A RS Model for Stock Market Forecasting and Portfolio Selection
Allied with Weight Clustering and Grey System Theories
Kuang Yu Huang, Chuen-Jiuan Jane, and Ting-Cheng Chang

Abstract—In this study, the weight clustering model which consists of GM(1,N) with K-means Clustering is combined with Grey Systems theory and Rough Set (RS) theory to create an automatic stock market forecasting and portfolio selection mechanism. In our proposed approach, financial data are collected every quarter and are inputted to an GM(1,1) predicting model to forecast the future trends of the collected data over the next quarter. Next, the forecasted data of financial statement is transformed into financial ratios using a GM(1,N) model and clustered by using a K-means clustering algorithm, and then supplied to a RS classified module which selects appropriate investment stocks by adopting a set of decision-making rules. Finally, a grey relational analysis technique is applied to specify an appropriate weighting of the selected stocks to maximize the portfolio’s rate of return. The validity of our proposed approach is demonstrated to use the electronic stock data extracted from the financial database maintained by the Taiwan Economic Journal (TEJ). The portfolio's results derived by using our proposed weight clustering model are compared with those portfolio’s results of a conventionally clustering method. It is found that our proposed method yielded a greater average annual rate of return (23.42%) on the selected stocks from 2004 to 2006 in Taiwan stock market.

I. INTRODUCTION

Predicting stock prices in today's volatile markets is notoriously difficult and represents a major challenge for traditional time-series-based forecasting mechanisms. A number of applications have been proposed in recent decades for predicting the market trends. Typical mechanisms include the use of genetic algorithms to choose optimal portfolios [1,2], the application of neural networks to predict real-world stock trends [3-5], the integration of fuzzy logic and forecasting techniques to create artificial intelligence systems for market tracking and forecasting purposes [6,7], the use of statistical approaches for the forecasting of economic indicators [8-12], the application of rough set (RS) theory to predict the S&P 100 index [13], and so on.

Manuscript received December 2, 2007. This work was supported in part by the National Science Council, Taiwan, Republic of China, under Grant NSC 94-2416-H-275-002..

Kuang Yu Huang is with the Department of Management Information System, Ling Tung University, Taiwan (phone number: 886-4-23892088 ext 3701; fax number: 886-4-24064101; e-mail: kyhuang@mail.ltu.edu.tw).

Chuen-Jiuan Jane is with the Department of Finance, Ling Tung University, Taiwan (e-mail: hj869@mail.ltu.edu.tw).

Ting Cheng Chang is with the Department of Commercial Technology & Management, Ling Tung University, Taiwan (e-mail: tchang@mail.ltu.edu.tw).

Rough set theory was introduced more than twenty years ago [14] and has emerged as a powerful technique for the automatic classification of objects [15]. The popularity of RS theory stems primarily from its operational processes, which adhere closely to the notions of knowledge discovery and data mining [16]. Furthermore, RS theory acts directly upon the data of interest and has no need for any other form of external information, such as probability data (as in statistical techniques), grade of membership information (as in fuzzy set theory [17-19]), over-fitting data (as in artificial neural networks (ANNs) [20]), and so on.

This study develops a mechanism for assisting investors in forecasting the future behavior of the stock market such that they can make the rational decisions as to how to best manage their portfolio. Our proposed approach combines the GM(1,1) grey prediction model, the multivariate GM(1,N) model [21], the K-means clustering technique, RS theory, grey relational analysis, and the investment guidelines prescribed by Buffett [22] to develop an algorithm for predicting financial data over a quarter and for forecasting the stock portfolio to maximize the rate of return. The stock data and market indices used for evaluation purposes are extracted from the financial database maintained by the Taiwan Economic Journal (TEJ). The performance of our proposed algorithm in clustering the future financial ratios and choosing an appropriate stock portfolio is compared with the performance of a conventional k-means based model.

The organization of the rest of the paper is as follows. Section 2 deals with the fundamental principles of the Grey system theory and RS theory, respectively. Section 3 demonstrates the integration of these concepts to construct an automatic forecasting and portfolio selection scheme. Section 4 compares the performance of the proposed method with the performance of a conventional k-means clustering model. Finally, Section 5 presents some briefly concluding remarks and indicates the intended direction of future research.

II. REVIEW OF RELATED METHODOLOGIES

A GM(1,1) Model

The output of the GM(1,1) model is predicted on the basis of an assumption that a system changes only gradually over time and can be constructed by less data. The grey predicting model has three operations: (a) accumulated generation, (b) inverse accumulated generation, and (c) grey modeling. This model uses the operation of accumulated generation to build a differential equation as
\[
ax^{(i)}(k) + \frac{b}{a}x^{(i)}(k) = bh^i, \quad \text{where } a \text{ and } b \text{ are the characteristic coefficients, } \quad x^{(i)}(k) = \sum_{m=1}^{k} x^{(i)}(m), \quad k = 1, 2, ..., n \] is the accumulated generation operation of \( x^{(i)} \). \( x^{(0)} \) = \{x^{(0)}(1), x^{(0)}(2), x^{(0)}(3), ..., x^{(0)}(n)\} is the original data sequence, and \( n \) is the number of quarters. The approximation coefficients are determined by using the least square method and symbolized as \( X_n = \left[ x^{(0)}(2), x^{(0)}(3), x^{(0)}(4), ..., x^{(0)}(n) \right]^T \) is part of \( x^{(0)} \).

Applying the inverse AGO, we then have
\[
\hat{x}^{(0)}(k) = \left( x^{(0)}(0) - \frac{b}{a} \right) \left( 1 - e^{-\frac{a}{b}} \right) e^{\frac{a}{b}(k-1)} \quad (1)
\]

where \( k = 2, 3, ..., n \) : \( \hat{x}^{(0)}(1), \hat{x}^{(0)}(2), ..., \hat{x}^{(0)}(n) \) and \( \hat{x}^{(0)}(n+1), \hat{x}^{(0)}(n+2), ... \) are called the GM(1,1) fitted sequence and the GM(1,1) forecasting values, respectively.

### Rough Set Theory

Rough set theory (RST) was introduced by Pawlak [14] in 1982. RST is a powerful mathematical tool for handling the vagueness and uncertainty inherent in many decision-making processes. The underlying principle of RST is the assumption that every object in the universe of discourse has a set of information (i.e. attributes) associated with it. Objects characterized by the same information are regarded as the indiscernibility. The indiscernible relationships generated amongst all the objects in the universe of discourse provide the basic mathematical basis for RST. The typical problems amenable to RST processing include the classifying sets of objects in terms of their attribute values, checking dependencies (full or partial) between attributes, reducing attributes, analyzing the significance of individual attributes, generating decision rules, and so on [27].

#### 1) Information systems

In RST, knowledge about the universe of discourse is represented using so-called information systems. A typical information system has the form \( S = (U, \Omega, \alpha_q) \), where \( U \) is a non-empty set of finite objects and \( \Omega \) is a non-empty finite set of attributes describing each object. Here, \( \Omega = C \cup D \), in which \( C \) is a finite set of conditional attributes and \( D \) is a finite set of decision-making attributes. For each \( q \in \Omega \), \( \alpha_q \) represents the domain of \( q \). Finally, \( \alpha_q \) is the information function and is given by \( f : U \rightarrow V_q \). The elements \( (X \subseteq U) \) in the information system represent individual cases, states, processes, patients or observations; for example, while the attributes \( (C \& D) \) can be regarded as the features, variables or characteristic conditions of these elements. The decision-making table (also known as an attribute-value table) is a particular RS information system in which the rows and columns represent elements in the universe of discourse and the attributes of these elements, respectively.

#### 2) Approximation of Sets

In RST, this indiscernibility of the elements is handled for using the concept of approximate sets. Assume that \( S = (U, \Omega, \alpha_q, f_q) \) is a decision table in which \( X \subseteq U \) and \( R \subseteq \Omega \). The upper and lower approximates of \( X \) are denoted as \( R^+(X) \) and \( R^-(X) \), respectively, and are defined as
\[
R^+(X) = \bigcup \{Y \in U / IND(R) : Y \cap X \neq \emptyset \} \quad (2)
\]
\[
R^-(X) = \bigcup \{Y \in U / IND(R) : Y \subseteq X \} \quad (3)
\]
where \( U / IND(R) \) expresses the equivalence of \( R \) and \( IND(R) \) denotes the indiscernibility of \( R \), i.e.
\[
IND(R) = \{(x, y) \in U^2 : \text{for every } a \in R, a(x) = a(y)\} \quad (4)
\]

The lower approximate set \( R_-(X) \) contains all elements \( X \) of the same rank when evaluated in terms of the \( Y \) decision-making attribute, while the upper approximate set \( R^+(X) \) contains the set of all possible same-rank elements \( X \) when processed in accordance with the \( Y \) decision-making attribute. Finally, the set \( BN_r(X) = R^+(X) - R^-(X) \) is referred to as the boundary set of \( X \).

#### C. GM(1,N) model

Imagine a system described by the sequences \( x^{(0)}(i), i = 1, 2, 3, ..., n \), where \( x^{(0)}(i) \) describes the main factor of interest and sequences \( x^{(0)}(k), x^{(0)}(k), ..., x^{(0)}(k) \) are the factors which influence this main factor. Such a system can be analyzed by using the following multivariate GM(1,N) grey model:
\[
x_i(k) + ax_i^{(i)}(k) = \sum_{j=1}^{k} b_jx_j^{(i)}(k) \quad (5)
\]

in which \( k = 2, 3, ..., n \) and \( x_i^{(0)}(k) = \sum_{j=1}^{k} x_j^{(0)}(i) \) and \( x_i^{(i)}(k) = 0.5x_i^{(i)}(k) + 0.5x_i^{(i)}(k-1), k \geq 2 \).

Substituting all possible \( x_i^{(i)}(k) \) terms into Equation (5) yields a matrix of the form as follows:
D. Grey relational analysis

According to Deng [21], the grey relational coefficient has the form

\[
\gamma(x_i(0), x_j(0)) = \frac{\Delta_0 + \xi \Delta_{\text{max}}}{\Delta_{\text{min}} + \xi \Delta_{\text{max}}} \tag{8}
\]

where \(i = 1, 2, \ldots, m\), \(k = 1, 2, \ldots, n\), \(x_i(k)\) is the reference value, and \(x_j(k)\) is the inspected value.

Furthermore,
\[
\Delta_{\text{min}} = \min_{\forall i, \forall k} \norm{x_0(k) - x_i(k)}
\]
\[
\Delta_{\text{max}} = \max_{\forall i, \forall k} \norm{x_0(k) - x_i(k)}
\]

Finally, \(\xi\) is the distinguishing coefficient with a value between 0 and 1. Note that \(\xi\) is generally assigned to a value of 0.5.

The grey relational grade represents the degree of correlation between two sequences and is generally defined as the average of their respective grey relational coefficients, i.e.

\[
\Gamma_{ij} = \frac{1}{n} \sum_{k=1}^{n} \gamma(x_0(k), x_i(k)) \tag{9}
\]

Having calculated the grey relational grades, the sequences can be ranked by using a so-called grey relational ranking procedure. For example, for the case of a reference sequence \(x_0(k)\), the grey relational rank of \(x_i(k)\) is greater than that of \(x_j(k)\) if \(\gamma(x_0, x_i) > \gamma(x_0, x_j)\). The corresponding ranking is denoted as \(x_i > x_j\).
C. Detailed processing steps of forecasting and stock selection model

The detailed processing steps in the forecasting and stock selection model are illustrated in Figure 1. The basic steps in this model can be summarized as follows:

*Step 1: Data collection and attribute determination*

In our current model, the conditional attributes should reflect the financial quality of a company. Therefore, our proposed model specifies the following attributes: the profitability, the capitalized cost ratio, the individual share ratio, the growth rate, the debt ratio, the operational leverage and all statutory financial ratios. The decision as to which of the selected companies should actually be processed by the RST portfolio selection mechanism is then made by processing the forecast data generated by the GM(1,1) model in accordance with the seven decision-making attributes specified in the previous section.

*Step 2: Data preprocessing*

Having collected the relevant financial data every quarter, a basic pre-processing operation is performed to improve the efficiency of the GM(1,1) forecasting model. For example, any data records containing missing fields (i.e. attributes) are immediately rejected. In addition, the problem of data outliers is addressed by using the Box Plots method [23] to establish an inter-quartile range such that any data falling outside this range can be automatically assigned to a default value depending on the interval within which it is located.

*Step 3: GM(1,1) prediction*

Our current model uses an GM(1,1) prediction model to forecast the future trends of the financial variables of each of the selected companies. In the current GM(1,1) model, forecasting is deliberately restricted to a one-step-ahead and a rolling mode to prevent the accumulation of errors from previous forecasting periods.

*Step 4: Information reduction using GM(1,N) multivariate model*

To improve the efficiency of the RST / Grey relational analysis process, the values of financial data can be transformed into the values of financial ratios in which the numbers of conditional attributes can be reduced to seven. In our proposed model, this transforming operation is performed by using the GM(1,N) method to identify the weighting of every financial data contributed to specified financial ratios.

*Step 5: K-means clustering*

Prior to submission to the RST stock selection mechanism, the forecast data of conditional attribute values \((C_1 \sim C_n)\) are clustered into three groups using a K-means clustering algorithm.

*Step 6: Selection of approximate sets*

Having clustered the forecast data, the Rough Set method is adopted to determine the lower approximate set. The generalized rules extracted by the low approximate set are all recognized rules or relationships in the investment industry. This indicates that Rough Set analysis is a feasible means of identifying top stock performers by classifying the contributions of the attributes. In other words, Rough Set theory avoids the
need for the blind and haphazard stock selection methods commonly which are employed by investors in the past.

**Step 7: Fund allocation**

Having identified suitable stocks for the investment, it is then necessary to determine an appropriate weight of the stocks in the portfolio in order to maximize the overall rate of return. In our current model, this fund allocation problem is processed by using a Grey Relational Analysis based on

\[
\text{stock weight}(i) = \frac{n - i + 1}{\sum_{k=1}^{n} k},
\]

where \(i\) is the grey relation order of each stock item, and \(n\) is the total number of invested stocks.

Having completed all the steps described above, a check is made of the overall rate of return on the investment. 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 iterated process. However, if the rate of return is unacceptable, the suitability of the decision attributes is reviewed and amended appropriately.

**IV. THE EVALUATION OF OUR PROPOSED MODEL**

**BY USING ELECTRONIC STOCK DATA**

**A. Data extraction**

The feasibility of our proposed forecasting and stock selection model was evaluated using electronic stock data extracted from the New Taiwan Economy database (TEJ). The period for data collection was from the first quarter in 2003 to the fourth quarter in 2006, given a total of 16 quarters in all. Meanwhile, the forecasting period extended from the first quarter in 2003 to the first quarter in 2007, given a total of 17 quarters.

In general, financial statements for a particular accounting period are subject to a considerable delay before they are actually published. For example, the annual reports are published after four months, the half-yearly reports are published after two months, and the first and third quarterly reports (without notarization) are published after at least one month.

Since the last quarter data every year can not be acquired until 31st May in the following year, the data can not be used by the GM(1,1) model to predict the financial trends over the first quarter of the year. In other words, the forecasting and investing process proposed in this study can only be conducted three times each year, i.e. 5/31–09/22, 9/22–11/15 and 11/15–05/31 next year. In addition, in the decision-making rules used in the Rough Set stock selection process, the Return on Equity (ROE) and constant EPS indicators are based on the full 12 months of the previous year. Thus, the forecasting period for investment purposes is actually reduced to the second quarter in 2004 to the fourth quarter in 2006.

**B. Verification of weight-base clustering performance**

Table 1 compares the stock companies selected by our proposed hybrid model with those chosen by an alternative model in which the weight clustering module was replaced by the conventional clustering of information reduction model. It is observed that these two stock selection models recommend the much different companies for investment purposes.

<table>
<thead>
<tr>
<th>Listed Company Code</th>
<th>Call (5/31)</th>
<th>Number</th>
<th>Put (9/22)</th>
<th>return rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2391</td>
<td>41.63</td>
<td>8</td>
<td>51.62</td>
<td>9.99%</td>
</tr>
<tr>
<td>6131</td>
<td>38.19</td>
<td>6</td>
<td>35.21</td>
<td>-2.98%</td>
</tr>
<tr>
<td>2441</td>
<td>22.11</td>
<td>9</td>
<td>19.68</td>
<td>-2.43%</td>
</tr>
<tr>
<td>3035</td>
<td>38.13</td>
<td>3</td>
<td>33.67</td>
<td>-4.46%</td>
</tr>
<tr>
<td>6257</td>
<td>22.21</td>
<td>3</td>
<td>17.24</td>
<td>-4.97%</td>
</tr>
<tr>
<td>Total return rate</td>
<td></td>
<td></td>
<td></td>
<td>1.26%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>listed company code</td>
<td>Call (5/31)</td>
<td>Number</td>
<td>Put (9/22)</td>
<td>return rate</td>
</tr>
<tr>
<td>3035</td>
<td>38.13</td>
<td>8</td>
<td>33.67</td>
<td>-4.46%</td>
</tr>
<tr>
<td>6131</td>
<td>38.19</td>
<td>6</td>
<td>35.21</td>
<td>-2.98%</td>
</tr>
<tr>
<td>2460</td>
<td>16.79</td>
<td>11</td>
<td>16.69</td>
<td>-0.10%</td>
</tr>
<tr>
<td>2391</td>
<td>41.63</td>
<td>3</td>
<td>51.62</td>
<td>9.99%</td>
</tr>
<tr>
<td>6141</td>
<td>27.91</td>
<td>2</td>
<td>21.86</td>
<td>-6.05%</td>
</tr>
<tr>
<td>Total return rate</td>
<td></td>
<td></td>
<td></td>
<td>-4.09%</td>
</tr>
</tbody>
</table>

Table 1 also shows that the rate of return obtained while investing in the companies which are recommended by our proposed model is better than the rate of return obtained while following the recommendations of the information reduction-based method.

Table 2 summarizes the quarterly and yearly rates of return obtained over the nine investment periods between 2004 and 2006 by using our proposed model and the information reduction-based model, respectively. As shown, our proposed model achieves a significantly higher average yearly rate of return than the information reduction-based model.
This study has presented a hybrid model based upon the Grey System theory and Rough Set theory for the automatic prediction of stock market trends and the selection of an optimal stock portfolio. The performance of the weight-based approach has been compared with that of an information reduction-based method using electronic stock data extracted from the New Taiwan Economy database (TEJ). The major findings and contributions of this study can be summarized as follows:

1. This study has confirmed the effectiveness of combining different clustering techniques to improve the efficiency and accuracy of automatic classification algorithms.

2. Since it is not necessary to concern about the relativity of the subsets, it is suitable for both certain and uncertain data system and it also owns both of data mining and information warning effects.

3. The process of information reduction usually obtained from the Rough Sets model is based on the principle of eliminating any redundant or unimportant information in the original data. In our current model, the reduction process does not conduct the process of elimination, but will transform the financial data into financial ratios as the conditional attributes. This transformation and reduction of conditional attributes performed prior to the RS classification process is conducted by using the GM(1,N) method.

4. It has been shown that the portfolio recommendations of the weight-based model are different from those portfolio recommendations of a conventional information reduction-based model, and our proposed model could yield a significantly improved rate of return.

Overall, the results presented in this study have confirmed that our proposed fusion model provides a promising method for stock portfolio management. Furthermore, the structure of our proposed model represents a suitable foundation for a broad range of derivatives in other research fields. In this current model, the forecasting task is performed by using the GM(1,1) predicting method proposed by Deng [37]. However, in future studies, the use of alternative predicting techniques will be considered in order to evaluate the potentiality for further improving the rate of return obtained from the selected stocks.

REFERENCES


<table>
<thead>
<tr>
<th>Investment period</th>
<th>Attributes Clustering Model</th>
<th>Weight-based model</th>
<th>Information reduction-based model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate of return</td>
<td>Quarter</td>
<td>Yearly</td>
<td>Quarter</td>
</tr>
<tr>
<td>2004 2nd quarter</td>
<td>1.26%</td>
<td>-4.09%</td>
<td>19.84%</td>
</tr>
<tr>
<td>3rd quarter</td>
<td>0.57%</td>
<td>15.61%</td>
<td>1.93%</td>
</tr>
<tr>
<td>4th quarter</td>
<td>13.78%</td>
<td>22.01%</td>
<td>-</td>
</tr>
<tr>
<td>2005 3rd quarter</td>
<td>-2.74%</td>
<td>21.51%</td>
<td>-0.62%</td>
</tr>
<tr>
<td>4th quarter</td>
<td>23.50%</td>
<td>17.32%</td>
<td>-2.52%</td>
</tr>
<tr>
<td>2006 3rd quarter</td>
<td>-1.64%</td>
<td>17.47%</td>
<td>-2.40%</td>
</tr>
<tr>
<td>4th quarter</td>
<td>30.98%</td>
<td>22.40%</td>
<td>33.15%</td>
</tr>
<tr>
<td>Average year rate of return</td>
<td>23.42%</td>
<td>18.89%</td>
<td></td>
</tr>
<tr>
<td>Accumulated 3 years rate of return</td>
<td>70.27%</td>
<td>56.67%</td>
<td></td>
</tr>
</tbody>
</table>


