African Journal of Business Management Vol. 5(15), pp. 6554-6568, 4 August, 2011
Available online at http://www.academicjournals.org/AJBM
DOI: 10.5897/AJBM11.069
ISSN 1993-8233 ©2011 Academic Journals

Full Length Research Paper

Optimized classification approach to modeling an expert system for selecting stock portfolio

Kuang Yu Huang

Department of Information Management, Ling Tung University, Nan-tun District, Taichung City, Taiwan.
E-mail: kyhuang@mail.ltu.edu.tw. Tel: +886-9-22621030. Fax: +886-4-23821912.

Accepted 17 May, 2011

An expert system for automatically selecting stock portfolio is presented. The expert system involved Grey Relational Analysis (GRA) model, the VPRS-index clustering / classification method, and Variable Precision Rough Set (VPRS) theory. The GRA model is applied to consolidate the 53 financial indices into six financial ratios (Grey Relational Grades (GRGs)) for each stock item. The VPRS-index method is used to determine the optimal number of clusters per GRG. VPRS theory is then applied to identify the stocks within the β-lower approximate sets. Finally, the GRGs of each candidate stock item are consolidated to a single GRG indicating the ability of the stock item to maximize the rate of return. The validity and effectiveness of the VPRS-index clustering / classification method is first evaluated prior to that of the expert system. After that, results of this study showed that this expert system yields a higher rate of return than those of several existing portfolio selection systems.

Key words: Fuzzy C-Means, VPRS, VPRS-index method, classification, stock portfolio.

INTRODUCTION

In recent decades, many applications have been proposed for predicting market trends selecting suitable stocks for investment purposes. These applications are typically based upon genetic algorithms (GAs) (Hassan et al., 2007), neural networks (Jandaghi et al., 2010), statistical forecasting mechanisms (Box and Jenkins, 1976; Tse, 1997), or rough set (RS) theory (Skalko, 1996). When using regression models to predict stock market trends, the results are determined not only by the financial indices of the stocks involved, but also by external factors such as the financial environment, political changes, changes in company strategy, variations in the demand/supply relationship, and so on. Consequently, the reliability of the prediction results cannot be secure. Additionally, in real-world stock market systems, the information associated with each data object is vague and uncertain. Therefore, the task of identifying the relationships between the independent and dependent variables is extremely challenging. Therefore, in endeavoring to maximize the rate of return on investment while simultaneously reducing the associated risk, investors are turning progressively toward the use of sophisticated computer modeling and forecasting techniques based on a variety of clustering, classification, and data analysis methods.

Accordingly, in a recent study (Huang et al., 2009; Huang, 2009b), the current author proposed investment portfolios selection systems in which the attribute reduction is obtained using GM(1,N) function and the number of clusters per attribute, obtained using GM(1,N) function, is just defined in advance as N = 3. However, these portfolios selection systems cannot clearly explain what the threshold parameter selection mechanism of attribute reduction is based on and why the number of clusters per attribute is defined in advance as N = 3. The purpose here is to explore a little further into these two issues in order to model a more robust expert system for selecting stock portfolio. Accordingly, an automatic stock selection system proposed in this study comprises two major components, namely (a) Data Processing and (b) Data Mining.

In the Data Processing component, first of all, we have to inquire into the issue of attribute reductions. In applying classification theory to classify such datasets, it is desirable to pre-process the dataset in order to eliminate the conditional attributes which have little or no effect on the classification decision. By doing so, the decision table is facilitated and the decision rules can be

more readily identified. Of all the available dimension reduction methods (e.g. principle component analysis, independent component analysis and GRA), GRA is particularly attractive since it consolidate attributes while other methods reduce attributes which have little or no effect on the classification decision. Thus, in the Data Processing component, 53 financial indices are gathered automatically for each stock item every quarter and a GRA (Grey Relational Analysis) model is used to consolidate these indices into just 6 predetermined financial ratios (GRGs). The GRA (Deng, 1985) is the most fundamental components of Grey System theory, which is proposed by Deng (1982) and is a powerful technique for handling systems characterized by poor, deficient and vague information, is used to quantify the respective effects of the various factors within the grey system in terms of GRGs. Basically, GRA function is an arithmetic mean (Wen, 2004), geometric mean (Huang et al., 2008) or p-norm function (Nagai et al., 2005) applied to a specified grouping conditional attributes. GRA provides the means to "weight" the various factors within a vague system according to their effects on the system outcome, and hence provides an ideal basis for classification systems.

Secondly, in the Data Mining component, an enhanced classification method is proposed. The literature contains many algorithms for automatic classification purposes, including decision-tree algorithms such as neural networks (Lin, 2010), support vector machines (Vapnik, 2000), Bayesian classifiers (Wang and Hsu, 2010), and so forth. These algorithms all have their own particular merits and have found widespread use in a diverse range of applications, including weather prediction, manufacturing process planning, medical diagnosis, and so on. However, they cannot deal effectively with continuous valued systems or systems characterized by uncertainty or missing information Thus, the Rough Set (RS) theory is employed in the enhanced classification method and is used to classify such systems.

(RS) theory was first introduced more than twenty years ago (Pawlak, 1982) and has applied to extract reliable classification rules (Huang and Jane, 2009; Pawlak, 1994; Huang, 2009b) in a diverse range of fields. However, the ability of RS techniques to correctly classify a dataset relies upon the availability of complete and certain information. To extend RS theory to perform a classification operation with a controlled degree of uncertainty or misclassification error, Variable Precision Rough Set (VPRS) (Ziarko, 1993) proposed by Ziarko is a methodology in which the records within the dataset were analyzed and classified in terms of their statistical tendencies rather than their functional patterns (Ziarko, 1993; Ziarko, 2001). In VPRS theory, the uncertain nature of the information within the dataset of interest is handled using the concept of β-lower and β-upper approximate sets. In the stock portfolio selection system, the values of financial indices are continuous; the performance of VPRS models is basically resulting from the quality of the original clustering results. Attributes clustering must be performed in prior to conduct a continuous valued dataset classification, and correct partitioning is the prelude to available classifications. Accordingly, when continuous valued stock datasets with uncertain or missing information, it is preferable to utilize VPRS theory for classification purposes, and to integrate the VPRS model with some form of cluster generation / cluster index evaluation procedure such that the optimal discretizing solution can be obtained. Thus, in the Data Mining component, the number of clusters of GRGs is optimized using a VPRS-index method, which is applied to optimizing the number of clusters each attribute of instances within a dataset and the classification results of this dataset, and VPRS theory is then applied to identify the stocks within the β - lower approximate sets. These stocks are then processed by the GRA consolidation model in order to establish a single financial indicator for each stock item on which to base the stock selection decision.

LITERATURE REVIEW

Grey Relational Analysis (GRA)

GRA functions provide an effective means of solving multiple-criteria decision problems by ranking the feasible solutions associated with their so-called GRGs such that the optimal solution can be readily decided (Huang et al., 2008). In the proposed expert system for selecting stock portfolio in the present study, the GRA function is used to ease the stock classification and selection procedures by consolidating the values the multiple attributes of each instance into a single integrated attribute value describing one specific financial ratio of the stock item or indicating the ability of the stock item..

Index function $I_{\max}$ (Huang, 2010a; b)

The VPRS-index method proposed in this study partitions the dataset according to the values of the individual data attributes rather than that of the data norms. Suppose that each object x_i in the dataset has m conditional attributes and the l-th attribute a_l can be partitioned into p_l clusters, then $C_{a_l}(x_i)$ gives the index of the cluster to which the l-th attribute a_l of object x_i belongs. Here $C_{a_l}(x_i)$ is given by:

$$C_{a_l}(x_i) = I_{\max}(\mu_j(x_i(a_l))) = Index(\max(\mu_j(x_i))) \quad for \quad 1 \le l \le m, 1 \le i \le n,$$

where $I_{\max}(\mu_j(x_i(a_l)))$ returns the index of the cluster

corresponding to the maximum value amongst the membership functions value of the l-th attribute of x_i.

VPRS theory

The VPRS operates on what may be represented as a knowledge-representation system, or information system (Ziarko, 1993). The basic principles and notations of information systems (S) and the applications of VPRS theory to the processing of such systems are represented thus:

β-lower and -upper approximate sets

For a given dataset, any records which are indistinguishable from one another when evaluated using a specific subset of all the attributes define an equivalence or indiscernibility relationship. In VPRS theory, this indiscernibility concept is operated using approximate sets. A representative information system has the form $S = (U, A, V_q, f_q)$, where U is a non-empty finite set of records, A is a non-empty finite set of attributes describing these records and $X \subseteq U$ and $R \subseteq A$. Generally speaking, the attributes in set A can be partitioned into a set of conditional attributes $C \neq \phi$ and a set of decision attributes $D \neq \phi$, i.e. $A = C \cup D$ and $C \cap D = \phi$. For each attribute, $q \in A$, V_q represents the domain of q, i.e. $V = \cup V_q$. Finally, $f_q : U \times A \to V$ is an information function defined such that $f(x, q) \in V_q$ for $\forall q \in A$ and $\forall x \in U$.

The VPRS method used in this study applies the systematic method presented by the current author in (Huang, 2009b) to decide a suitable value of the threshold parameter β, i.e., the value of β at which a certain proportion of the records in a specific conditional class are classified into the same decision class. When processing an information system using a VPRS model with $0.5 < \beta \leq 1$, the objective is to recognize the β-lower and β-upper approximate sets in terms of each cluster of the decision attribute. In general, the β-lower approximation of sets $X \subseteq U$ and $P \subseteq C$ is given by:

$$_{\beta}\underline{R}_P(X) = \{x \in U : P(X/[x]_P) \geq \beta\} = \cup\{[x]_P : P(X/[x]_P) \geq \beta\}.$$

Similarly, the β-upper approximation of sets $X \subseteq U$ and $P \subseteq C$ can be expressed as:

$$_{\beta}\overline{R}_P(X) = \{x \in U : P(X/[x]_P) > 1-\beta\} = \cup\{[x]_P : P(X/[x]_P) > 1-\beta\}.$$

Note that $P(X/Y) = |X \cap Y|/|Y|$ if $|Y| > 0$, and $P(X/Y) = 1$ otherwise. Note also that $|x|$ indicates the cardinality of set X. In the specific case of $\beta = 1$, $_{\beta}\underline{R}_P(X)$ and $_{\beta}\overline{R}_P(X)$ are equivalent to the lower and upper approximate sets in RS theory. In other words, the VPRS model reverts to the traditional RS model.

Accuracy of VPRS classification results

The accuracy of the VPRS classification results can be quantified as follows:

$$_{\beta}\alpha_c = \left|_{\beta}\underline{R}_P(X)\right| \Big/ \left|_{\beta}\overline{R}_P(X)\right|,$$

where $X = \{x : C_d(x) = c, \forall x \in U\}$; and $|_{\beta}\underline{R}_P(X)|$ and $\left|_{\beta}\overline{R}_P(X)\right|$ are the cardinalities of the β-lower and β-upper approximate sets, respectively, when classifying the records (x) associated with the cth cluster of the decision attribute d.

Overview of PBMF and VP Cluster Index Functions

PBMF -index function

The PBMF cluster validity index function (Pakhira et al., 2004) assures the formation of a small number of compact clusters within the dataset and maximizes the separation distance between at least two of these clusters. The PBMF-index function is formulated as $PBMF(K) = \left(\frac{1}{K} \times \frac{\overline{E}_1}{J_{m'}} \times D_K\right)$, where K is the number of clusters, $J_{m'} = \sum_{k=1}^{K}\sum_{j=1}^{n} \mu_{kj}^{m'} \left\| x_j - z_k \right\|$, $\overline{E}_1$ is constant for a given dataset in which the instances belong to only one cluster and is set in such a way as to prevent the second term from vanishing, and $D_K = \max_{i,j=1}^{K} \left\| z_i - z_j \right\|$ is the maximum separation distance among all possible pairs of cluster center points in the dataset. In addition, n is the total number of objects in the dataset, $U(X) = [\mu_{kj}]_{K \times n}$ is a partition matrix, m' is the fuzzification parameter and z_k is the centroid of the k-th cluster. In applying the PBMF-index function in data clustering applications, the objective is to find the value of K which maximizes the index value.

VP- index function

In contrast to the PBMF-index function (Pakhira et al., 2004) which is based on the FCM clustering approach, the VP-index function applies the VPRS classification

scheme to extend applicability of the PBMF-index function to deal with the classification issue of vague information system. The VP- index function proposed in this study has the form:

$$VP(N_d, {}_{\beta}\alpha_c) = (\frac{1}{N_d} \times \frac{\overline{E_1}}{{}_{\beta}F'_{N_d}} \times D'_{\mathrm{N}_d}),$$

where N_d is the number of clusters of the conditional and decision attributes, and ${}_{\beta}\alpha_c$ is the accuracy of VPRS classification when evaluated associated with the c-th cluster of the decision attribute. In addition, ${}_{\beta}F'_{N_d}$ is obtained by accumulating the value of ${}_{\beta}E'_c$ for each cluster of the decision attribute (d), where ${}_{\beta}E'_c$ is given by ${}_{\beta}E'_c = \sum_{j=1}^{n} \left\| x'_{jc} \right\| \Big/ {}_{\beta}\alpha_c$, where $\mu_{cj}(x_j(d))$ is the membership function of instance x_j in the c-th cluster of the decision attribute d and z'_c is the multi-dimensional centroid of the lower approximate sets in terms of the c-th cluster of the decision attribute d and is obtained by calculating the mean values of the conditional and decision attributes of each record within the corresponding sets. Moreover, m' is the fuzzification parameter and n is the whole number of records in the dataset. Finally, the value of D'_{N_d} is equal to the maximum separation distance amongst the centroids of all the lower approximate sets in terms of the different clusters of the decision attribute, that is, $D'_{\mathrm{N}_d} = \max_{i,j=1}^{N_d} \left\| z'_i - z'_j \right\|$. Note that the value of D'_{N_d} is upper limited by the maximum separation distance amongst all possible pairs of records in the dataset.

Note that parameter ${}_{\beta}F'_{N_d}$ in the VP-index function differs slightly from parameter $J_{m'}$ in the PBMF-index function represented in earlier. The value of ${}_{\beta}F'_{N_d}$ depends on ${}_{\beta}\alpha_c$ in VP-index function, while the value of $J_{m'}$ does not consider the effect of α_c on the PBMF-index function value.

Comparison between VP- and PBMF-index functions

Table 1 summarizes the major components of the VP-index function and the PBMF-index function in order to highlight the differences between them. At a high level, three principal differences exist, namely (i) the VP-index function clusters the individual attributes of each instance within the dataset, whereas the PBMF-index function clusters the data based upon the norms of each instance; (ii) the VP-index function is based on z'_c, that is, the centroids of the lower approximate sets in terms of each cluster c of the decision attribute, whereas the PBMF-index function is based on z_k, i.e., the centroid of the k-th cluster obtained when clustering the dataset using the FCM method; and (iii).the VP-index function clearly considers the classification accuracy when measuring the optimality of the clustering results, whereas the PBMF-index function takes only account of the optimal number of clusters within the dataset.

METHODOLOGY

The VPRS-index method proposed in this study integrates the FCM clustering scheme, variable precision rough set (VPRS) theory and a modified form of the PBMF index function, designated as the VP-index function, in order to optimize both the number of clusters within the dataset and the corresponding classification accuracy. In the VPRS-index method, each attribute (both conditional and decision) is supposed to have an identical number of clusters and the objective is to map each attribute of element (X_i) in U to a suitable cluster amongst all the clusters in terms of the conditional ($C_1 \sim C_n$) or decision (d) attributes. The procedure of the VPRS-index method is showed in the subsequent sections.

Details of VPRS-index method

Figure 1 illustrates the basic structure of the proposed VPRS-index method. The details of each processing step are summarized in the following paragraphs.

Step 1: Specify number of clusters per attribute in an interval [2, $\mathrm{N}_{\max}$]

The VPRS-index method utilizes an iterative process to optimize the number of clusters designated to the conditional and decision attributes within the system of interest. (Notice that the number of decision attributes is defined by default as one.) The conditional and decision attributes are partitioned into an identical number of clusters, N, where N is limited by the interval [2, $\mathrm{N}_{\max}$], where 2 symbolizes the minimal number of clusters per attribute (and is the default setting) and $\mathrm{N}_{\max}$ symbolizes the maximum allowed number of clusters per attribute.

Step 2: Fuzzify attributes of information system using FCM method

In general, a continuous-valued information system can only be transformed into an equipollent fuzzy information system when a classified fuzzy set has been obtained. In the FCM is clustered by process performed in the VPRS-index method, the interval values $[\alpha, \beta]$ of the entire conditional and decision attributes are designated to p_l fuzzy clusters. The continuous-valued information system (U, A, V_q, f_q) is then transformed into the fuzzy information system $(U, \tilde{A}, \Phi, d)$, in which $\Phi = \{\tilde{A}_{lj} \mid l \le m, j \le p_l\}$,

Table 1. Detailed definitions of VP-index and PBMF-index.

Function	VP-index	PBMF-index
	$VP(N_d, \beta\alpha_c) = (\frac{1}{N_d} \times \frac{\overline{E_1}}{\beta F'_{N_d}} \times D'_{N_d})$	$PBMF(K) = \left(\frac{1}{K} \times \frac{E_1}{J_{m'}} \times D_K\right)$
	Cluster each attributes of a data	Cluster the data in a data set
	N_d is the number of clusters assigned to the conditional and decision attributes	K is the number of clusters of data set
How to cluster the data	$\beta F'_{N_d} = \sum_{c=1}^{N_d} \beta E'_c$, $\beta E'_c = \sum_{j=1}^{n} \overline{\mu}_{cj}^{m'}(x_j(d)) \left\| x_j - z'_c \right\| \Big/ \beta\alpha_c$ (1) $\overline{\mu}_{cj}(x_j(d))$ is the membership function of data object x_j in the c-th cluster of the decision attribute d. (2) z'_c is the multi-dimensional centroid of the lower approximate sets in terms of the is clustered by c-th cluster of the decision attribute *d* and is obtained by computing the mean values of the conditional and decision attribute values of each within the corresponding sets. (3.1) $\left\| x_j - z'_c \right\|$ is length of the vector (norm) between the x_j object and z'_c. (3.2) $\beta E'_c = \sum_{j=1}^{n} \left\| x'_{jc} \right\| \Big/ \beta\alpha_c$, where $\left\| x'_{jc} \right\| = \overline{\mu}_{cj}^{m'}(x_j(d)) \left\| x_j - z'_c \right\|$. (3.3) $\beta\alpha_c$ is the classification accuracy and indicates the cardinality proportion of β-lower approximates in β-upper approximates when evaluated in terms of the c-th cluster of decision-making attribute d.	$J_{m'} = \sum_{k=1}^{K} E_k$, $E_k = \sum_{j=1}^{n} \mu_{kj}^{m'} \left\| x_j - z_k \right\|$ (1) $\mu_{kj}^{m'}$ is the membership functions of the j-th data object. (2) z_k is the centroid of the k-th cluster obtained when the dataset using the FCM method. (3,1) $\left\| x_j - z_k \right\|$ is length of the vector (norm) between the x_j data object and z_k. (3.2) $E_k = \sum_{j=1}^{n} \left\| x_{jk} \right\|$, where $\left\| x_{jk} \right\| = \mu_{kj}^{m'} \left\| x_j - z_k \right\|$.
	$D'_{N_d} = \max_{i,j=1}^{N_d} \left\| z'_i - z'_j \right\|$ is equal to the maximum separation distance amongst the centroids of all the lower approximate sets associated with the different clusters of the decision attribute	$D_K = \max_{i,j=1}^{K} \left\| z_i - z_j \right\|$ is equal to the maximum separation distance between the cluster centroids

where $\tilde{A}_{lj} = \mu_j(x_i(a_l))$ denotes the values of the membership functions in terms of the l-th conditional attribute a_l of the i-th object.

Step 3:: Assign each attribute of each instance to appropriate conditional or decision attribute cluster

Applying the index function $I_{\max}(\mu_j(x_i(a_l)))$ for $1 \le l \le m, 1 \le i \le n$, the membership functions of each attribute of each instance are processed in order to decide the conditional or decision attribute cluster to which they belong.

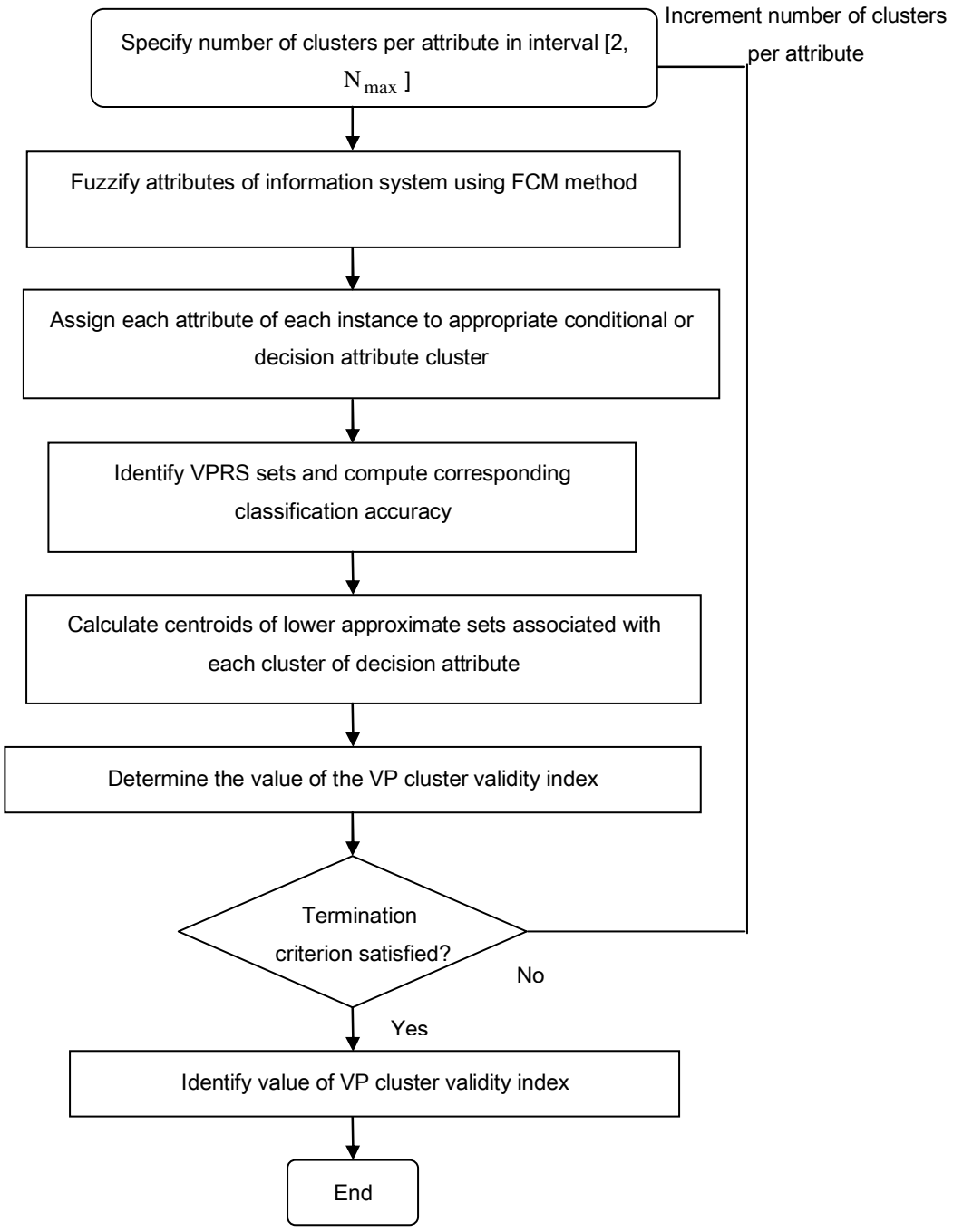

Figure 1. Flow chart showing basic steps in proposed VPRS-index method.

Step 4: Identify VPRS sets and compute corresponding classification accuracy

Having mapped the attribute values of all the instances to the suitable conditional or decision attribute clusters, the β-lower and β-upper approximate sets in terms of each cluster *c* of the decision attribute *d* are extracted according to the definitions presented in earlier. The accuracy of VPRS classification in terms of each cluster of the decision attribute is then obtained by computing the cardinality ratio of the corresponding β-lower approximate sets to the β-upper approximate sets.

Step 5: Calculate centroids of lower approximate sets associated with each cluster of decision attribute

The multi-dimensional centroids of the lower approximate sets in terms of each cluster of the decision attribute *d* are obtained by

Table 2. Membership function values of each attribute of each instance.

Code of instance	Conditional attribute				Decision attribute	
	a_1		a_2		d	
1	0.025	0.975	0.061	0.939	0.010	0.990
2	0.063	0.937	0.025	0.975	0.015	0.985
3	0.988	0.012	0.939	0.061	0.985	0.015
4	0.992	0.008	0.974	0.026	0.990	0.010

Table 3. β-lower and -upper approximate sets associated with c -th decision attribute.

Code of instance	β-lower approximate sets $\beta\underline{R}(X:C_d(\mathrm{x})=c, x\in X)$			
1	2	2	2	$\beta\underline{R}(X:C_d(\mathrm{x})=2, x\in X)$
2	2	2	2	
3	1	1	1	$\beta\underline{R}(X:C_d(\mathrm{x})=1, x\in X)$
4	1	1	1	

Each of the β -lower approximate sets $\beta\underline{R}(X:C_D(\mathrm{x})=c, x\in X)$ is equal to the corresponding β -upper approximate set $\beta\overline{R}(X:C_D(x)=c, x\in X)$.

calculating the mean attribute values (both conditional and decision) of all of the instances within the corresponding sets.

Step 6: Determine the value of the VP cluster validity index

Having determined the accuracy of VPRS classification and centroids of the lower approximate sets, the optimality of the clustering and classification results is assessed using the VP- index function.

Step 7: Check termination criterion

A check is made to see whether *N* is equal to the upper bound value (*N*=39) when the value of the index function had calculated for the current number of clusters per attribute *N*. In the event that *N* is not equal to $\mathrm{N}_{\max}$, the value of *N* is increased by 1, then the FCM, VPRS and cluster validity index computation procedures are iterated. The iteration procedure terminates and the computational process moves to the final step when the termination criterion is satisfied.

Step 8: Identify value of VP cluster validity index

Once the termination criterion has been satisfied, the values of the VP-index function obtained for *N* = 2~ $\mathrm{N}_{\max}$ are compared. The maximum value of the index function is taken as the VP cluster validity index and corresponds to the clustering solution which optimizes both the number of clusters per attribute and the entire accuracy of VPRS classification of the dataset.

A step-by-step example showing calculation of VP-index value

This section illustrates the derivation of the VP-index value for a simple hypothetical dataset comprising just four entries. An assumption is made that each entry has two conditional attributes, a_1, a_2, and one decision attributes, d. Let the four instances be defined as $x_1(1.90,1.30,0.75)$, $x_2(2.10,1.20,0.65)$, $x_3(2.45,1.45,0.30)$ and $x_4(2.55,1.55,0.20)$, respectively.

According to the VPRS-index method, a repeated process is applied. Initialize that each conditional and decision attribute is partitioned into 2 clusters. Then, the continuous-valued data in the hypothetical dataset are discretized using the FCM technique. The membership function values of each attribute of each instance are summarized in Table 2. The attribute values of each instance are then appointed to suitable conditional or decision attribute clusters by applying the index function $I_{\max}$ to the corresponding membership function values. The mapping results are shown in Table 3. As shown, the discretized vectors of the four instances x_i (I_{a_1}, I_{a_2}, I_d) have the form $x_1(2,2,2)$, $x_2(2,2,2)$, $x_3(1,1,1)$, and $x_4(1,1,1)$, respectively. The β-lower and β-upper approximate sets in terms of each cluster of the decision attribute are computed according to the formulation given earlier of (Huang, 2009b) and are also shown in Table 3.

Moreover, the threshold parameter β in terms of first and second clusters of the decision attribute are determined according to the procedure given earlier of Ref (Huang, 2009b) and are 0.939 and 0.974, respectively. Therefore, the β-upper and β-lower approximate sets obtained using VPRS are the same as the upper and lower approximate sets obtained using RS. The accuracy of VPRS classification in terms of each cluster of the decision attribute is obtained by counting the cardinality ratio of the corresponding β-lower approximate sets to the β-upper approximate sets. In the present example, the accuracies of VPRS classification are therefore equal to α_1=2/2=1.000 and α_2 = 2/2=1.000, respectively.

Table 4. Values of $\left\|x'_{jc}\right\| (= \mu^2_{cj}(x_j(d)) \times \left\|x_j - z'_c\right\|)$.

x_j	z'_c	
j	c=1	c=2
1	0.000	0.120
2	0.000	0.119
3	0.084	0.000
4	0.085	0.000
$\sum_{j=1}^{4} \left\|x'_{jc}\right\|$	0.169	0.239

Then, the RS procedure is applied to determine the multi-dimensional centroids of the lower approximate sets in terms of each cluster of the decision attribute by computing the mean attribute values (both conditional and decision) of all the instances within the corresponding sets. Thus, in the present example, the centroids of the lower approximate sets in terms of the two cluster of the decision attribute are obtained as $z'_2 = mean(x \mid x \in \underline{R}(X), C_d(x) = 2) = mean(x \mid x \in \{x_1, x_2\}) =$ $((1.90+2.10)/2, (1.30+1.20)/2, (0.75+0.65)/2) =$ $(2.00, 1.25, 0.70)$ and $z'_1 = mean(x \mid x \in \underline{R}(X), C_d(x) = 1) =$ $mean(x \mid x \in \{x_3, x_4\}) = (2.50, 1.50, 0.25)$, respectively.

Having decided the membership function values of all the instances, the accuracy of VPRS classification, and the centroids of the lower approximate sets, the optimality of the discretization / classification consequence is evaluated using the VP-index function (that is, $VP(N_d, \beta\, \alpha_c) = (\frac{1}{N_d} \times \frac{\overline{E_1}}{\beta\, F'_{N_d}} \times D'_{N_d})$). In picturing the derivation of $\beta\, F'_{N_d}$ (where $\beta\, F'_{N_d} = \sum_{c=1}^{N_d} \beta\, E'_c$), the subsequent discussions arbitrarily consider the computation of $\beta\, E'_1$. (Note, that $\beta\, E'_2$ is computed in a similar manner.). The first instance in the dataset, x_1, has attribute values of $x_1(1.90, 1.30, 0.75)$. In addition, the centroid of the lower approximate sets in terms of the first cluster of the decision attribute is given by $z'_1(2.50, 1.50, 0.25)$. As a result, $(x_1(a_1) - z'_1(a_1)) = (1.90 - 2.50) = -0.60$, $(x_1(a_2) - z'_1(a_2)) =$ $(1.30 - 1.50) = -0.20$, and $(x_1(d) - z'_1(d)) = (0.75 - 0.25) = 0.50$. Thus, the vector of $x_{11} = x_1 - z'_1$ has the form $[x_{11}(a_1), x_{11}(a_2), x_{11}(d)] = [-0.60, -0.20, 0.50]$, and the corresponding norm is equal to $\left\|x_1 - z'_1\right\| = \sqrt{x_{11}(a_1)^2 + x_{11}(a_2)^2 + x_{11}(d)^2} = \sqrt{(-0.60)^2 + (-0.20)^2 + 0.50^2} = 0.806$. Let the fuzzification parameter m' be defined as 2.0. Applying the notation $\left\|x'_{j1}\right\| = \mu^2_{1j}(x_j(d)) \times \left\|x_j - z'_1\right\|$, the effect of instance x_1 on z'_1, that is, $\left\|x'_{11}\right\|$, is obtained by multiplying $\left\|x_1 - z'_1\right\|$ by the square of the corresponding membership function value, i.e., $\mu^2_{11}(x_1(d)) = 0.010^2 = 0.000$. Thus, $\left\|x'_{11}\right\|$ has a value of 0.000. $\left\|x'_{21}\right\|$, $\left\|x'_{31}\right\|$ and $\left\|x'_{41}\right\|$ are calculated using the same procedure. The corresponding results are shown in Table 4. The value of $\beta\, E'_1$ is thus obtained as $\beta\, E'_1 =$ $(\sum_{j=1}^{4} \mu^2_{1j}(x_j(d)) \left\|x_j - z'_1\right\|) \Big/ \beta\, \alpha_1 = (\sum_{j=1}^{4} \left\|x'_{j1}\right\|) \Big/ \beta\, \alpha_1 =$ $(\left\|x'_{11}\right\| + \left\|x'_{21}\right\| + ... + \left\|x'_{41}\right\|) / \beta\, \alpha_1 =$ (0.000+0.000+0.084+0.085)/ 1.000 = 0.169. Utilizing the same approach to that described above, the value of $\beta\, E'_2$ is obtained as 0.239. $\beta\, F'_{N_d}$ is thus found to have a value of $\beta\, F'_2 = \sum_{c=1}^{2} \beta\, E'_c$ =0.408.

Factor $\overline{E_1}$ in the VP-index function is a constant for a given dataset in which the instances belong to only one cluster. Consequently, the attribute values of the centroid z_1 of the illustrative dataset can be calculated using the arithmetic mean function $mean(x \mid x \in \{x_i\}, i = 1,2,...,4)$ as $((1.90 + 2.10 + 2.45 + 2.55), (1.30 + 1.20 + 1.45 + 1.55), (0.75 + 0.65 + 0.30 + 0.20))$ $= z_1(2.25, 1.375, 0.475)$. Based on the vector of centroid z_1, it can be shown that $(x_1(a_1) - z_1(a_1)) = (1.90 - 2.250) = -0.350$, $(x_1(a_2) - z_1(a_2)) = (1.30 - 1.375) = -0.075$, and $(x_1(d) - z_1(d)) = (0.75 - 0.475) = 0.275$. Thus, the vector of $x_{11} = x_1 - z_1$ has the form $[x_{11}(a_1), x_{11}(a_2), x_{11}(d)] = [-0.350, -0.075, 0.275]$, and the corresponding norm is equal to $\left\|x_1 - z_1\right\| = \sqrt{x_{11}(a_1)^2 + x_{11}(a_2)^2 + x_{11}(d)^2} = \sqrt{(-0.350)^2 + (-0.075)^2 + 0.275^2} =$ 0451. Similarly, the norms of $\left\|x_2 - z_1\right\|$, $\left\|x_3 - z_1\right\|$ and $\left\|x_4 - z_1\right\|$ are found to be 0.289 and 0.443, respectively. The value of

$\overline{E_1}$ in the VP-index function is then obtained by summing the norms of $\|x_j - z_1\|$ where $j = 1,2,...,4$, yielding a value of $\overline{E_1}$ = 1.460.

The value of D'_{N_d} in the VP-index function is obtained by calculating the maximum separation distance between the centroids of the lower approximate sets in terms of the first and second clusters of the decision attribute. In the present example, these centroids are gven by $z'_1(2.50,1.50,0.25)$ and $z'_2(2.00,1.25,0.70)$, respectively. Thus, the vector of $z_{12} = z'_1 - z'_2$ which maximizes the value of $D'_{N_d} = \max_{i,j=1}^{N_d} \|z'_i - z'_j\|$ has the form $[z_{12}(a_1), z_{12}(a_2), z_{12}(d)] = [-0.50, -0.25, 0.45]$. The corresponding norm is thus equal to $\sqrt{(-0.50)^2 + (-0.25)^2 + 0.45^2} = 0.718$.

Given the parameter values specified / derived above (that is, $N_d = 2$, $\overline{E_1}$ = 1.460, $\beta F'_2$ =0.408 and $D'_{N_d} = 0.718$), the VP-index function $(VP(N_d, \beta\alpha_c) = (\frac{1}{N_d} \times \frac{\overline{E_1}}{\beta F'_{N_d}} \times D'_{N_d}))$ returns a value of 1.284.

RESULTS

Modeling an expert system for selecting stock portfolio

The VPRS-index method is combined with a GRA model and a VPRS classification scheme to yield an expert system for automatic selecting stock portfolio. In the following paragraphs, the detailed processing steps are discussed and the performance of the proposed expert system for selecting stock portfolio is evaluated.

GRA attribute consolidation / reduction mechanism

In the proposed expert system for automatic selecting stock portfolio, a GRA model is used in two stages: (1) it is used initially to consolidate the attributes (financial indices) of the stock items to ease the clustering task in the VPRS-index method; and (2) it is then used to consolidate the six attributes of the stock items filtered according to Buffet's principles to yield a single performance measure upon which to evaluate the merit of each stock item within the selected stock portfolio.

In this study, this expert system is assumed to have the form $S = (U, A, V_q, f_q)$, where U is a non-empty finite set of objects (stock items) and A is a finite set of attributes (financial indices) describing these objects. Following the application of the GRA model, a modified expert system with the form $S = (U, \hat{A}, V_q, f_q)$ is obtained, in which $\hat{A}$ is a set of six consolidated attributes (financial ratios) representing the same set of objects, and the residual notations are as represented previously. The six financial ratios are clustered using the VPRS-index method and the cluster indices coinciding with the optimal clustering solution are then processed using VPRS theory in order to recognize the corresponding β-lower approximate sets. Supposing that U is the domain of discourse and R is the set of equivalences of U, the VPRS problem can be formulated as $X \subseteq U$ is : $(_{\beta}\overline{R}(X), _{\beta}\underline{R}(X))$.

As represented earlier, the GRA model is also used to consolidate the six financial ratios of each stock item remaining after the stocks within the β-lower approximate set have been filtered using the general investment principles directed by Buffet. In this case, the GRA model takes the six consolidated financial ratios (GRGs) of each stock item as the input and generates a single GRG which denotes the global performance of the corresponding stock item. The GRGs are ranked in descending order such that the stock items with a better financial performance are placed above those with a poorer performance and the ranked sequence is then taken as the input to the final stock selection decision.

Filtering of stock items according to basic investment principles

To ease the workload of the GRA model in consolidating the six GRGs of each stock item to a single performance indicator, the stocks within the β-lower approximate sets are filtered according to a set of decision-making attributes which are defined according to the general investment principles specified by Buffett and formalized by Hagstrom (Hagstrom et al., 2005). Buffett debated that reducing costs is necessary for enterprises seeking to hone their competitive ability and to rival their competitors in terms of price, while high profit margins and a high inventory turnover are both reliable indicators of the financial prosperity of a company. Buffet further affirmed that only companies with all three attributes can be certain of survival and will possess the means to earn profit for their shareholders. For this reason, in the present study, the stock items within the β-lower approximate sets recognized by the VPRS classification model are filtered according to the threshold values of attributes "return on asset (after tax) greater than zero", "return on equity greater than zero", "gross profit ratio greater than zero", "equity growth rate greater than zero" and "constant EPS greater than zero".

Data extraction

In this study, the feasibility of the proposed expert system

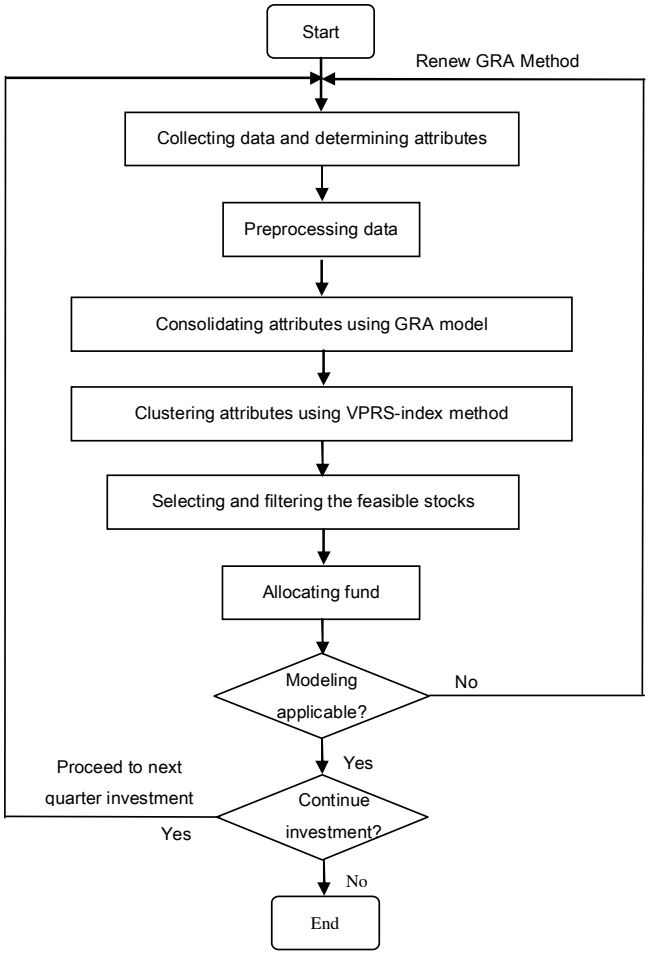

Figure 2. Flow chart of proposed expert system for selecting stock portfolio.

for selecting stock portfolio was assessed using electronic stock data extracted from the TEJ database over the period between the first quarter of 2004 to 6/1/2009. In general, financial statements for a particular accounting period are subject to a certain delay before publication. A detailed description of submission deadlines for the financial statements maintained in the TEJ database are presented in (Huang and Jane, 2009). Since the financial data relating to the last quarter in every year is not available until May 31st in the following year, the data cannot be used by this expert system to select suitable investment stocks in the first quarter. Consequently, the expert system can only be performed three times in every 12 month period, namely 5/31-09/22, 9/22 - 11/15 and 11/15 - 05/31 the following year.

Detailed processing steps in VP-index function based expert system for selecting stock portfolio

The detailed processing steps in the proposed expert system for selecting stock portfolio are illustrated in Figure 2 and are summed up as follows:

Step 1: Collecting data and determining attributes

In each quarter, the 53 attributes of each specified stock item within the TEJ database are collected automatically, and the user is given the opportunity (1) to modify the choice of financial ratios used for attribute consolidation / reduction in the initial GRA process; (2) to select a new GRA model for attribute consolidation / reduction purposes; and (3) to modify the decision-making attributes used to filter the stocks in the β-lower approximate set prior to their further consolidation using the GRA model.

Step 2: Preprocessing data

Having collected the pertinent financial data for each quarterly period, a basic pre-processing operation is implemented to improve the efficiency of the GRA attribute

Table 5. Mapping of 53 financial indices to 6 financial ratios.

Profitability	Rate Per Share	Growth Rate	Credit Capacity	Operating Capacity	Statutory Ratio
Return On Assets %-EBIT	BPS (A)	YOY%-Sales	Current Ratio	Inv.&A-R /Equity	Sales Per Employee
Return on Equity %	EPS-Net Income	%Gross Margin Growth	Acid Test	Total Asset Turnover	Operation Inc./Emply
Gross Margin %	PS-Cashflow	YOY%-Real. GM	Interest Exp. %	A/R&N/R Turnover	Fixed Assets/Emply
Real. Gross Profit %	PS-Sales	YOY%-Oper. Income	D/E Ratio	Days-A/R Turnover	PBR
Operating Income %	PS-Operating Income	YOY%-Pre-Tax Income	Liabilities %	Inventory Turnover	
Pre-Tax Income	PS-Pretax Income	YOY%-Ordin. Income	Equity/TA %	Days-Inventory Turn.	
Net Non-op.Inc./Rev. %		Net Income Gth%- After Tax	(L-T Liab.+SE)/FA %	Fixed Asset Turnover	
Net Income%-Exc Disp (After Tax)		YOY%-Total Assets	Debt/Equity %	Equity Turnover	
		YOY%-Total Equity	Oper. Income/Capital	Days-A/P Turnover	
		Depreciation YOY%-Fixed Assets	Pre Tax Income/Capital	Net Operating Cycle	
		YOY%-Return on TA			
		Retention Ratio			
		QOQ%-Sales			
		QOQ%-Operating Inc.			
		QOQ%-Net Income			

GRA attribute consolidation / reduction process. Particularly, the data instances containing missing fields (that is, missing financial indices) are immediately discarded, and the Box Plots method (Chakravarti, 1967) is applied to resolve the data outlier problem by establishing an inter-quartile range such that any data points falling outside this range can be automatically designated a default value relying on the interval within which they fall.

Step 3: Consolidating attributes using GRA model

For the stock records which are left after the pre-processing operation, the GRA model normalizes the values of each of the 53 financial indices and then calculates the six corresponding financial ratios according to the mapping given in Table 5. Note that the 53 financial indices are categorized into 6 predetermined financial ratios.

Step 4: Clustering attributes using VPRS-index method

In order to recognize the optimal number of clusters per attribute (conditional and decision) and the corresponding set of cluster indices, the VPRS-index method is applied to processing the values of the six financial ratios obtained in Step 3 (that is, five conditional attributes $C_1 \sim C_5$ and one decision attribute D_1).

Step 5: Selecting and filtering the feasible stocks

To recognize the stock items within the β-lower approximate sets, the VPRS-index method is used to process the optimal set of cluster indices. Then, in order to identify the final set of stocks for possible inclusion within the investment portfolio, these stock items are filtered according to the general investment guidelines proposed by Buffett.

Step 6: Allocating fund

The GRA model is again applied to consolidate the six financial ratios of each stock item remaining after the filtering operation to a single GRG (i.e., a global performance indicator). The GRGs of all the surviving stock items are then arranged in descending order and the first five stock items are chosen for stock portfolio.

Table 6. Illustrative financial ratio values (GRGs) obtained using GRA model to consolidate the attributes (financial indices) of financial data extracted from TEJ database for second quarter in 2007.

Company code	(a)	(b)	(c)	(d)	(e)	(f)
1	0.8689	0.5572	0.8142	0.8618	0.8677	0.8416
2	0.8915	0.5583	0.8164	0.8590	0.8658	0.8069
3	0.8558	0.5595	0.8156	0.8613	0.8661	0.8113
4	0.8544	0.5495	0.8094	0.8617	0.8659	0.8091
…	…	…	…	…	…	…
595	0.8932	0.5658	0.8134	0.8617	0.8753	0.8085
596	0.8714	0.5457	0.8093	0.8644	0.9074	0.8030
597	0.8704	0.5797	0.8156	0.8624	0.8605	0.8101
598	0.8930	0.5528	0.8099	0.8601	0.8847	0.8165

The attributes of columns are (a) Rate Per Share (b) Growth Rate (c) Credit Capacity (d) Operating Capacity (e) Statutory Ratio (f) Profitability.

Step 7: Checking the validity of modeling

The rate of return on the stock portfolio composed at the end of quarter *k* is compared at the end of quarter *k+1* with the average rate of return implied by the variation in the Taiwan Stock Exchange Capitalization Weighted Stock Index (TAIEX) over the equivalent financial period. A decision is made as to whether or not the model should be run for a further quarter using the existing GRA model when the rate of return is acceptable. However, the suitability of the GRA model will be reviewed and a new GRA model will be adopted if appropriate when the rate of return is deemed unacceptable. In the next two instances, the GRA model is used to compute the following financial ratios: (1) the profitability, (2) the rate per share, (3) the growth rate, (4) the credit capacity, (5) the operating capacity, and (6) the statutory ratio, where ratio (1) is treated as the decision attribute of the stock system and ratios (2)-(5) are treated as the conditional attributes. (Note that the mapping of the 53 financial indices to the six consolidated ratios is summarized in Table 5).

Performance evaluation of VPRS-index classification method

The classification results obtained through VPRS theory could profoundly affect he performance of the proposed expert system for selecting stock portfolio. Therefore, this section commences with an example showing the validity and effectiveness of the VPRS-index method and the performance of the proposed expert system is then evaluated in next section. The validity and effectiveness of the proposed VPRS-index method is evaluated by an illustrative example relating to electronic stock data extracted from the financial database maintained by the Taiwan Economic Journal (TEJ) (Huang and Jane, 2009; Huang, 2009b) for the second quarter of 2007. In the proposed classification method, a specified set of stock items are collected automatically every quarter and 53 financial indices associated with each stock item are consolidated into 6 normalized financial ratios (GRGs) using a GRA model. A total of 598 GRG records were obtained (Table 6 for indicative values of each ratio for a selected subset of these 598 records) as the records for which some of the financial data was incomplete had deleted.

In performing the evaluations, the effectiveness of the proposed classification method is explored by comparing the classification results with the results obtained from pseudo-supervised classification method. The VPRS-index method provides the means to discretize the continuous values of the separate attributes within a dataset and to classify datasets where the records do not provide any category information. In contrast, supervised classification methods cluster attributes based on a consideration of category information. There are currently no classifiers available for the supervised classification of datasets with no category (class) information. Consequently, it is not possible to set up a straight comparison between the classification performance obtained by the VPRS-index method and those obtained from a supervised method. Therefore, in this illustrative example, the classification performance of the VPRS-index method is compared with those of pseudo-supervised decision-tree classification method, in which pseudo-category information is joined to a dataset which originally lacks category information. The pseudo-category information is obtained by applying the VPRS-index method to the target dataset in order to recognize the optimal number of clusters for the decision attribute. The $I_{\max}$ function presented earlier is then used to acquire the suitable decision attribute cluster for each record in the dataset. The resulting cluster index is then treated as pseudo-category information for the record. In this illustrative example, the VPRS-index method and the pseudo-

Table 7. Comparison of classification accuracy (CA) obtained from VPRS-index method and pseudo-supervised decision-tree classification method for 10-fold subsamples.

	VPRS-index method		#Pseudo-supervised decision-tree classification method	
i th subsample	Training dataset	Testing dataset	Training dataset	Testing dataset
1	0.9814	1.0000	0.2881	0.1000
2	0.9814	1.0000	0.2714	0.1000
3	0.9814	1.0000	0.3717	0.1000
4	0.9926	1.0000	0.2695	0.1167
5	0.9851	1.0000	0.2844	0.0667
6	0.9814	1.0000	0.2546	0.0500
7	0.9926	1.0000	0.3271	0.0667
8	0.9851	1.0000	0.3178	0.0500
9	0.9814	1.0000	0.3135	0.1525
10	0.9629	1.0000	0.3117	0.0678
Average CA	0.9825	1.0000	0.3010	0.0870
Deviation of CA (%)	0.82	0.0	3.4	3.2

supervised decision-tree classification method are used to classify training and testing datasets based upon a common 10-fold subsample of the stock market dataset. The optimal number of clusters for the decision attribute in this dataset is equal to 12, and therefore the pseudo-category information joined to the dataset to facilitate discretizing using the decision-tree classification method has a value in the interval [1,12]. A common k-fold subsample (^{k}=10) was used to verify the performance of a classification method. Of the k subsamples, one subsample was held for use as validation data in testing the method, while the remaining $k-1$ subsamples were used as training data.

The classification performance of the two methods is assessed in terms of the classification accuracy (CA). For the case of the VPRS-index method, CA is defined as the ratio of the entire cardinality of the β-lower approximation sets in terms of each cluster of the decision attribute to the total number of samples in the dataset, that is. $\sum_{c=1}^{N_d} \left| \beta \underline{R}_P(X) \right| \Big/ |U|$. Meantime for the pseudo-supervised decision-tree classification method, the CA is defined as the ratio of the number of records for which the measured category information is identical to the added pseudo-category information to the total number of records in the dataset.

The CA, the average CA and the deviation of the CA obtained for the training and testing datasets by the VPRS-index method and pseudo-supervised decision-tree classification method are shown in Table 7. It can be found that the CA obtained for each training data and testing data obtained by the VPRS-index method is higher than those by the pseudo-supervised decision-tree classification method. Meanwhile, it can be seen that the VPRS-index method yields an average CA of 0.9825 for the training dataset and 1.0000 for the testing dataset. In contrast, the pseudo-supervised decision-tree classification method yields average CAs of 0.3010 and 0.0870 for the training dataset and testing dataset, respectively. In other words, the average CA obtained by the VPRS-index method is higher than those obtained by the pseudo-supervised decision-tree classification method for both datasets, respectively. In addition, it is seen that the lowest CA values obtained through the VPRS-index method for the training and testing datasets (i.e. 0.9814 and 1.0000, respectively) are higher than those obtained through the pseudo-supervised decision-tree classification method. Consequently, the performance of the VPRS-index method in optimizing the accuracy of VPRS classification using a VPRS classification model is superior to those of the pseudo-supervised classification decision-tree method where a pseudo number of clusters is assigned to the decision attribute, respectively.

Performance evaluation of proposed expert system for selecting stock portfolio

The validity and effectiveness of the proposed expert system is evaluated by comparing the rate of return on the investment portfolios selected in the 15 investment periods between 2004 and 2009 with the rate of return on the equivalent investment portfolios selected using a system where (1) the VPRS-index method is replaced by a Fuzzy C-Means clustering scheme where the number of clusters per attribute, obtained using GM(1,N) function, is just defined in advance as N = 3; and (2) the VPRS classification method is replaced by the RS classification method. Additionally, the rate of return obtained using the two stock selection schemes is compared with the average rate of return predicted by the variation in the TAIEX index over the equivalent investment periods. The

Table 8. Rates of return of TAIEX, GM (1, N)-based reduction attributes method, MVAR-MRR method, pre-determined cluster based stock selection scheme, and VPRS-based stock selection scheme.

Investment period	TAIEX	GM(1,N)-based reduction attributes method in Ref (Huang and Jane, 2009)	MVAR -MRR method in Ref (Huang, 2009a)	Clusters pre-determined	VPRS index function based
04/05/31~04/09/21	-0.48	-2.58	-5.50	-0.19	20.22
04/09/21~04/11/15	-0.72	-0.92	0.96	-4.33	5.75
04/11/15~05/05/31	1.78	21.02	24.13	-0.47	7.39
05/05/31~05/09/21	0.93	11.88	1.07	8.60	8.74
05/09/21~05/11/15	-0.60	0.67	-8.64	-1.73	-0.30
05/11/15~06/06/01	13.96	13.35	28.41	14.56	18.81
06/06/01~06/09/21	0.25	10.35	6.67	-11.21	8.99
06/09/21~06/11/15	5.04	1.15	-3.47	-7.86	-5.92
06/11/15~07/05/31	12.55	27.48	42.59	82.01	44.23
07/05/31~07/09/21	11.79			-1.57	17.73
07/09/21~07/11/15	-2.20			2.90	4.31
07/11/15~08/06/02	-2.03			2.29	-10.72
08/06/02~08/09/22	-29.96			-22.36	-24.73
08/09/22~08/11/17	-27.34			-11.62	29.04
08/11/17~09/06/01	56.63			49.4	50.22
Accumulated rate of return	39.60	82.45	86.22	98.42	115.68

corresponding results are showed in Table 8. It can be shown that the accumulated rate of return attained using the proposed VP-index function based mechanism (115.68%) is higher than that attained using the pre-determined clustering based scheme (98.42%) and is also higher than the accumulated rate of return implied by the variation in the TAIEX index (39.60%). Further, in the period 2004 to 2006, the accumulated rate of return attained using the VP-index function based mechanism (107.93%) is higher than that attained using the GM(1,N) attribute reduction based scheme (Huang and Jane, 2009) (82.45%) or the MVAR-MRR method (Huang, 2009a) (86.22%). In the meantime, the rates of return attained in 2004, 2005 and 2006 using the VP-index function based mechanism are 33.36, 27.27 and 47.30, respectively. By contrast, the rates of return attained using the GM (1, N) attribute reduction based scheme are 17.57, 25.90 and 38.98, respectively; while those attained using the MVAR-MRR method based scheme are 19.59, 20.84 and 45.79, respectively. In other words, the rates of return attained using the proposed stock selection scheme are higher than those obtained using the GM(1,N) attribute reduction scheme or the MVAR-MRR based scheme. Thus, the overall viability and effectiveness of the proposed Stoczk selection system is validated.

Conclusions

This study has presented an expert system for automatically selecting stock portfolio based on a Grey' Relational Analysis (GRA) model, a modified form of the PBMF- index function (designated as the VP-index function), and VPRS theory. In this paper, we present a GRA consolidation method, confirm the validity of the VPRS-index method and demonstrate the feasibility of the proposed expert system for selecting stock portfolio. Two of these findings are worth summarizing:

(1) The VPRS-index method is applicable to continuous valued datasets where the records do not provide any category information and may be imprecise and vague. It is not possible to establish a direct comparison between the classification results of the VPRS-index method and those of supervised methods since supervised methods rely on categorical information to cluster the attributes. However, it has been shown that the accuracy of VPRS classification of the VPRS-index method is better than those of pseudo-supervised decision-tree classification method when applied to a dataset to which pseudo-category information is joined to each record in order to facilitate classification.
(2) The expert system for selecting stock portfolio based on the VPRS-index method yields a higher rate of return than several existing portfolio selection systems. Moreover, the rate of return on the selected stock portfolio is noticeably higher than that predicted by the total variation in the TAIEX index over the equivalent investment period.

On the whole, the results presented in this study show that the proposed VPRS-index method provides an effective tool for optimizing both the number of attribute clusters and the accuracy of VPRS classification when

applied to the partitioning and classification of complex, real-world knowledge-based systems. A significant improvement in classification results was obtained. As a result, the VPRS-index method provides an ideal basis for such expert systems as automatic portfolio selection mechanisms (proved in this study), daily electrical peak load forecasting, seismic pattern discovery or remote sensing image of a city.

REFERENCES

Box GEP, Jenkins GM (1976). Time series analysis: forecasting and control. San Fransisco, CA: Holden-Day.

Chakravarti IM, Laha RG (1967). Handbook of Methods of Applied Statistics. John Wiley and Sons.

Deng J (1982). Control problems of grey system. Syst. Control Lett., 1(5): 288-294.

Deng J (1985). Relational space of grey systems. Fuzzy Math., 2: 1-10.

Hagstrom RG, Miller B, Fisher KL (2005). The Warren Buffett Way: Investment Strategies of the World's Greatest Investor. John Wiley and sons (ASIA) PTE LTD.

Hassan MR, Baikunth N, Kirley M (2007). A fusion model of HMM, ANN and GA for stock market forecasting. Expert Syst. Appl., 33: 171-180.

Huang KY, Jane CJ, Chang TC (2008). A novel approach to enhance the classification performances of grey relation analysis. J. Inform. Optim. Sci., 29: 1169-1191.

Huang KY, Jane CJ (2009). A Hybrid Model for Stock Market Forecasting and Portfolio Selection Based on ARX- Grey System and RS Theories. Expert Syst. Appl., 36: 5387-5392.

Huang KY (2009a). A Hybrid GRA / MV Model for the Automatic Selection of Investment Portfolios with Minimum Risk and Maximum Return. J. Grey Syst., 21(2): 149-166.

Huang KY (2009b). Application of VPRS model with enhanced threshold parameter selection mechanism to automatic stock market forecasting and portfolio selection. Expert Syst. Appl., 36: 11652-11661.

Huang KY (2010a). Applications of an Enhanced Cluster Validity Index method based on the Fuzzy C-means and Rough Set Theories to Partition and Classification. Expert Syst. Appl., 37: 8757-8769.

Huang KY (2010b). A Hybrid Particle Swarm Optimization Approach for Clustering and Classification of Datasets. Knowledge-Based Systems, doi: 10.1016/j.knosys.2010.12.003.

Jandaghi G, Tehrani R, Hosseinpour D, Gholipour R, Shadkam SAS (2010). Application of Fuzzy-neural networks in multi-ahead forecast of stock price. A. J. Bus. Manage., 4(6): 903-914.

Lin SL (2010). A two-stage logistic regression-ANN model for the prediction of distress banks: Evidence from 11 emerging countries. Afr. J. Bus. Manage., 4(14): 3149-3168.

Nagai MT, Yamaguchi DS, Li GD (2005). Grey structural modeling. J. Grey Syst., 8(2): 119-130.

Pakhira MK, Bandyopadhyay S, Maulik U (2004). Validity index for crisp and fuzzy clusters. Pattern Recognit., 37: 487-501.

Pawlak Z (1982). Rough sets. Int. J. Inform. Comput. Sci., 11(5): 341-356.

Pawlak Z (1994). Rough set approach to multi-attribute decision analysis. Eur. J. Oper. Res., 72: 443-459.

Skalko C (1996). Rough sets help time the OEX. J. Comput. Intel. Financ., 4(6): 20-27.

Tse RYC (1997). An application of the ARIMA model to real estate prices in Hong Kong. J. Property Financ., 8(2): 52-163.

Vapnik VN (2000). The nature of statistical learning theory. New York: Springer.

Wang CH, Hsu LC (2010). The influence of dynamic capability on performance in the high technology industry: The moderating roles of governance and competitive posture. Afr. J. Bus. Manage., 4(5): 562-577.

Wen KL (2004). Grey Systems: modeling and prediction. Yang's Scientific Research Institute, AZ. USA.

Ziarko W (1993). Variable precision rough set model. J. Comput. Syst. Sci., 46: 39-59.

Ziarko W (2001). Probabilistic decision tables in the variable precision rough set model. Comput. Intel., 17(3): 593-603.