

Application of enhanced cluster validity index function to automatic stock portfolio selection system

Kuang Yu Huang · Shiuan Wan

Published online: 23 June 2011
© Springer Science+Business Media, LLC 2011

Abstract This paper presents an automatic stock portfolio selection system. In the proposed approach, 53 financial indices are collected for each stock item and are consolidated into six financial ratios [Grey relational grades (GRGs)] using a Grey relational analysis model. The GRGs are processed using a modified form of the PBMF index method (designated as the Huang index function) to determine the optimal number of clusters per GRG. The resulting cluster indices are then processed using rough set theory to identify the stocks within the lower approximate sets. Finally, the GRGs of each stock item in the lower approximate sets are consolidated into a single GRG, indicating the ability of the stock item to maximize the rate of return. It is demonstrated that the proposed stock selection mechanism yields a higher rate of return than several existing portfolio selection systems.

Keywords Rough set · Huang index function · PBMF index function · Classification · Cluster · Portfolio

1 Introduction

Predicting the future behavior of the stock market is of great interest to investors, speculators, and industries alike. However, it is extremely difficult to predict future stock prices and construct profitable stock portfolios in today's volatile markets. Many researchers have attempted to

predict stock price movements on a daily, weekly or monthly basis using artificial intelligence (AI) techniques [1–3] or statistical methods [4, 5]. When the problem of stock market prediction is considered, the following three fundamental factors inevitably arise:

- (1) The trading cost increases as the frequency at which the stocks are traded increases. This has a significant impact on the stock prediction problem on a daily basis.
- (2) When the future behavior of the stock market is forecasted, large volumes of external information are required, e.g., probability data (for statistical techniques); membership grade information (for fuzzy set theory techniques [6–8]); available training data (for artificial neural networks (ANNs) [9]), and so on.
- (3) When regression models are adopted to predict stock market trends, the results are not only determined by the financial indices of the stocks involved, but also by external factors such as the financial environment, political changes, changes in company strategy, fluctuations in the demand/supply relationship, and so on. Consequently, the predicted results cannot be guaranteed to be reliable.

In this study, it is argued that the effects of political, economic and other changes are only temporary. Thus, the goal of the study is to analyze the underlying value of each company's stock from an investor's perspective, irrespective of the short term impact of external effects. In practice, if the stock value can be predicted rationally and consistently for each company, it is expected that an evaluation process can be constructed as a basis for selecting stocks for investment purposes. In real-world stock market systems, the information associated with each data object may be imprecise and uncertain. Therefore, the task of

K. Y. Huang (✉) · S. Wan
Department of Information Management, Ling Tung University,
1 Ling Tung Road, Taichung City 408, Taiwan
e-mail: kyhuang@mail.ltu.edu.tw

S. Wan
e-mail: shiuan123@mail.ltu.edu.tw

identifying the relationships between the independent and dependent variables is highly challenging. In this study, the inherent ambiguity of the stock market data is dealt with by applying rough set (RS) theory to process the clustered data in order to identify suitable stocks for investment purposes. However, as described above, to simplify the clustering and stock filtering processes by applying GRA (Grey relational analysis) to datasets, it is first necessary to pre-process the data to identify and consolidate independent variables among the collected information.

Accordingly, the stock selection mechanism proposed in this study comprises two major components, namely (a) data processing and (b) data mining. In the data processing component, a GRA model is used to consolidate the 53 financial indices associated with each stock item into just 6 predetermined financial ratios known as Grey relational grades (GRGs). In the data mining component, the number of clusters per attribute is optimized using a modified form of the PBMF index function, and then RS theory is applied to identify the stocks within the lower approximate sets. These stocks are then processed by a GRA reduction model to establish a single financial indicator for each stock item on which to base the stock selection decision.

Real-world datasets invariably comprise a large number of objects and attributes. Thus, when applying RS theory to classify such datasets, it is desirable to pre-process the dataset to eliminate any conditional attributes which have little or no effect on the classification decision. This simplifies the decision table, and the decision rules can be more easily identified. Among all the available dimension reduction methods (e.g., principle component analysis, independent component analysis and GRA), GRA is particularly attractive since it effectively consolidates attributes.

Grey system theory, proposed by Deng [10] in 1982, is a powerful technique for dealing with systems characterized by poor, incomplete and uncertain information. One of the most fundamental components of Grey system theory is that of GRA [11], in which information from the Grey system is used to quantify the respective effects of the various factors within the system in terms of Grey relational grades (GRGs). In other words, GRA provides the means to *weigh* the various factors within an uncertain system in accordance with their effects on the system outcome, and therefore, it provides an ideal basis for classification systems. GRA requires a smaller volume of data and can process a large number of factors simultaneously, even when the relationship among these factors is uncertain or complex. As a result, GRA provides an ideal tool to analyze the complex inter-relationships amongst the individual parameters in systems with multiple performance characteristics [12–15], and has therefore been

widely applied in a variety of optimization, decision-making and classification problems in the fields of finance, business, economics, design, manufacturing and production [13–22]. In the stock selection mechanism proposed in this study, GRA is initially used to reduce the 53 financial performance indices of each stock item (collected factors) to just 6 core attributes (system parameters) in order to simplify the clustering process whilst retaining the underlying interrelationship between the conditional and decision attributes of the stock system. The GRA model is then reapplied at a later stage of the stock selection process to consolidate the 6 core attributes of each candidate stock item to a single performance indicator, which is then applied to select the stocks for inclusion in the stock portfolio. The data objects in the stock system are then clustered using a fuzzy clustering scheme, and the resulting cluster indices are analyzed using RS theory [16, 23] to identify the stocks within the lower approximate sets. In general, when applying RS theory to categorize real numbers into different classes, it is necessary to eliminate any redundant factors (attributes) and to determine an appropriate number of clusters for each attribute. As described above, the problem of removing the redundant factors is resolved in the proposed stock selection mechanism by using a GRA reduction model.

When any typical clustering problem is considered, two fundamental questions invariably arise: (1) “how many clusters are actually present within the dataset?” and (2) “to what extent do the clustering results reflect the true partitions within the dataset and enable the extraction of reliable decision-making rules? [24]” Traditionally, the problem of evaluating the optimality of the clustering results obtained for a particular dataset is referred to as the cluster validity problem [25]. Many methods have been proposed to assess the performance of fuzzy clustering schemes [26]. Among these methods, early indices, such as the partition coefficient [27, 28] and the classification entropy index [25, 29], are based simply on the membership values of the items within the dataset, and are therefore easily computed. However, recent studies have shown that the performance of clustering schemes can be improved by considering not only the values of the data objects within the dataset, but also the matrix U used to partition the data [16, 30–36]. Existing clustering methods typically cluster the dataset in accordance with the norms of the instances rather than the values of the individual attributes of the instances. However, in most real-world datasets, the instances within the dataset may have multiple attributes, whereby each attribute represents an independent parameter of the corresponding instance. Consequently, the clustering results obtained using traditional methods fail to take sufficient account of the complex interrelationships among the various attributes of the

dataset. Therefore, when devising a method to optimize both the clustering results and the classification accuracy, it is necessary to apply some form of classification-defined knowledge to the attribute values of the instances so that the complex interrelationships between the various attributes can be properly taken into account. In the stock selection mechanism proposed in this study, this is achieved using a new index function designated as the Huang cluster validity index function.

The remainder of this paper is organized as follows. Section 2 presents the fundamental principles of GRA theory, the FCM method, RS theory and the conventional PBMF index method. Section 3 describes the integration of these concepts to create the proposed Huang index function. Section 4 compares the performance of the Huang index function with that of the PBMF (FCM-based) clustering function when applied to a hypothetical dataset and a real-world stock market system. Section 5 describes the proposed stock portfolio selection system and evaluates its performance. Finally, Sect. 6 provides some brief concluding remarks.

2 Review of related methodologies

2.1 Grey relational analysis

Basically, a Grey relational analysis (GRA) function is an arithmetic mean, geometric mean or p-norm function applied to specified groupings of conditional attributes. GRA functions provide effective means of resolving multiple-criteria decision-making problems by ranking the potential solutions in terms of their so-called Grey relational grades (GRGs) so that the optimal solution can be easily determined [37]. In the stock portfolio selection system proposed in this study, the GRA method is used to simplify the stock classification and selection processes by consolidating the values of the multiple attributes of each data object into a limited number of attribute values, each representing one particular sub-system of the total stock system.

2.2 Fuzzy C-means (FCM) clustering [38, 39]

The fuzzy C-means (FCM) clustering method, developed by Dunn in 1973 [40] and later refined by Bezdek [27], has many applications, ranging from feature analysis to clustering and classifier design. The FCM clustering method consists of two basic procedures, namely (1) calculating the cluster centroids within the dataset, and (2) determining the cluster memberships of each data object. This two-step procedure is repeated iteratively until the centroids of all the clusters within the dataset converge.

2.3 Index function I_{\max}

It is assumed that each object x_i in the dataset has just one conditional attribute, and that this attribute can be partitioned into p groups (i.e., p clusters). As a result, each data object has a total of p membership functions $\mu_j(x_i) \quad j = 1, 2, \dots, p$. In the Huang index function, the data objects are mapped to the p clusters in accordance with the following index function:

$$I_{\max}(\mu_j(x_i)) = \text{Index}(\max(\mu_j(x_i))) = C(x_i) = c, \quad 1 \leq c \leq p.$$

For example, it is supposed that the conditional attribute is partitioned into 3 clusters and the membership functions of the first object x_1 in each of these 3 clusters are given by $\mu_1(x_1) = 0.35$, $\mu_2(x_1) = 0.63$, and $\mu_3(x_1) = 0.02$, respectively. In this particular example, the index function returns a value of $C(x_1) = I_{\max}(\mu_j(x_1)) = 2$, and thus, the conditional attribute of the first object is mapped to the second cluster. This approach is easily extended to the case of data objects with multiple attributes. For example, if every object 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

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

where $I_{\max}(\mu_j(x_i(a_l)))$ returns the index of the cluster corresponding to the maximum value among all the membership functions of the l -th attribute of x_i .

2.4 Rough set theory

Rough set (RS) theory was introduced by Pawlak [35] as a means of handling the vagueness and uncertainty inherent in the real-world decision-making process. RS theory is based on the assumption that every object in the universe of discourse is associated with a particular set of information (i.e., attributes). Objects characterized by the same information are regarded as being indiscernible. The indiscernibility relationship among all the objects in the universe of discourse provides the basic mathematical basis for RS theory.

2.4.1 Approximate sets

In RS theory, this indiscernibility of the data objects is handled using approximate sets. It is assumed that the information system $S = (U, A, V_q, f_q)$ is represented in the form of a decision table in which $X \subseteq U$ and $R \subseteq A$.

The upper and lower approximate sets of X are denoted as $\bar{R}(X)$ and $\underline{R}(X)$, respectively, and are defined as

$$\begin{aligned}\bar{R}(X) &= \{x \in U \mid [x]_p \cap X \neq \emptyset\}, \\ \underline{R}(X) &= \{x \in U \mid [x]_p \subseteq X\},\end{aligned}$$

where $[x]_p$ denotes the equivalence class determined by x with respect to P , i.e., $[x]_p = \{y \in U : (x, y) \in I_p\}$. The lower approximate set $\underline{R}(X)$ contains all of the elements (x) which have the same rank when evaluated in terms of the X -th decision attribute, while the upper approximate set $\bar{R}(X)$ contains the set of all elements (x) which may have the same rank when processed in accordance with the X -th decision attribute.

Having determined the upper and lower approximate sets, the accuracy of the classification results can be evaluated in accordance with:

$$\alpha_c = \frac{|\underline{R}(X)|}{|\bar{R}(X)|},$$

where $X = \{x : C_d(x) = c, \forall x \in U\}$, and $|\underline{R}(X)|$ and $|\bar{R}(X)|$ are the cardinalities of the lower and upper approximate sets, respectively, when the elements (x) are ranked in terms of the c -th cluster of the decision attribute d .

2.5 PBMF cluster validity index function

The PBMF cluster validity index function [24] ensures 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 $\text{PBMF}(K) = \left(\frac{1}{K} \times \frac{\bar{E}_1}{J_{m'}} \times D_K\right)$, where K is the number of clusters, $J_{m'} = \sum_{k=1}^K \sum_{j=1}^n \mu_{kj}^{m'} \|x_j - z_k\|$, \bar{E}_1 is constant for a given dataset and is set in such a way as to prevent the second term from vanishing, and $D_K = \max_{i,j=1}^K \|z_i - z_j\|$. 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 centroids of the k -th cluster. When the PBMF index function is applied to data clustering applications, the objective is to find the value of K which maximizes the index value.

3 Huang index function

The performance of RS theory in categorizing real numbers into different classes is critically dependent on the number of clusters used in the clustering process. In other words, an inappropriate choice as to the number of clusters may lead to a significant degradation in the classification performance. Accordingly, in the Huang index function proposed in this study, the FCM clustering scheme, RS theory, and a

modified form of the PBMF index function are integrated to optimize both the number of clusters within the dataset and the corresponding classification accuracy. Assuming that U is the domain of discourse and R is the set of equivalences of U , the RS classification problem can be formulated as follows:

$$X \subseteq U \quad \text{is:} \quad (\underline{R}_P(X), \bar{R}_P(X), \text{BND}_P(X)),$$

where X is the set of elements; U/I_P is the quotient set of U ; I_P is the indiscernibility of R ; \emptyset is the zero set; R is the attribute set of X and comprises the conditional attribute set (C) and the decision attribute set (D), $P \subseteq C$; $\underline{R}_P(X)$ is the lower approximate set of X ; $\bar{R}_P(X)$ is the upper approximate set of X ; and $\text{BND}_P(X)$ is the boundary set of X . It should be noted that every element in the domain of discourse U ($X \subseteq U$) has an attribute set (R) which describes the particular value of X .

As discussed in Sect. 2.3, the Huang index function is applied to cluster the attribute values of the data objects within the dataset, rather than the norms. Thus, in contrast to the conventional PBMF index function, the proposed approach takes better account of the intrinsic interrelationships among various parameters of the information system. In the Huang index function, each attribute (both conditional and decision) is assumed to have an equal number of clusters, and the objective is to map each attribute of element (X_i) in U to an appropriate cluster among all the clusters associated with the conditional ($C_1 \sim C_n$) or decision (d) attributes. The detailed parameters of the Huang index function are presented in the following section.

3.1 Parameters of the Huang index function

The Huang index function proposed in this study has the following form:

$$H(C, \alpha_c) = \left(\frac{1}{C} \times \frac{\bar{E}_1}{F'_C} \times D'_C\right),$$

where C is the number of clusters assigned to the conditional and decision attributes and α_c is the corresponding classification accuracy when evaluated in terms of the c -th cluster of the decision attribute d . In addition, F'_C is obtained by accumulating the value of E'_c for each cluster of the decision attribute (d), where E'_c is given by $E'_c = \sum_{j=1}^n \bar{\mu}_{cj}^{m'}(x_j(d)) \|x_j - z'_c\| / \alpha_c$, in which $\bar{\mu}_{cj}(x_j(d))$ is the membership function of data object x_j in the c -th cluster of the decision attribute d , and z'_c is the multi-dimensional centroids of the lower approximate sets associated with the 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 data item within the corresponding sets. Furthermore, m' is the fuzzification parameter and n is

the total number of data objects in the dataset. Finally, the value of D'_C is equal to the maximum separation distance among the centroids of all the lower approximate sets associated with the different clusters of the decision attribute, i.e., $D'_C = \max_{i,j=1}^C \|z'_i - z'_j\|$. It should be noted that the value of D'_C is upper bounded by the maximum separation distance among all possible pairs of data points within the dataset.

Parameter F'_C in the Huang index function differs from the term $J_{m'}$ in the PBMF index function (see Sect. 2.5) in that its value depends on E'_c and therefore, takes classification accuracy into account.

3.2 Tendencies of terms within the Huang index function

As discussed in the previous section, the Huang index function has the form $H(C, \alpha_c) = \left(\frac{1}{C} \times \frac{\bar{E}_1}{F'_C} \times D'_C\right)$. In other words, the index function comprises three terms: $1/C$, \bar{E}_1/F'_C and D'_C . Clearly, the value of the first term decreases as the number of clusters assigned to the conditional and decision attributes, C , increases. In other words, the value of the index function falls as the value of C rises. In the second term of the index function, the value of \bar{E}_1 is constant for a given dataset and is equal to \bar{E}_1 in the PBMF index function. As discussed in the previous section,

F'_C represents the sum of all E'_c ; each of which includes the classification accuracy α_c when evaluated in terms of the c -th cluster of the decision attribute d . Hence, the Huang index function increases as E'_c decreases. The third term in the index function, D'_C , measures the maximum separation distance between the centroid of the lower approximate sets associated with the different clusters of the decision attribute, and increases as C increases. Thus, the contribution of D'_C to the value of the Huang index function increases as the number of decision attribute clusters increases.

3.3 Comparison between the Huang index function and the PBMF index function

Table 1 compares the major components of the Huang index function and the PBMF index function. At a high level, three major differences exist between the two functions, namely (1) the Huang index function clusters the individual attributes of each data object within the dataset, whereas the PBMF index method clusters the data based on the norms of the data objects; (2) the Huang index function is based on z'_c , i.e., the centroid of the lower approximate sets associated with 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 (3) the Huang index

Table 1 Detailed definitions of Huang and PBMF indices

Functions	Huang index $H(C, \alpha_c) = \left(\frac{1}{C} \times \frac{\bar{E}_1}{F'_C} \times D'_C\right)$	PBMF index $PBMF(K) = \left(\frac{1}{K} \times \frac{\bar{E}_1}{J_{m'}} \times D_K\right)$
How to cluster the data?	Cluster all attributes of data objects C is the number of clusters assigned to the conditional and decision attributes	Cluster all data in a dataset K is the number of clusters of a dataset
	$F'_C = \sum_{c=1}^C E'_c, \quad E'_c = \sum_{j=1}^n \bar{\mu}_{cj}^{m'}(x_j(d)) \ x_j - z'_c\ / \alpha_c$	$J_{m'} = \sum_{k=1}^K E_k, \quad E_k = \sum_{j=1}^n \mu_{kj}^{m'} \ x_j - z_k\ $
	(1) $\bar{\mu}_{cj}(x_j(d))$ is the membership function of data object x_j in the c -th cluster of the decision attribute d .	(1) $\mu_{kj}^{m'}$ is the membership function of the j -th data object.
	(2) z'_c is the multi-dimensional centroid of the lower approximate sets associated with the 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 data item within the corresponding sets	(2) z_k is the centroid of the k -th cluster obtained when clustering the dataset using the FCM method
	(3.1) $\ x_j - z'_c\ $ is the length of the vector (norm) between the x_j data object and z'_c	(3.1) $\ x_j - z_k\ $ is the length of the vector (norm) between the x_j data object and z_k
	(3.2) $E'_c = \sum_{j=1}^n \ x_{jc}\ $, where $\ x_{jc}\ = \bar{\mu}_{cj}^{m'}(x_j(d)) \ x_j - z'_c\ $	(3.2) $E_k = \sum_{j=1}^n \ x_{jk}\ $, where $\ x_{jk}\ = \mu_{kj}^{m'} \ x_j - z_k\ $.
	(3.3) α_c is the classification accuracy and indicates the cardinality proportion of lower approximates in upper approximates when evaluated in terms of the c -th cluster of decision-making attribute d	
	$D'_C = \max_{i,j=1}^C \ z'_i - z'_j\ $ is equal to the maximum separation distance among the centroids of all the lower approximate sets associated with the different clusters of the decision attribute	$D_K = \max_{i,j=1}^K \ z_i - z_j\ $ is equal to the maximum separation distance between the cluster centroids

function takes explicit account of the classification accuracy when evaluating the optimality of the clustering results, whereas the PBMF index function considers only the optimal number of clusters within the dataset.

3.4 Details of the Huang index function

Figure 1 illustrates the basic structure of the Huang index function [41] and summarizes each processing step.

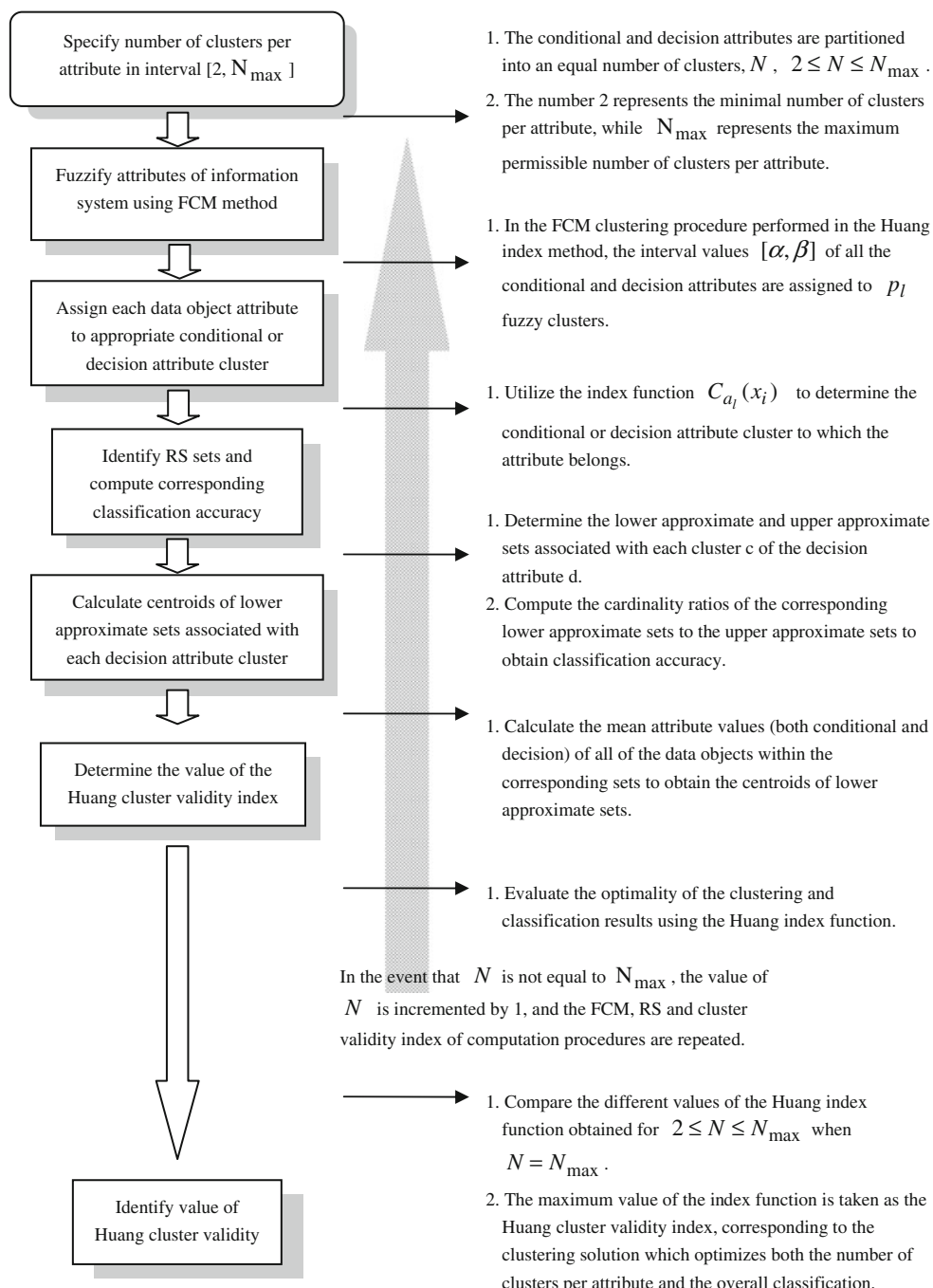


Fig. 1 Basic steps of the Huang index function

4 Performance evaluation of the Huang index method

This section commences by presenting a step-by-step example showing the calculation of the Huang index. The validity of the proposed Huang index method is then evaluated by considering an illustrative example related to electronic stock data extracted from the financial database maintained by the Taiwan Economic Journal (TEJ) [16]. When evaluations are performed, the effectiveness of the

Table 2 Ten instances in a hypothetical dataset

Code of instances	Conditional attributes		Decision attribute
	a_1	a_2	
1	1.0044	0.9896	1.9941
2	1.6411	0.6098	2.2509
3	1.6998	0.7001	2.3999
4	1.8968	0.3775	2.2743
5	1.8789	0.2440	2.1229
6	1.2290	0.5401	1.7690
7	1.5648	0.2555	1.8202
8	1.0157	0.2267	1.2423
9	1.1612	0.7161	1.8773
10	1.5732	0.4154	1.9886

proposed index method is measured by comparing the partitioning and classification results with those obtained from the conventional PBMF index (FCM-based) method.

4.1 Step-by-step example showing the calculation of the Huang index value

In the following example, steps 2–6 of the Huang index method are demonstrated for a hypothetical dataset in which each instance has two conditional attributes, i.e., a_1 , a_2 , and one decision attribute, i.e., d . The instances in the hypothetical dataset are shown in Table 2.

4.1.1 Step 2: fuzzify attribute values of instances using the FCM method

The continuous data in the hypothetical dataset are clustered using the FCM method. An assumption is made that each conditional attribute can be partitioned into 2 clusters. The membership function values of each attribute of each

Table 3 Membership function values of each attribute of each instance

Code of instances	Conditional attributes				Decision attribute	
	a_1	a_2	a_1	a_2	d	d
1	0.9811	0.0189	0.8700	0.1300	0.6071	0.3929
2	0.0166	0.9834	0.8656	0.1344	0.9923	0.0077
3	0.0004	0.9996	0.9951	0.0049	0.9217	0.0783
4	0.0518	0.9482	0.0345	0.9655	0.9840	0.0160
5	0.0446	0.9554	0.0190	0.9810	0.9577	0.0423
6	0.9357	0.0643	0.5984	0.4016	0.0073	0.9927
7	0.0912	0.9088	0.0138	0.9862	0.0503	0.9497
8	0.9846	0.0154	0.0278	0.9722	0.2053	0.7947
9	0.9888	0.0112	0.9992	0.0008	0.1647	0.8353
10	0.0791	0.9209	0.0999	0.9001	0.5846	0.4154

Table 4 Centroids of each cluster of attributes

i -th Cluster	Conditional attributes		Decision attribute
	a_1	a_2	
1	1.6705	0.6550	2.3254
2	1.0157	0.2267	1.2423

instance are summarized in Table 3. The centroids of each attribute cluster are shown in Table 4.

4.1.2 Step 3: assign each data object attribute to appropriate conditional or decision attribute cluster

The appropriate conditional and decision attribute clusters are obtained for each instance by applying the index function I_{max} to the membership function values shown in Table 3. The corresponding results are presented in Table 5.

4.1.3 Step 4: identify RS sets and compute the corresponding classification accuracy

The upper and lower approximate sets for each cluster c of the decision attribute d are shown in Table 6. The classification accuracy associated with each cluster of the decision attribute is obtained by computing the cardinality ratio of the corresponding lower approximate sets to the upper approximate sets. The results are shown in Table 7.

4.1.4 Step 5: calculate the centroids of the lower approximate sets associated with each cluster of the decision attribute

The multi-dimensional centroids of the lower approximate sets associated with each cluster of the decision attribute d

Table 5 Discretization results of data in Table 2

Code of instances	Conditional attributes		Decision attribute
	a_1	a_2	
1	1	1	1
2	2	1	1
3	2	1	1
4	2	2	1
5	2	2	1
6	1	1	2
7	2	2	2
8	1	2	2
9	1	1	2
10	2	2	1

are obtained by calculating the mean attribute values (both conditional and decision) of all of the instances within the corresponding sets. Thus, the centroids of the lower approximate sets associated with the two clusters of decision attribute d are obtained as follows:

$$z'_1 = \text{mean}(x|x \in \underline{R}(X), C_d(x) = 1) = \text{mean}(x|x \in \{x_2, x_3\})$$

$$= ((1.6411 + 1.6998), (0.6098 + 0.7001), (2.2509 + 2.3999)) = (1.6705, 0.6550, 2.3254), \text{ and}$$

$$z'_2 = \text{mean}(x|x \in \underline{R}(X), C_d(x) = 2) = \text{mean}(x|x \in \{x_8\})$$

$$= (1.0157, 0.2267, 1.2423).$$

4.1.5 Step 6: determine the value of the cluster validity index

Having determined the classification accuracy and centroids of the lower approximate sets, the optimality of the clustering and classification results is evaluated using the Huang index function (i.e., $H(C, \alpha_c) = (\frac{1}{c} \times \frac{E'_1}{F'_c} \times D'_C)$).

The membership functions of the first instance x_1 in the two clusters associated with the decision attribute d are given by $\bar{\mu}_{11}(x_1(d)) = 0.6071$, $\bar{\mu}_{21}(x_1(d)) = 0.3929$, respectively (see Table 3). It should be noted that the first instance x_1 has attribute values of $x_1(1.0044, 0.9896, 1.9941)$ (see Table 2 and the centroid of the lower approximate sets associated with the second cluster of the decision attribute is given by $z'_2(1.0157, 0.2267, 1.2423)$ (see Table 4). As a result, $(x_1(a_1) - z'_2(a_1)) = (1.0044 - 1.0157) = -0.0113$, $(x_1(a_2) - z'_2(a_2)) = (0.9896 - 0.2267) = 0.7629$, $(x_1(a_3) - z'_2(a_3)) = (1.9941 - 1.2423) = 0.7518$. Therefore, the vector of $x_{12} = x_1 - z'_2$ has the form $[x_{12}(a_1), x_{12}(a_2), x_{12}(a_3)] = [-0.0113, 0.7629, 0.7518]$, and the corresponding norm is equal to $\|x_1 - z'_2\| = \sqrt{x_{12}(a_1)^2 + x_{12}(a_2)^2 + x_{12}(a_3)^2} = \sqrt{(-0.0113)^2 + 0.7629^2 + 0.7518^2} = 1.0711$. Let the

Table 6 Lower and upper approximate sets

Lower approximate sets $\underline{R}(X : C_d(x) = c, x \in X)$				
2	2	1	1	$\underline{R}(X : C_d(x) = 1, x \in X)$
3	2	1	1	
8	1	2	2	$\underline{R}(X : C_d(x) = 2, x \in X)$
Upper approximate sets $\bar{R}(X : C_d(x) = c, x \in X)$				
1	1	1	1	$\bar{R}(X : C_d(x) = 1, x \in X)$
6	1	1	2	
9	1	1	2	
2	2	1	1	
3	2	1	1	
4	2	2	1	
5	2	2	1	
7	2	2	2	
10	2	2	1	
1	1	1	1	$\bar{R}(X : C_d(x) = 2, x \in X)$
6	1	1	2	
9	1	1	2	
4	2	2	1	
5	2	2	1	
7	2	2	2	
10	2	2	1	
8	1	2	2	

Table 7 Classification accuracy associated with each cluster of decision attribute $\alpha_c = |\underline{R}(X)|/|\bar{R}(X)|$ where $X = \{x : C_d(x) = c, \forall x \in U\}$

α_1	α_2
$2/9 = 0.2222$	$1/8 = 0.1250$

fuzzification parameter m' be specified as 2.0. The effect of instance x_1 on z'_2 , $\|x_{12}\|$, is obtained by multiplying $\|x_1 - z'_2\|$ by the square of the corresponding membership function, i.e., $\bar{\mu}_{21}^2(x_1(d)) = 0.3929^2 = 0.1544$. Thus, the value of $\|x_{12}\|$ is 0.1653. The effect of instance x_j on z'_i is shown in Table 8.

The value of E'_2 in the Huang index function is computed using the vectors $\|x_{j2}\| = \bar{\mu}_{2j}^2(x_j(d)) \times \|x_j - z'_2\|$ and the classification accuracy α_2 presented in Tables 7 and 8, respectively. Specifically, E'_2 is determined by summing up the products of the norms and the squares of the corresponding membership functions for each of the instances, and then dividing the result by α_2 . Thus, the value of E'_2 is obtained as

$$E'_2 = \left(\sum_{j=1}^{10} \bar{\mu}_{2j}^2(x_j(d)) \|x_j - z'_2\| \right) / \alpha_2$$

$$= \left(\sum_{j=1}^{10} \|x_{j2}\| \right) / \alpha_2 = (\|x_{12}\| + \|x_{22}\| + \dots + \|x_{102}\|) / \alpha_2$$

$$= (0.1653 + 0.0001 + \dots + 0.1640) / 0.1250 = 18.1462.$$

Table 8 Values of $\|x_j - z'_c\| (= \mu_{c_j}^2(x_j(d)) \times \|x_j - z'_c\|)$

x_j	z'_c	
	$c = 1$	$c = 2$
1	0.3007	0.1653
2	0.0905	0.0001
3	0.0781	0.0087
4	0.3503	0.0003
5	0.4616	0.0022
6	0.0000	0.6395
7	0.0016	0.7195
8	0.0563	0.0000
9	0.0185	0.5685
10	0.1451	0.1640

Similarly, the value of E'_1 is obtained as 6.7620. The value of F'_C is then found to be $F'_C = \sum_{c=1}^C E'_c = 24.9082$.

The factor \bar{E}_1 in the Huang index function is a constant term for a dataset in which the instances belong to only one cluster. Thus, the centroid of the illustrative dataset is given by $z_1 = \text{mean}(x|x \in \{x_i\}, i = 1, 2, \dots, 10)$. As a result, the centroid z_1 calculated by the arithmetic mean function $\text{mean}(x|x \in \{x_i\}, i = 1, 2, \dots, 10)$ has attribute values of $\text{mean}(x|x \in \{x_i\}, i = 1, 2, \dots, 10) = ((1.0044 + 1.6411 + 1.5732), (0.9896 + 0.6098 + \dots + 0.4154), (1.9941 + 2.2509 + \dots + 1.9886)) = z_1(1.4665, 0.5075, 1.9740)$. Based on the vector of centroid z_1 , it can be shown that $(x_1(a_1) - z_1(a_1)) = (1.0044 - 1.4665) = -0.4621$, $(x_1(a_2) - z_1(a_2)) = (0.9896 - 0.5075) = 0.4821$, and $(x_1(a_3) - z_1(a_3)) = (1.9941 - 1.9740) = 0.0201$. Therefore, the vector of $x_{12} = x_1 - z'_2$ has the form $[x_{11}(a_1), x_{11}(a_2), x_{11}(a_3)] = [-0.4621, 0.4821, 0.0201]$, and the corresponding norm is equal to $\|x_1 - z_1\| = \sqrt{x_{11}(a_1)^2 + x_{11}(a_2)^2 + x_{11}(a_3)^2} = \sqrt{(-0.4621)^2 + 0.4821^2}$

$\sqrt{+0.0201^2} = 0.6681$. Similarly, the norms of $\|x_2 - z_1\|, \|x_3 - z_1\|, \dots, \|x_{10} - z_1\|$, are found to be 0.3431, 0.5224, ..., 0.1418, respectively. The value of \bar{E}_1 is then obtained by summing up the norms of $\|x_j - z_1\|$ where $j = 1, 2, \dots, 10$, yielding a value of $\bar{E}_1 = 4.6403$.

The value of D'_C in the Huang index function is acquired by calculating the maximum separation distance between the centroids of the lower approximate sets associated with the first and second clusters of the decision attribute. As shown in Table 4, these centroids are given by $z'_1(1.6705, 0.6550, 2.3254)$ and $z'_2(1.0157, 0.2267, 1.2423)$, respectively. Thus, the vector of $z_{12} = z'_1 - z'_2$ has the form $[z_{12}(a_1), z_{12}(a_2), z_{12}(a_3)] = [0.5611, -0.1543, 0.4069]$, and the corresponding norm is determined to be $\|z'_1 - z'_2\| = \sqrt{z_{12}(a_1)^2 + z_{12}(a_2)^2 + z_{12}(a_3)^2} = \sqrt{0.5611^2 + (-0.1543)^2} = \sqrt{+0.4069^2} = 1.3361$.

Finally, the Huang index $((H(C, \alpha_c)) = (\frac{1}{C} \times \frac{E'_1}{F'_C} \times F'_C))$ is found to have a value of 0.1245, where $C = 2, \bar{E}_1 = 4.6403, F'_C = 24.9082$ and $D'_C = 1.3361$.

4.2 An illustrative example

In this section, the performance of the Huang index function in partitioning and classifying real-world complex datasets is evaluated using stock data extracted from the TEJ database for the first quarter of 2008. The TEJ database comprises 53 financial indices (attributes) for each stock item (data object). However, for reasons of practicality, the performance evaluations were restricted to just 1 decision attribute and 2 conditional attributes. Having deleted records in which some of the data was incomplete, a total of 327 records were obtained. (See Table 9 for representative values of each index for a selected subset of these 327 records.)

Table 9 Illustrative financial data extracted from TEJ database for first quarter 2008

Company codes	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)	(m)
1	-2.17	-3.78	-1.61	-0.01	-0.04	-0.07	96.85	252.7	6.8	342.7	0	97.51	0.16
2	-1.58	0.34	1.92	0.01	-0.38	0.08	90.42	56.26	38.85	107.8	-0.04	37.07	0.73
3	-1.58	82.01	80.67	2.42	0.38	23.33	60.36	91.89	38.72	113.5	0.04	141.5	-0.07
4	37.16	31.04	-6.12	2.6	31.07	25.69	85.92	84.46	57.44	112.9	3.11	55.07	9.16
...
324	8.17	9.37	1.2	1.35	11.68	13.4	102.0	26.22	44.16	84.02	1.21	53.21	6.18
325	7.38	7.37	-0.01	1.05	10.56	10.54	43.8	21.31	29.23	35.87	1.06	20.97	3.73
326	0.08	-0.97	-1.05	-0.07	0.06	-0.69	102.0	15.24	88.44	28.77	0.01	33.46	3.13
327	1.99	3.41	1.42	0.32	1.84	3.16	46.88	129.5	63.08	113.3	0.18	192.8	2.24

(a) Business profit rate, (b) Pre-tax income %, (c) Net non-op. Inc./rev., (d) PS-pre_tax income, (e) Operational income/capital, (f) Pre-tax income/capital, (g) Days-A/R turnover, (h) Days-inventory turnover, (i) Days-A/P turnover, (j) Net operating cycle, (k) PS-operating income, (l) D/E ratio, (m) EPS-net income

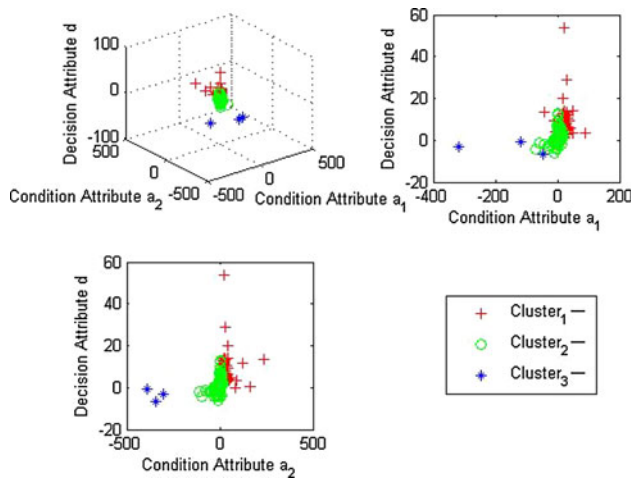


Fig. 2 Data partitioning results obtained by the PBMF index function when clustering stock market data based on two conditional attributes and one decision attribute

In this illustrative example, the performance of the Huang index function is compared with that of the PBMF index function for a case in which the clustering process is based on just two conditional attributes (i.e., business profit rate and pre-tax net profit rate) and the single decision attribute (i.e., EPS net income). In this example, the dataset is partitioned into three clusters in the PBMF index function and into three clusters per attribute in the Huang index function. The corresponding clustering results are presented in Figs. 2 and 3, respectively. In this example, Fig. 2 contains all of the data points within the dataset, whereas Fig. 3 includes only those data points belonging to the lower approximate sets associated with each cluster of the decision attribute. (Note: Fig. 3 only shows results for one cluster of the decision attribute, since the other two clusters were both found to have upper approximate sets

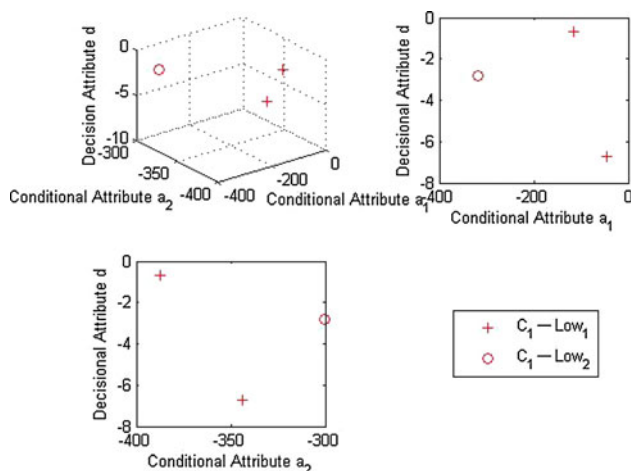


Fig. 3 Data partitioning results obtained by the Huang index function when clustering stock market data based on two conditional attributes and one decision attribute

only.) Figure 2 shows that the partitioning results obtained using the PBMF index function are dominated by the value of the second conditional attribute. In other words, most of the data points within the dataset are assigned to a single cluster. In addition, it can be seen that some of the data points in the first cluster overlap those in the second. Simply put, the PBMF index function yields a poor partitioning and classification performance when applied to real-world information systems characterized by a limited number of attributes. Furthermore, it can be seen in Fig. 3 that the lower approximate sets generated by the Huang index function contain very few data points, yielding very little information for generating reliable decision-making rules when it is applied to complex information systems with a limited number of attributes per cluster.

5 Evaluation of proposed portfolio selection model

In this section, the Huang index function is combined with a GRA dimension reduction model and an RS classification scheme to obtain an automatic stock portfolio selection system. In the proposed system, a specified set of stock items is collected automatically every quarter, and the 53 financial indices associated with each stock item are consolidated into 6 normalized financial ratios using a GRA model. The stock items are then clustered in accordance with their financial ratio values using the Huang index function, and the cluster indices associated with the optimal clustering solution (i.e., the clustering solution which maximizes the value of the Huang cluster validity index) are then processed using an RS classification model in order to identify the stocks within the lower approximate sets. These stock items are then filtered in accordance with Buffet's general investment principles [42] in order to determine stocks for possible inclusion within the portfolio. Finally, the GRA model is re-applied to consolidate the 6 normalized financial ratios of the filtered stocks into a single GRG, indicating the potential of each stock item to maximize the rate of return on the stock portfolio.

The major concepts of the proposed system are described in Sects. 5.1–5.3. The detailed processing steps within the system are then discussed in Sect. 5.4. Finally, the performance of the proposed system is evaluated in Sect. 5.5.

5.1 GRA dimension reduction mechanism

The GRA model is used to compute the following financial ratios: (1) profitability, (2) rate per share, (3) growth rate, (4) credit capacity, (5) operating capacity, and (6) statutory ratio, where ratio (1) is taken as the decision attribute of the stock system and ratios (2)–(5) are taken as the conditional

attributes. (Note: the mapping of the 53 financial indices to the six consolidated ratios is summarized in Table 10).

The stock market 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 information system with the form $S = (U, \hat{A}, V_q, f_q)$ is obtained, where \hat{A} is a set of six consolidated attributes (financial ratios) describing the same set of objects. The six financial ratios are clustered using the Huang index function, and cluster indices corresponding to the optimal clustering solution are then processed using RS theory to identify the corresponding lower approximate sets. Assuming that U is the domain of discourse and R is the set of equivalences of U , the RS problem can be formulated as $X \subseteq U : (\bar{R}(X), \underline{R}(X)), BN_R(X)$ (see Sect. 3).

As described above, the GRA model is also used to reduce the six financial ratios of each remaining stock item after the stocks within the lower approximate sets have been filtered using the general investment principles prescribed by Buffet. In this case, the GRA model takes the six consolidated financial ratios (GRGs) of each stock item as inputs. It then outputs a single GRG describing the overall performance of the corresponding stock item. The GRGs are ranked in descending order so 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 for the final stock selection decision.

5.2 Filtering of stock items in accordance with basic investment principles

To simplify the workload of the GRA model in reducing the six GRGs of each stock item to a single performance indicator, the stocks within the lower approximate sets are filtered in accordance with a set of decision-making attributes, which are defined in accordance with the general investment principles specified by Buffett and formalized by Hagstrom [42]. Buffett argued that reducing costs is essential for enterprises seeking to hone their competitive ability and rival their competitors in terms of price, while high profit margins and a high inventory turnover are both reliable indicators of the financial well-being of a company. Buffet further asserted that only companies with all three attributes can be certain of survival and possess the means to earn profit for their shareholders. Accordingly, in the present study, the stock items within the lower approximate sets identified by the RS classification model are filtered in accordance with the following thresholds: (1) return on asset (after tax) > 0, (2) return on equity > 0, (3) gross profit ratio > 0, (4) equity growth rate > 0, and (5) constant EPS > 0.

Table 10 Mapping of 53 financial indices–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	Percentage of gross margin growth	Acid test	Total asset turnover	Operation Inc./employee
Gross margin %	PS-cash flow	YOY %-real. GM	Interest Exp. %	A/R & N/R turnover	Fixed assets/employee
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-pre_tax 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			

5.3 Data extraction

In this study, the feasibility of the proposed stock selection mechanism was evaluated using electronic stock data extracted from the TEJ database over the period extending from the first quarter of 2003–6/1/2009. In general, financial statements for a particular accounting period are subject to a certain delay before publication. For example, annual reports are published after 4 months, half-yearly reports after 2 months, and first and third quarterly reports (without notarization) after a minimum of 1 month. The submission deadlines for the financial statements maintained in the TEJ database are as follows:

(1) Annual report: the submission deadline laid down by the Security Superintendence Commission is 4 months after the closing balance day. However, companies listed in previous years (TSE and OTC) can delay filing until 5/31.

(2) Half-yearly report: the submission deadline laid down by the Security Superintendence Commission is 2 months after the closing balance day. However, companies listed in previous years (TSE and OTC) can delay filing until 9/21.

(3) First-quarter report: the submission deadline laid down by the Security Superintendence Commission is 1 month after the closing balance day. However, companies listed in previous years (TSE and OTC) can delay filing until 5/31.

(4) Third-quarter report: the submission deadline laid down by the Security Superintendence Commission is 1 month after the closing balance day. However, companies listed in previous years (TSE and OTC) can delay filing until 11/15.

Since financial data relating to the last quarter of every year is not available until May 31st of the following year, the data cannot be used by the proposed stock selection system to select suitable investment stocks in the first quarter. As a result, the stock selection system can only be executed three times in every 12 months period, namely 5/31–09/22, 9/22–11/15 and 11/15–05/31 of the following year.

5.4 Detailed processing steps in the Huang index function-based stock selection system

The detailed processing steps in the proposed stock selection system are illustrated in Fig. 4, and summarized below.

5.4.1 Step 1: data collection and attribute determination

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 to modify the choice of financial ratios used for attribute reduction in the initial GRA process, to select a new GRA model for attribute reduction purposes,

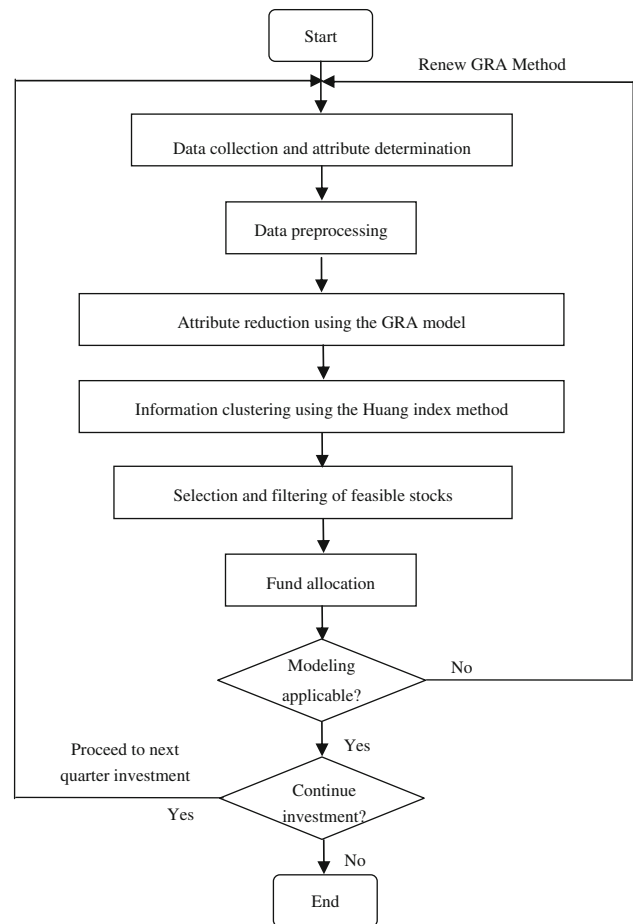


Fig. 4 Proposed stock selection system flow chart

and to modify the decision-making attributes used to filter the stocks in the lower approximate set prior to their further consolidation using the GRA model.

5.4.2 Step 2: data pre-processing

Having collected the relevant financial data for each quarterly period, a basic pre-processing operation is performed to improve the efficiency of the GRA attribute reduction process. Specifically, the data records containing missing fields (i.e., missing financial indices) are deleted, and the box plots method [43] is applied to resolve the data outlier problem by establishing an inter-quartile range so that any data points falling outside this range are automatically assigned a default value depending on the interval within which they fall.

5.4.3 Step 3: information consolidation using the GRA model

When stock records remain after the pre-processing operation completes, the GRA model normalizes the values of

each of the 53 financial indices and then computes the six corresponding financial ratios in accordance with the mapping given in Table 10.

5.4.4 Step 4: information clustering using the Huang index method

The values of the six financial ratios obtained in Step 3 (i.e., five conditional attributes $C_1 \sim C_5$ and one decision attribute D_1) are processed using the Huang index method in order to identify the optimal number of clusters per attribute (conditional and decision) and the corresponding set of cluster indices.

5.4.5 Step 5: selection and filtering of feasible stocks

The optimal set of cluster indices generated by the Huang index method is processed using RS theory to identify the stock items within the lower approximate sets. These stock items are then filtered in accordance with the general investment guidelines proposed by Buffett in order to identify a set of stocks for possible inclusion within an investment portfolio.

5.4.6 Step 6: fund allocation

The six financial ratios of each stock item remaining after the filtering operation are consolidated to a single GRG (i.e., an overall performance indicator) by the GRA model. The GRGs of all the surviving stock items are then arranged in descending order, and the first five stock items are chosen for investment purposes.

5.4.7 Step 7: check the validity of the modeling

The rate of return on the stock portfolio constructed 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. If the rate of return is acceptable, a decision is made as to whether or not the model should be run for a further quarter using the existing GRA model. However, if the rate of return is deemed unacceptable, the suitability of the GRA model is reviewed and a new GRA model is adopted if appropriate.

5.5 Performance evaluation of the Huang index function-based stock selection system

In [16], an automatic stock market forecasting and portfolio selection mechanism is constructed by integrating a moving average autoregressive exogenous (ARX) prediction

model with a GM (1, N) attribute reduction model and RS theory. Meanwhile, in [44], financial data was collected automatically and input to a GM (1, 1) prediction model in order to forecast the future trends of the collected data. The forecast data was then reduced using a GM (1, N) model, classified using a k-means clustering algorithm, and processed by an RS classification module to select suitable investment stocks. Finally, a minimum variance at risk (MVAR) scheme comprising 5 GRA models and a Markowitz mean–variance (MV) model was employed to identify an efficient frontier, which achieved the optimal tradeoff between the minimum risk and the maximum rate of return. The MVAR–MRR model was used to evaluate the expected risk at the point of the maximum expected return rate for the five frontier curves generated by the five GRA models. The stock allocation with the lowest risk value was then selected as being the optimal solution.

In this section, the validity and effectiveness of the proposed stock selection mechanism is initially evaluated by comparing the rate of return on the investment portfolios selected in the 17 investment periods between 2003 and 2009 with the rate of return on the equivalent investment portfolios constructed using a system in which the Huang index function is replaced by a fuzzy-based clustering scheme, with the number of clusters per GRG specified simply as $N = 3$. The rate of return obtained using the two stock selection schemes is then compared with: (1) the predicted average rate of return implied by the variation in the TAIEX index over the equivalent investment periods; (2) the average rate of return achieved by the portfolio selection method proposed in [16]; and (3) the average rate of return achieved by the MVAR–MRR portfolio selection method proposed in [44].

The corresponding results are presented in Table 11. It can be seen that the accumulated rate of return achieved using the proposed Huang index function-based mechanism (116.90%) is higher than that achieved using the pre-determined clustering based scheme (103.21%). The accumulated rate of return obtained through the proposed method is also higher than that implied by the variation in the TAIEX index (44.90%). Additionally, in the period 2004–2006, the accumulated rate of return achieved using the Huang index function-based mechanism (107.93%) was higher than that achieved using the GM(1,N) attribute reduction based scheme [16] (82.45%) or the MVAR–MRR method [44] (86.22%). Meanwhile, the rates of return achieved in 2004–2006 using the Huang index function-based mechanism are 33.36, 27.27 and 47.30, respectively. In contrast, the rates of return achieved using the GM (1, N) attribute reduction-based scheme are 17.57, 25.90 and 38.98, respectively, while those achieved using the MVAR–MRR method-based scheme are 19.59, 20.84 and 45.79, respectively. In other words, the rates of return

Table 11 Rates of return of TAIEX, GM (1, N)-based reduction attributes method, MVAR-MRR method, pre-determined cluster-based stock selection scheme, and the Huang-based stock selection scheme

Investment period	TAIEX	GM(1, N)-based reduction attributes method in Ref. [16]*	MVAR -MRR method in Ref. [44] ^a	Cluster pre-determined	Huang index function-based
03/09/22–03/11/17	4.87			11.06	1.91
03/11/17–04/05/31	0.43			−6.27	−0.69
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	44.90	82.45	86.22	103.21	116.90

* This column gives the results from a paper published by the current authors in Ref. [16]

^a This column gives the results from the paper published by the current authors in Ref. [44]

achieved using the proposed stock selection scheme are higher than those obtained using the GM (1, N) attribute reduction scheme or the MVAR-MRR method-based scheme. Thus, the overall viability and effectiveness of the proposed stock selection system is confirmed.

6 Conclusions

This study has presented an automatic stock portfolio selection system based on a Grey relational analysis (GRA) model, a modified form of the PBMF index method (designated as the Huang index method), and rough set (RS) theory. In the proposed approach, 53 financial indices were collected automatically for each stock item every quarter and a GRA model was used to consolidate these indices into six predetermined financial ratios [Grey relational grades (GRGs)]. The GRGs were then processed using the Huang index function in order to determine the optimal number of clusters per GRG and the corresponding values of the cluster indices for each stock item. The cluster indices were then processed using an RS classification model in order to identify the stock items within the lower approximate sets of the stock system. These items were filtered in accordance with established investment principles and the six GRGs of

each surviving stock item were then consolidated into a single GRG, indicating the performance of the corresponding stock item in terms of its ability to maximize the likely rate of return. Finally, the top five stock items were chosen for investment purposes. The general validity of the Huang index function has been confirmed by comparing the clustering results obtained for a real-world database containing stock market information with the corresponding results obtained using the conventional PBMF index (FCM-based) method. Finally, the real-world feasibility of the proposed stock portfolio selection system has been demonstrated by comparing the rate of return on the selected portfolio with that obtained by three alternative stock selection schemes. The results presented in this study support the following major conclusions:

(1) In the PBMF index function, the optimality of the clustering results is evaluated in accordance with (a) the distance between each data object and the cluster centroids, and (b) the maximum separation distance between the cluster centroids. In contrast, in the Huang index function, the optimality of the clustering results is evaluated in terms of (a) the distance between the data objects and the centroids of the lower approximate sets associated with each cluster of the decision attribute, (b) the maximum distance between the centroids of the lower approximate sets associated with the different clusters of the decision attribute,

and (c) the classification accuracy of the clustering results (i.e., the cardinality ratio of the lower approximate sets to the upper approximate sets for each cluster of the decision attribute). In other words, in contrast to the conventional PBMF index function, which simply determines the optimal clustering of the data within the dataset given a specified number of clusters, the Huang index function determines the optimal number of attribute clusters within the dataset, which maximizes the separation distance between the attribute values and simultaneously optimizes the classification accuracy.

(2) The PBMF index function clusters the norms of the data objects within the dataset, whereas the Huang index function clusters the attribute values. As a result, the Huang index method takes better account of the intrinsic interrelationships among the various conditional and decision attributes, and prevents the clustering results from being dominated by any attribute(s) with a higher order of magnitude.

(3) In the case of a more complex dataset, e.g., a dataset in which the attributes have relatively homogeneous values, the PBMF index function clusters all of the data points within the dataset, but yields a relatively poor partitioning performance. In contrast, the Huang index function achieves a better partitioning performance, but only classifies a limited subset of the data points. In other words, the Huang index function is only able to extract a limited amount of useful and correct information from the dataset, and is therefore of limited use in defining accurate and reliable decision-making rules.

(4) The stock portfolio selection system based on the Huang index function yields a higher rate of return than a system in which the clustering process is based on a pre-determined number of clusters per attribute. In addition, the rate of return on the selected stock portfolio is considerably higher than that predicted by the overall variation in the TAIEX index over the equivalent investment period. Finally, the rate of return on the selected stock portfolio is superior to that obtained by the GM(1,N) attribute reduction-based scheme in [16] or the MVAR-MRR based scheme in [44].

In fact, the goal of this study is to achieve a possible tendency through our stock portfolio selection system. Incidentally, since the prices in the stock market sometimes fluctuate dramatically, we chose not to measure stock prices over the short term, instead investigating their performance over the long term. In summary, the results presented in this study demonstrate that the Huang index function is an effective tool for optimizing both the number of attribute clusters and the classification accuracy when applied to the partitioning and classification of complex, real-world datasets. As a result, the Huang index function provides an ideal basis for such applications as automatic

portfolio selection mechanisms (demonstrated in this study), landslide detection, daily electrical peak load forecasting, and so on.

References

- Hassan MR, Nath B, Kirley M (2007) A fusion model of HMM, ANN and GA for stock market forecasting. *Expert Syst Appl* 33:171–180
- Cao Q, Leggio KB, Schniederjans MJ (2005) A comparison between fama and French's model and artificial neural networks in predicting the Chinese stock market. *Comput Oper Res* 32:2499–2512
- Kuo RJ, Chen CH, Hwang YC (2001) An intelligent stock trading decision support system through integration of genetic algorithm based fuzzy neural network and artificial neural network. *Fuzzy Sets Sys* 118:21–45
- Tse RYC (1997) An application of the ARIMA model to real estate prices in Hong Kong. *J Prop Finance* 8(2):52–163
- Pankratz A (1983) Forecasting with univariate box-jenkins models: concepts and cases. Wiley, NY
- Deng T, Chen Y, Xu W, Dai Q (2007) A novel approach to fuzzy rough sets based on a fuzzy covering. *Inf Sci* 177:2308–2326
- Liu M, Chen D, Cheng W, Li H (2006) Reduction method based on a new fuzzy rough set in fuzzy information system and its applications to scheduling problems. *Comput Math Appl* 51:1571–1584
- Qin K, Pei Z (2005) On the topological properties of fuzzy rough sets. *Fuzzy Sets Syst* 151:601–613
- Romahi Y, Shen Q (2000) Dynamic financial forecasting with automatically induced fuzzy associations. In: *Proceedings of the 9th international conference on fuzzy systems*. San Antonio, pp 493–498
- 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
- Wang Z, Zhu L, Wu JH (1996) Grey relational analysis of correlation of errors in measurement. *J Grey Syst* 8:73–78
- Zhu F, Yi M, Ma L, Du J (1996) The grey relational analysis of the dielectric constant and others. *J Grey Syst* 8:287–290
- Tan X, Yang Y, Deng J (1998) Grey relational analysis factors in hypertensive with cardiac insufficiency. *J Grey Syst* 10:75–80
- Xu G, Tian W, Qian L, Zhang X (2007) A novel conflict reassignment method based on grey relational analysis (GRA). *Pattern Recognit Lett* 28:2080–2087
- Huang KY, Jane C-J (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
- Wu JH, Chen CB (1999) An alternative form for grey relational grades. *J Grey Syst* 11:7–12
- Kung C-Y, Wen K-L (2007) Applying grey relational analysis and grey decision-making to evaluate the relationship between company attributes and its financial performance: a case study of venture capital enterprises in Taiwan. *Decis Support Syst* 43: 842–852
- Chiang K-T, Chang F-P (2006) Application of grey-fuzzy logic on the optimal process design of an injection-molded part with a thin shell feature. *Int Commun Heat Mass Transf* 33:94–101
- Lina Y-C, Laib H-H, Yeh C-H (2007) Consumer-oriented product form design based on fuzzy logic: a case study of mobile phones. *Int J Ind Ergon* 37:531–543
- Wong CC, Lai HR (2000) A new grey relational measurement. *J Grey Syst* 12:341–346

22. Hu Y-C (2007) Grey relational analysis and radial basis function network for determining costs in learning sequences. *Appl Math Comput* 184:291–299
23. Huang KY (2009) Application of VPRS model with enhanced threshold parameter selection mechanism to automatic stock market forecasting and portfolio selection. *Expert Syst Appl* 36(9): 11652–11661
24. Pakhira MK, Bandyopadhyay S, Maulik U (2004) Validity index for crisp and fuzzy clusters. *Pattern Recognit* 37:487–501
25. Bezdek JC (1974) Cluster validity with fuzzy sets. *J. Cybern* 3: 58–74
26. Wang W, Zhang Y (2007) On fuzzy cluster validity indices. *Fuzzy Sets Syst* 158:2095–2117
27. Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Plenum Press, NY
28. Trauwaert E (1988) On the meaning of Dunn's partition coefficient for fuzzy clusters. *Fuzzy Sets Syst* 25:217–242
29. Bezdek JC (1974) Numerical taxonomy with fuzzy sets. *J Math Biol* 1:57–71
30. Gath I, Geva AB (1989) Unsupervised optimal fuzzy clustering. *IEEE Trans Pattern Anal Mach Intell* 11:773–781
31. Wu KL, Yang MS (2005) A cluster validity index for fuzzy clustering. *Pattern Recognit Lett* 26:1275–1291
32. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning internal representations by error propagation. In: Rumelhart DE, McClelland JL (eds) *Parallel distributed processing: explorations in the microstructure of cognition*. MIT Press, Cambridge, pp 318–363
33. Vapnik VN (2000) *The nature of statistical learning theory*. Springer, NY
34. Friedman N, Geiger D, Goldsmid M (1997) Bayesian network classifiers. *Mach Learn* 29(2):131–163
35. Pawlak Z (1982) Rough sets. *Int J Inf Comput Sci* 11(5):341–356
36. Pawlak Z (1994) Rough set approach to multi-attribute decision analysis. *Eur J Oper Res* 72:443–459
37. Huang KY, Jane C-J, Chang T-C (2008) A novel approach to enhance the classification performances of grey relation analysis. *J Inf Optim Sci* 29:1169–1191
38. Cox E (2005) *Fuzzy modeling and genetic algorithms for data mining and exploration*. Elsevier, USA
39. Glackina C, Maguirea L, McIvorb R, Humphreysb P, Hermana P (2007) A comparison of fuzzy strategies for corporate acquisition analysis. *Fuzzy Sets Syst* 158:2039–2056
40. Dunn JC (1973) A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J Cybern* 3:32–57
41. Huang KY (2010) 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(12): 8757–8769
42. Hagstrom RG, Miller B, Fisher KL (2005) *The Warren Buffett way: investment strategies of the world's greatest investor*. Wiley, USA
43. Chakravarti IM, Laha RG, Roy J (1967) *Handbook of methods of applied statistics*, vol 1. Wiley, USA
44. Huang KY (2009) A hybrid GRA/MV model for the automatic selection of investment portfolios with minimum risk and maximum return. *J Grey Syst* 21:149–166 (ISSN: 0957–3720)