support of a rule, and the third is the compact of a rule. From the partial rule examples listed in Table 4, four rules with the same strength rate (2.67%) and strength object number (2), have accuracy rate is 100%, and the number of condition attributes is identical for each rule in decision class 1 ($d = 1$). The same criterion makes it more difficult for decision makers to identify the most interesting (important) rule. The compactness rates of the rules provide an alternative useful selection index. For instance, the rule ranking shows that rule number 4 is the most interesting rule, with a compactness rate of 6.54, and rule number 7 is the least interesting rule, with a compactness rate of 4.665, among the four rules in decision class 1. The decision maker can choose the rule with the highest compactness rate. The same process applies in decision class 3 ($d = 3$), where rule number 38 is the most interesting rule with a compactness rate of 4.57, and rule number 39 is the least interesting rule with a compactness rate of 3.725 among the two rules.

*4.4.2. The threshold evaluation with compactness rate*

This study applied the proposed method used the compactness rate to provide three thresholds during data preprocessing. Rules with compactness rates below the user-specified threshold are considered uninteresting (unimportant). Moreover, objects with compactness rates below the user-specified threshold are considered outliers. The thresholds have the advantage of eliminating the data amount and increasing data analysis. The first threshold relates to the frequency of the sub-attribute for each condition attribute. The decision-maker can remove objects that are not satisfied by the threshold. For instance, in Table 2, if the threshold is 10, objects with sub-attribute frequencies below the threshold are eliminated. Specifically, objects with the value 5 for attribute $c_1$, value 2 for attribute $c_3$ and values 3 and 7 for attribute $c_7$ are eliminated. Table 2 lists the relative value domains in bold type.

The second threshold relates to object compactness rate, which is indicated by the weight of the condition attribute of the object. The weight can be considered the important (or the interesting) rate of the value domain for each attribute. The compactness rate of the object indicates the weight of each attribute of that object. For example, in this study, the compactness rate of the highest object is 3.96 and that of the smallest object is 1.35. Table 3 lists the partial results of object compactness rates. Decision-makers can establish the threshold of object using the compactness rate to eliminate objects with small compactness rate.

The third threshold involves the decision rules. Each decision rule has its own compactness rate, which is calculated by totaling the compactness rates of the objects covered in the rule. In this study, the highest compactness rate for a rule is 16.005 and the smallest is 1.35. A decision-maker can set the threshold to eliminate decision rules with small compactness rates. This arrangement has the advantage of solving the problem too many decision rules.

The above thresholds should be set by experts based on real needs. The thresholds can be determined and special analysis performed to identify outlier data. Interesting rules can be identified by calculating the compactness rate of the rule. The compactness rate can help decision-makers select suitable rules from among those with few support objects and that share the same strength rate. An interesting rule should support popular objects with a higher compactness rate. These popular objects have general characteristics. Those objects express a popular knowledge domain that matches modern life.

Furthermore, the proposed compactness method is based on the value domain of the condition attributes. Calculating the weight for each condition attribute using the value domain identifies interesting objects. The interesting objects mean that most people have the same characteristics.

## 5. Conclusion

The new measurement calculates the compactness rate using the value domain of the attributes that provide an alternative method to find more interesting (important) rules supported by important (interesting) objects. The proposed method overcomes the problem of too many rules in classical RST, which previously created implementation difficulties. Furthermore, the proposed method also provides a pruning process with three thresholds to reduce the data amount and eliminate unimportant rules. The proposed method also requires calculating the data gathering degree. The proposed method thus is a scientific method suitable for application to different data types or combination with other soft computing methods. Decision-makers can use the proposed method to perform more precise analyses.

## Appendix A.

The original attribute specification for personal analysis described in Table A1.

**Table A1**
Attribute specification for the personal analysis.

| Attribute name | Attribute values | Attribute value sets |
|---|---|---|
| Condition attributes | | |
| Age ($c_1$) | <30; 30~39; 40~49; 50~59; 60~ | {1,2,3,4,5} |
| Gender ($c_2$) | Female; male | {1,2} |
| Marriage ($c_3$) | Single; marry; divorce | {1,2,3} |
| Education ($c_4$) | Under Junior High School; High School; College; University; Graduate School | {1,2,3,4,5} |
| Working years ($c_5$) | Under 4 years; 5–9 years; 10–14 years; 15–19 years; 20–24 years; 25–29 years; 30 years~ | {1,2,3,4,5,6,7} |
| Professional ($c_6$) | 1–6 (categorized according the insurance industry's listed profession) | {1,2,3,4,5,6} |
| Monthly salary ($c_7$) | Under NT$30,000; NT$30,001–NT$80,000; NT$80,001–NT$120,000; NT$120,001–NT$200,000; NT$200,001~ | {1,2,3,4,5} |
| Decision attribute | | |
| Portfolio type ($d$) | Conservative; moderate; aggressive | {1,2,3} |

*Note*: the insurance industry's listed profession – the report of Accounting and Statistics, Directorate-General of Budget, Executive Yuan.

## Appendix B.

The original rules generated from ROSE2 was described in Table B1 .

**Table B1**
Original rules generated from ROSE2.

Rule 1. $(c_2 = 2)$ & $(c_3 = 2)$ & $(c_5 = 3)$ & $(c_6 = 1) \Rightarrow (d = 1)$; [5, 5, 6.67%, 100.00%][5,0,0][{2, 41, 110, 116, 145}, {}, {}]
Rule 2. $(c_3 = 1)$ & $(c_5 = 2)$ & $(c_7 = 1) \Rightarrow (d = 1)$; [2, 2, 2.67%, 100.00%][2,0,0][{57, 109}, {}, {}]
Rule 3. $(c_5 = 3)$ & $(c_7 = 5) \Rightarrow (d = 1)$; [3, 3, 4.00%, 100.00%][3,0,0][{41, 116, 193}, {}, {}]
Rule 4. $(c_1 = 1)$ & $(c_4 = 2)$ & $(c_6 = 1) \Rightarrow (d = 1)$; [2, 2, 2.67%, 100.00%][2,0,0][{40, 154}, {}, {}]
Rule 5. $(c_1 = 3)$ & $(c_2 = 1)$ & $(c_4 = 3)$ & $(c_6 = 1)$ & $(c_7 = 2) \Rightarrow (d = 1)$; [2, 2, 2.67%, 100.00%][2,0,0][{114, 181}, {}, {}]
Rule 6. $(c_1 = 2)$ & $(c_3 = 1)$ & $(c_4 = 3)$ & $(c_7 = 1) \Rightarrow (d = 1)$; [2, 2, 2.67%, 100.00%][2,0,0][{27, 189}, {}, {}]
Rule 7. $(c_4 = 1)$ & $(c_5 = 2)$ & $(c_7 = 2) \Rightarrow (d = 1)$; [2, 2, 2.67%, 100.00%][2,0,0][{45, 135}, {}, {}]
Rule 8. $(c_1 = 2)$ & $(c_3 = 1)$ & $(c_4 = 4)$ & $(c_5 = 3) \Rightarrow (d = 1)$; [1, 1, 1.33%, 100.00%][1,0,0][{33}, {}, {}]
Rule 9. $(c_1 = 2)$ & $(c_5 = 1)$ & $(c_6 = 1)$ & $(c_7 = 1) \Rightarrow (d = 1)$; [3, 3, 4.00%, 100.00%][3,0,0][{27, 118, 200}, {}, {}]
Rule 10. $(c_5 = 2)$ & $(c_6 = 3) \Rightarrow (d = 1)$; [2, 2, 2.67%, 100.00%][2,0,0][{180, 195}, {}, {}]
Rule 11. $(c_1 = 5)$ & $(c_2 = 2) \Rightarrow (d = 1)$; [1, 1, 1.33%, 100.00%][1,0,0][{144}, {}, {}]
Rule 12. $(c_2 = 1)$ & $(c_3 = 1)$ & $(c_6 = 2)$ & $(c_7 = 2) \Rightarrow (d = 1)$; [1, 1, 1.33%, 100.00%][1,0,0][{18}, {}, {}]
Rule 13. $(c_1 = 2)$ & $(c_3 = 2)$ & $(c_6 = 2)$ & $(c_7 = 1) \Rightarrow (d = 1)$; [1, 1, 1.33%, 100.00%][1,0,0][{95}, {}, {}]
Rule 14. $(c_3 = 1)$ & $(c_4 = 1) \Rightarrow (d = 1)$; [1, 1, 1.33%, 100.00%][1,0,0][{22}, {}, {}]
Rule 15. $(c_1 = 1)$ & $(c_4 = 4)$ & $(c_5 = 2) \Rightarrow (d = 1)$; [2, 2, 2.67%, 100.00%][2,0,0][{109, 198}, {}, {}]
Rule 16. $(c_1 = 4)$ & $(c_2 = 1)$ & $(c_7 = 1) \Rightarrow (d = 2)$; [1, 1, 1.72%, 100.00%][0,1,0][{}, {150}, {}]
Rule 17. $(c_1 = 4)$ & $(c_4 = 2)$ & $(c_7 = 2) \Rightarrow (d = 2)$; [1, 1, 1.72%, 100.00%][0,1,0][{}, {42}, {}]
Rule 18. $(c_2 = 2)$ & $(c_4 = 5)$ & $(c_7 = 5) \Rightarrow (d = 2)$; [1, 1, 1.72%, 100.00%][0,1,0][{}, {50}, {}]
Rule 19. $(c_5 = 2)$ & $(c_7 = 4) \Rightarrow (d = 2)$; [1, 1, 1.72%, 100.00%][0,1,0][{}, {60}, {}]
Rule 20. $(c_1 = 3)$ & $(c_4 = 4)$ & $(c_6 = 2) \Rightarrow (d = 2)$; [2, 2, 3.45%, 100.00%][0,2,0][{}, {53, 62}, {}]
Rule 21. $(c_2 = 2)$ & $(c_3 = 2)$ & $(c_4 = 4)$ & $(c_7 = 1) \Rightarrow (d = 2)$; [2, 2, 3.45%, 100.00%][0,2,0][{}, {21, 137}, {}]
Rule 22. $(c_2 = 2)$ & $(c_3 = 1)$ & $(c_6 = 2)$ & $(c_7 = 2) \Rightarrow (d = 2)$; [3, 3, 5.17%, 100.00%][0,3,0][{}, {6, 96, 112}, {}]
Rule 23. $(c_1 = 2)$ & $(c_2 = 1)$ & $(c_3 = 1)$ & $(c_5 = 1)$ & $(c_6 = 1) \Rightarrow (d = 2)$; [2, 2, 3.45%, 100.00%][0,2,0][{}, {43, 172}, {}]
Rule 24. $(c_4 = 1)$ & $(c_5 = 3) \Rightarrow (d = 2)$; [1, 1, 1.72%, 100.00%][0,1,0][{}, {142}, {}]
Rule 25. $(c_1 = 3)$ & $(c_4 = 3)$ & $(c_7 = 1) \Rightarrow (d = 2)$; [3, 3, 5.17%, 100.00%][0,3,0][{}, {136, 152, 190}, {}]
Rule 26. $(c_1 = 3)$ & $(c_3 = 1)$ & $(c_5 = 4) \Rightarrow (d = 2)$; [1, 1, 1.72%, 100.00%][0,1,0][{}, {108}, {}]
Rule 27. $(c_2 = 1)$ & $(c_5 = 2)$ & $(c_6 = 2) \Rightarrow (d = 2)$; [1, 1, 1.72%, 100.00%][0,1,0][{}, {48}, {}]
Rule 28. $(c_3 = 2)$ & $(c_4 = 3)$ & $(c_5 = 4)$ & $(c_7 = 2) \Rightarrow (d = 2)$; [1, 1, 1.72%, 100.00%][0,1,0][{}, {158}, {}]
Rule 29. $(c_4 = 3)$ & $(c_5 = 2)$ & $(c_6 = 1)$ & $(c_7 = 3) \Rightarrow (d = 2)$; [1, 1, 1.72%, 100.00%][0,1,0][{}, {197}, {}]
Rule 30. $(c_1 = 2)$ & $(c_4 = 4)$ & $(c_5 = 4) \Rightarrow (d = 2)$; [1, 1, 1.72%, 100.00%][0,1,0][{}, {111}, {}]
Rule 31. $(c_4 = 2)$ & $(c_6 = 3) \Rightarrow (d = 2)$; [1, 1, 1.72%, 100.00%][0,1,0][{}, {123}, {}]
Rule 32. $(c_1 = 3)$ & $(c_3 = 1)$ & $(c_4 = 3) \Rightarrow (d = 2)$; [1, 1, 1.72%, 100.00%][0,1,0][{}, {74}, {}]
Rule 33. $(c_2 = 1)$ & $(c_3 = 2)$ & $(c_4 = 5) \Rightarrow (d = 3)$; [8, 8, 11.94%, 100.00%][0,0,8][{}, {}, {149, 164, 167, 177, 183, 185, 186, 187}]
Rule 34. $(c_1 = 4)$ & $(c_5 = 3) \Rightarrow (d = 3)$; [1, 1, 1.49%, 100.00%][0,0,1][{}, {}, {168}]
Rule 35. $(c_3 = 2)$ & $(c_4 = 3)$ & $(c_7 = 3) \Rightarrow (d = 3)$; [6, 6, 8.96%, 100.00%][0,0,6][{}, {}, {34, 54, 148, 153, 170, 194}]
Rule 36. $(c_1 = 2)$ & $(c_2 = 1)$ & $(c_4 = 3)$ & $(c_7 = 2) \Rightarrow (d = 3)$; [7, 7, 10.45%, 100.00%][0,0,7][{}, {}, {3, 44, 107, 113, 134, 188, 192}]
Rule 37. $(c_1 = 3)$ & $(c_2 = 2)$ & $(c_3 = 2)$ & $(c_5 = 2)$ & $(c_6 = 1) \Rightarrow (d = 3)$; [1, 1, 1.49%, 100.00%][0,0,1][{}, {}, {171}]
Rule 38. $(c_5 = 4)$ & $(c_7 = 4) \Rightarrow (d = 3)$; [2, 2, 2.99%, 100.00%][0,0,2][{}, {}, {138, 165}]
Rule 39. $(c_4 = 2)$ & $(c_6 = 2) \Rightarrow (d = 3)$; [2, 2, 2.99%, 100.00%][0,0,2][{}, {}, {63, 146}]
Rule 40. $(c_3 = 2)$ & $(c_4 = 1)$ & $(c_5 = 1) \Rightarrow (d = 3)$; [1, 1, 1.49%, 100.00%][0,0,1][{}, {}, {156}]
Rule 41. $(c_3 = 1)$ & $(c_4 = 3)$ & $(c_5 = 3)$ & $(c_7 = 2) \Rightarrow (d = 3)$; [3, 3, 4.48%, 100.00%][0,0,3][{}, {}, {107, 134, 160}]
Rule 42. $(c_1 = 1)$ & $(c_4 = 5) \Rightarrow (d = 3)$; [1, 1, 1.49%, 100.00%][0,0,1][{}, {}, {155}]
Rule 43. $(c_1 = 3)$ & $(c_5 = 4)$ & $(c_7 = 5) \Rightarrow (d = 3)$; [1, 1, 1.49%, 100.00%][0,0,1][{}, {}, {75}]
Rule 44. $(c_4 = 3)$ & $(c_5 = 3)$ & $(c_6 = 2) \Rightarrow (d = 3)$; [4, 4, 5.97%, 100.00%][0,0,4][{}, {}, {54, 113, 194, 196}]
Rule 45. $(c_3 = 2)$ & $(c_4 = 4)$ & $(c_5 = 2)$ & $(c_7 = 2) \Rightarrow (d = 3)$; [2, 2, 2.99%, 100.00%][0,0,2][{}, {}, {59, 65}]
Rule 46. $(c_3 = 1)$ & $(c_4 = 3)$ & $(c_5 = 4) \Rightarrow (d = 3)$; [2, 2, 2.99%, 100.00%][0,0,2][{}, {}, {71, 151}]
Rule 47. $(c_2 = 2)$ & $(c_4 = 1)$ & $(c_7 = 1) \Rightarrow (d = 3)$; [1, 1, 1.49%, 100.00%][0,0,1][{}, {}, {19}]
Rule 48. $(c_1 = 4)$ & $(c_6 = 2) \Rightarrow (d = 3)$; [3, 3, 4.48%, 100.00%][0,0,3][{}, {}, {143, 146, 147}]
Rule 49. $(c_4 = 2)$ & $(c_5 = 4)$ & $(c_7 = 5) \Rightarrow (d = 3)$; [1, 1, 1.49%, 100.00%][0,0,1][{}, {}, {169}]
Rule 50. $(c_1 = 3)$ & $(c_3 = 3) \Rightarrow (d = 3)$; [1, 1, 1.49%, 100.00%][0,0,1][{}, {}, {140}]
Rule 51. $(c_1 = 2)$ & $(c_2 = 1)$ & $(c_3 = 2)$ & $(c_5 = 3)$ & $(c_7 = 2) \Rightarrow (d = 3)$; [4, 4, 5.97%, 100.00%][0,0,4][{}, {}, {44, 47, 113, 192}]
Rule 52. $(c_1 = 1)$ & $(c_3 = 2)$ & $(c_4 = 3) \Rightarrow (d = 3)$; [2, 2, 2.99%, 100.00%][0,0,2][{}, {}, {4, 39}]
Rule 53. $(c_1 = 3)$ & $(c_2 = 1)$ & $(c_3 = 1) \Rightarrow (d = 3)$; [1, 1, 1.49%, 100.00%][0,0,1][{}, {}, {76}]
Rule 54. $(c_3 = 2)$ & $(c_4 = 5)$ & $(c_7 = 2) \Rightarrow (d = 3)$; [5, 5, 7.46%, 100.00%][0,0,5][{}, {}, {9, 149, 177, 185, 187}]
Rule 55. $(c_1 = 2)$ & $(c_4 = 4)$ & $(c_6 = 2)$ & $(c_7 = 1) \Rightarrow (d = 3)$; [1, 1, 1.49%, 100.00%][0,0,1][{}, {}, {99}]
# Approximate rules
Rule 56. $(c_2 = 2)$ & $(c_3 = 1)$ & $(c_4 = 4)$ & $(c_5 = 1)$ & $(c_6 = 1)$ & $(c_7 = 1) \Rightarrow (d = 1)$ OR $(d = 3)$; [5, 5, 26.32%, 100.00%][4,0,1][{14, 120, 161, 162}, {}, {1}]
Rule 57. $(c_3 = 2)$ & $(c_4 = 4)$ & $(c_5 = 1)$ & $(c_7 = 2) \Rightarrow (d = 1)$ OR $(d = 3)$; [3, 3, 15.79%, 100.00%][1,0,2][{175}, {}, {66, 179}]
Rule 58. $(c_1 = 2)$ & $(c_2 = 2)$ & $(c_3 = 1)$ & $(c_4 = 3)$ & $(c_5 = 2) \Rightarrow (d = 1)$ OR $(d = 3)$; [5, 5, 26.32%, 100.00%][4,0,1][{56, 70, 72, 174}, {}, {49}]
Rule 59. $(c_1 = 3)$ & $(c_2 = 2)$ & $(c_3 = 1)$ & $(c_5 = 3) \Rightarrow (d = 1)$ OR $(d = 3)$; [2, 2, 10.53%, 100.00%][1,0,1][{73}, {}, {163}]
Rule 60. $(c_2 = 1)$ & $(c_4 = 4)$ & $(c_5 = 4)$ & $(c_6 = 1)$ & $(c_7 = 2) \Rightarrow (d = 1)$ OR $(d = 3)$; [2, 2, 10.53%, 100.00%][1,0,1][{61}, {}, {67}]
Rule 61. $(c_1 = 1)$ & $(c_2 = 1)$ & $(c_4 = 4)$ & $(c_5 = 1)$ & $(c_7 = 2) \Rightarrow (d = 1)$ OR $(d = 3)$; [2, 2, 10.53%, 100.00%][1,0,1][{7}, {}, {199}]
Rule 62. $(c_1 = 1)$ & $(c_2 = 1)$ & $(c_4 = 4)$ & $(c_7 = 1) \Rightarrow (d = 1)$ OR $(d = 2)$; [10, 10, 23.26%, 100.00%][9,1,0][{78, 119, 121, 124, 126, 127, 157, 173, 184}, {69}, {}]
Rule 63. $(c_1 = 1)$ & $(c_2 = 2)$ & $(c_3 = 1)$ & $(c_4 = 3) \Rightarrow (d = 1)$ OR $(d = 2)$; [16, 16, 37.21%, 100.00%][10,6,0][{11, 31, 37, 46, 88, 92, 93, 115, 129, 159}, {17, 36, 83, 86, 117, 176}, {}]
Rule 64. $(c_4 = 3)$ & $(c_5 = 1)$ & $(c_7 = 2) \Rightarrow (d = 1)$ OR $(d = 2)$; [6, 6, 13.95%, 100.00%][3,3,0][{16, 29, 159}, {35, 38, 176}, {}]
Rule 65. $(c_1 = 2)$ & $(c_2 = 2)$ & $(c_4 = 4)$ & $(c_5 = 1)$ & $(c_6 = 1) \Rightarrow (d = 1)$ OR $(d = 2)$; [4, 4, 9.30%, 100.00%][1,3,0][{102}, {52, 103, 182}, {}]
Rule 66. $(c_4 = 3)$ & $(c_5 = 1)$ & $(c_6 = 3) \Rightarrow (d = 1)$ OR $(d = 2)$; [3, 3, 6.98%, 100.00%][1,2,0][{87}, {23, 85}, {}]
Rule 67. $(c_1 = 3)$ & $(c_2 = 2)$ & $(c_3 = 2)$ & $(c_5 = 4)$ & $(c_7 = 2) \Rightarrow (d = 1)$ OR $(d = 2)$; [2, 2, 4.65%, 100.00%][1,1,0][{104}, {106}, {}]
Rule 68. $(c_1 = 1)$ & $(c_2 = 2)$ & $(c_6 = 2)$ & $(c_7 = 1) \Rightarrow (d = 1)$ OR $(d = 2)$; [8, 8, 18.60%, 100.00%][5,3,0][{5, 88, 92, 93, 129}, {17, 28, 86}, {}]
Rule 69. $(c_1 = 2)$ & $(c_3 = 2)$ & $(c_4 = 3)$ & $(c_6 = 1)$ & $(c_7 = 1) \Rightarrow (d = 1)$ OR $(d = 2)$; [2, 2, 4.65%, 100.00%][1,1,0][{101}, {178}, {}]
Rule 70. $(c_4 = 4)$ & $(c_6 = 1)$ & $(c_7 = 3) \Rightarrow (d = 2)$ OR $(d = 3)$; [4, 4, 33.33%, 100.00%][0,2,2][{}, {105, 166}, {8, 100}]
Rule 71. $(c_1 = 2)$ & $(c_2 = 2)$ & $(c_3 = 2)$ & $(c_4 = 3)$ & $(c_5 = 2)$ & $(c_7 = 2) \Rightarrow (d = 2)$ OR $(d = 3)$; [4, 4, 33.33%, 100.00%][0,3,1][{}, {55, 77, 141}, {51}]
Rule 72. $(c_1 = 1)$ & $(c_2 = 2)$ & $(c_4 = 4)$ & $(c_6 = 1)$ & $(c_7 = 2) \Rightarrow (d = 2)$ OR $(d = 3)$; [4, 4, 33.33%, 100.00%][0,2,2][{}, {97, 133}, {64, 128}]
Rule 73. $(c_2 = 1)$ & $(c_4 = 3)$ & $(c_5 = 1)$ & $(c_6 = 1)$ & $(c_7 = 1) \Rightarrow (d = 1)$ OR $(d = 2)$ OR $(d = 3)$; [20, 20, 68.97%, 100.00%][6,6,8][{26, 80, 81, 84, 91, 125}, {10, 20, 30, 79, 82, 122, 130, 139}, {15, 25, 32, 89, 98, 191}]
Rule 74. $(c_2 = 1)$ & $(c_4 = 3)$ & $(c_5 = 1)$ & $(c_6 = 2) \Rightarrow (d = 1)$ OR $(d = 2)$ OR $(d = 3)$; [6, 6, 20.69%, 100.00%][1,1,4][{13, 24, 90, 94}, {131}, {12}]
Rule 75. $(c_2 = 2)$ & $(c_3 = 1)$ & $(c_4 = 4)$ & $(c_5 = 2)$ & $(c_7 = 2) \Rightarrow (d = 1)$ OR $(d = 2)$ OR $(d = 3)$; [3, 3, 10.34%, 100.00%][1,1,1][{132}, {68}, {58}]

## Appendix C.

The most important factor for investment revenue is asset allocation in an investment portfolio. Investors should modify their investment portfolios frequently and managed effectively in order to increase personal wealth. An investment portfolio is composed of many investment products which can be divided into two categories: risk products and non-risk products. Risk investments include stocks, mutual funds, foreign exchange, land, houses, and investment insurance. Non-risk investments include bank deposits, traditional insurance (such as life insurance, AD&D), and government bonds.

The asset allocation model usually can be divided into five categories or types of personal asset allocation portfolio: conservative portfolio, moderately conservative portfolio, moderate portfolio, moderately aggressive portfolio, and aggressive portfolio. In this study simplify the analysis combined the portfolios into three types: non-aggressive (conservative) portfolio, moderate portfolio, and aggressive portfolio.

The empirical study is following the paper "Using FSBT technique with Rough Set Theory for personal investment portfolio analysis" [20]. In this study, the decision attributes already have been simplified to three decision classes (categories: conservative portfolio, moderate portfolio, aggressive portfolio) in order to concise the paper structure.

In the paper [20], there are eight condition attributes. The attributes of children was a superfluous attribute after a reduct process. The reason for the attribute removed was due to the social benefit and education grant for the children which may reduce the family expenditure. For simplifying the processes, the children attribute did not appear in this study.

Here, 31 validation sample data sets as test data collected in 2009 are added to the hit test to check the feasibility of the decision rules in this study. The results in Table C1 showed that the hit rate reaches 55%. The results are clear that more than half new objects can fit into classes among the decision classes. More condition classes will decrease the hit rate and increase the decision rules. Another results of the reliability test based on Cronbach's Alpha for training data (data number are 200 which collected in 2006) and total data (data number are 231 which combined the training data and test data), these two constructs were 0.726 and 0.705, respectively.

The hit test between the training data and test data was described in Table C1.

**Table C1**
The hit rates of new data.

| # of original objects | Decision rules | | # of new test objects | Hit decision | | | | | | | | | | | | | | | | | | | | | | | | | | | Hit object | Hit rate (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Class # | # of new test objects | | Decision # | Class | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | | | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | | |
| 77 | Class 1 | 13 | | Class 1 | 3 | 2 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 10 | 77% |
| 53 | Class 2 | 9 | | Class 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 56% |
| 71 | Class 3 | 8 | | Class 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 25% |
| 201 | Total | 30 | | | 4 | 3 | 0 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 3 | 0 | 0 | 0 | 0 | 17 | 57% |

## References


[1] M. Beynon, M. Peel, Variable precision rough set theory and data discretisation: an application to corporate failure prediction, Omega 29 (6) (2001) 561–576.
[2] A.I. Dimitras, R. Slowinski, R. Susmaga, C. Zopounidis, Business failure prediction using rough sets, European Journal of Operational Research 114 (2) (1999) 263–280.
[3] S. Greco, B. Matarazzo, R. Slowinski, Rough sets theory for multicriteria decision analysis, European Journal of Operational Research 129 (1) (2001) 1–47.
[4] K. Hanisch, Reasons people retire and their relations to attitudinal and behavioral correlates in retirement, Journal of Vocational Behavior 45 (1) (1994) 1–16.
[5] T. Harrison, Understanding the behaviour of financial services consumers: a research agenda, Journal of Financial Service Marketing 8 (1) (2003) 6–9.
[6] T. Hennig-Thuran, K. Klee, The impact of customer satisfaction and relationship quality on customer retention: a critical reassessment and model development, Psychology and Marketing 14 (8) (1997), 797-764.
[7] Y. Hu, R. Chen, G. Tzeng, Finding fuzzy classification rules using data mining techniques, Pattern Recognition Letters 24 (1–3) (2003) 509–519.
[8] H. Li, M. Chen, Induction of multiple criteria optimal classification rules for biological and medical data, Computers in Biology and Medicine 38 (2008) 42–52.
[9] H. Markowitz, Portfolio selection, Journal of Finance Paper 7 (1) (1952) 77–91.
[10] Z. Pawlak, Rough sets, International Journal of Computer and Information Science 11 (5) (1982) 341–356.
[11] Z. Pawlak, Rough classification, International Journal of Man-Machine Studies 20 (5) (1984) 469–483.
[12] Z. Pawlak, Rough sets, decision algorithms and Bayes' theorem, European Journal of Operational Research 136 (1) (2002) 181–189.
[13] D Plath, H. Stevenson, Financial services consumption behavior across Hispanic American consumers, Journal of Business Research 58 (8) (2005) 1089–1099.
[14] B. Predki, R. Slowinski, J. Stefanowski, R. Susmaga, S. Wilk, ROSE—Software Implementation of the Rough Set Theory, in: L. Polkowski, A. Skowron (Eds.), Rough Sets and Current Trends in Computing, Proc. RSCTC'98, Lecture Notes in Artificial Intelligence, vol. 1424, Springer, Berlin, 1998, pp. 605–608.
[15] J. Shyng, F. Wang, G. Tzeng, K. Wu, Rough set theory in analyzings the attributes of combination values for the insurance market, Expert System with Application 32 (1) (2007) 56–64.
[16] J. Shyng, G. Tzeng, S. Hsieh, F. Wang, Data mining for multi-domain decision-making based on Rough Set Theory, in: IEEE International Conference on Systems, Man, and Cybernetics, Taipei, Taiwan, 2006.
[17] M. Teichman, P. Cecconi, B. Bernheim, N. Novarro, M. Monga, D. DaRosa, M. Resnick, How do residents manage personal finances? The American Journal of Surgery 189 (2) (2005) 134–139.
[18] T. Beaubouefa, F.E. Petryb, R. Ladnerb, Spatial data methods and vague regions: a rough set approach, Applied Soft Computing 7 (1) (2007) 425–440.
[19] X. Shaoa, X. Chua, H. Qiua, L. Gaoa, J. Yanb, An expert system using rough sets theory for aided conceptual design of ship's engine room automation, Expert Systems with Applications 36 (2) (2009) 3223–3233.
[20] J. Shyng, H. Shieh, G. Tzeng, S. Hsieh, Using FSBT technique with Rough Set Theory for personal investment portfolio analysis, European Journal of Operational Research 201 (2010) 601–607.
[21] B. Walczak, D. Massart, Tutorial rough sets theory, Chemometrics and Intelligent Laboratory Systems 47 (1) (1999) 1–16.
[22] The Wall Street Journal, http://asia.wsj.com/home-page.
[23] Dalal Street Investment Journal, http://www.dalalstreetjournal.com.
[24] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Simon Fraser University, Morgan Kaufmann Publishers, San Francisco, 2001.
[25] D. Grzymala-Busse, J. Grzymala-Busse, On the usefulness of machine learning approach to knowledge acquisition, Computational Intelligence 11 (2) (1995) 268–279.
[26] R. Quinlan, Induction of decision trees, Machine Learning 1 (1) (1986) 81–106.
[27] L. Zadeh, Fuzzy sets, Information and Control 8 (3) (1965) 338–353.
[28] R. Li, Z. Wang, Mining classification rules using rough sets and neural networks, European Journal of Operational Research 157 (2) (2004) 439–448.
[29] C. Cheng, T. Chen, L. Wei, A hybrid model based on rough sets theory and genetic algorithms for stock price forecasting, Information Sciences 180 (9) (2010) 1610–1629.