International Journal of u- and e- Service, Science and Technology
Vol.9, No. 6 (2016), pp.153-168
http://dx.doi.org/10.14257/ijunesst.2016.9.6.15

A Hybrid Stock Selection Model Based on Forecasting, Classification and Feature Selection

Shiliang Zhang[1,2] and Tingcheng Chang[1]

[1]*Department of Computer science, Ningde Normal University*
[2]*Institute of Remote Sensing and Geographical Information Systems and Beijing Key Lab of Spatial Information Integration and Its Applications, Peking University*
[1]*shiliangzh@163.com*, [2]*tcchang@teamail.ltu.edu.tw*

Abstract

The basic aim of this paper is to provide a model to explain stock performance paramount level. To reach this purpose, this research proposes that rough set theory (RS), allied with the use of Grey Prediction, Semi-Supervised Graph Regularized Non-negative Matrix Factorization (SGNMF), K-means and Grey Relation, can out-perform the more standard approaches that are employed in economics. This study focuses on stock to select the optimal stock portfolio out applying the financial statement datum from the New Taiwan Economy database (TEJ). Firstly, this study collects relative financial ratio datum as the conditional attributes selection and then uses GM(1,1) for forecasting, SGNMF for choosing the more important conditional attributes, and rough set for figuring the best portfolio out. Finally, the Grey relational analysis is used to reduce the investment risk for fund allocation. This study will demonstrate that rough sets model is applicable to stock portfolio. The empirical result in Taiwan: During five years (2009-2013), the average annual rate of return was 20.41%, the accumulated rate of return for 9 quarter was 61.22%. The portfolio determined by the model is a promising alternative to the conventional methods for economic and financial prediction..

Keywords*: Grey Forecasting, SGNMF, K-means, Rough-Set, Grey Relational Analysis, Stock Portfolio*

1. Introduction

Developed economies make increasing use of data-intensive technologies. There are 4.6 billion mobile-phone subscriptions worldwide and there are between 1 billion and 2 billion people accessing the internet. Between 1990 and 2005, more than 1 billion people worldwide entered the middle class which means more and more people who gain money will become more literate which in turn leads to information growth. The world's effective capacity to exchange information through telecommunication net works was 281 petabytes in 1986, 471 petabytes in 1993, 2.2 exabytes in 2000, 65 exabytes in 2007 and it is predicted that the amount of traffic flowing over the internet will reach 667 exabytes annually by 2014 [1-2].From the financial sector to manufacturing operations, more and more companies are relying on the analysis of huge amount of data to compete. A data mining method has the goal to finds new patterns in the system. Following this goal, many techniques have been applied to a variety of domains in business data mining, including marketing, finance, banking, manufacturing and telecommunications. There are more and more applications appearing in the area of economic prediction, such as using genetic algorithms to choose optimal portfolio [2], selecting the real-world stocks using neural networks [3] and predicting the S&P 100 index using rough sets [4]. These new techniques become alternatives to conventional statistical methods and even perform

ISSN: 2005-4246 IJUNESST

better in some cases[5]. Among these methods, the Rough Sets Theory of Pawlak(1982) - as a knowledge discovery tool with many advantages-inspired many scholars to do research this theory to the business domain. So far they have achieved many promising results.

The rough sets model has the following advantages [6-9]:

1. It is based on the original data only and does not need any external information, unlike probability in statistics or grade of membership in the fuzzy set theory.

2. The rough sets model is a tool suitable for analyzing not only quantitative attributes but also qualitative ones.

3. It discovers important facts hidden in data and expresses them in the natural language of decision rules.

4. The set of decision rules derived by the rough sets model gives a generalized description of the knowledge contained in the financial information tables, eliminating any redundancy typical of the original data.

5. The decision rules obtained from the rough sets model are based on facts, because each decision rule is supported by a set of real examples.

6. The results of the rough sets model are easy to understand, while the results from other methods(credit scoring, utility function, outranking relation) need an interpretation of the technical parameters, with which the user may not be familiar.

Most rough sets applications are focused on classification problem. In those applications, the data sets applied to rough sets are temporal independent, which means the data set can be shuffled without affecting the final results. However, for time series, such as the medical history of the patients or historical data of a stock, the sequential information is important. How can rough sets be adapted to extract information from them? To better improve the time series capability, the grey prediction model is incorporated with the rough set.

The application of Rough Set Theory often was combined with other theory. The combined models derived from it, like LERS [10-11], LEM2 [12], RSES [13-15].The grey model was first proposed by Deng in 1982 [16-17] and [28]. The Variable Precision Rough Sets Model [18]; Rough Classifier [19]; ProbRough [20-22]; TRANCE [23-24];VCR [25-26], is all combined with other theory. A grey system is a partially known and partially unknown system. This system provides a method which uses only a few data just used to predict the next output from the plant. Without resorting to forming a knowledge base, the grey modeling scheme constructs a differential equation to characterize the system. Therefore the next output from the system can be obtained by solving the differential equation.

NMF (Nonnegative Matrix Factorization, NMF) by Lee and Seung [27] is published in "Nature". The biggest difference between the matrix factorization method and the traditional decomposition will be in its original matrix and its result matrices are non-negative. The nature of the non-negative matrix factorization (NMF) is very suitable for handling real-life non-negative data compared to other methods and it is easier to understand. In addition, NMF has to break down a lot of information from which it obtained and the results are interpreted. It accounts for less storage space and without parameters it can still provide good classification performance, so the subject received much academic attention. There are many scholars that follow-up with its improvement [28-30]. Isometric Mapping's (Isomap) non-cash nature of dimensionality reduction methods consisting of Tenenbaum *et al.* [31] was published in "Science". Isomap can be regarded as the combination of multi-dimensional scaling (MDS) and analysis of the linear distance (CDA), mainly based on the foundations of MDS to find the nearest neighboring points for each point, and geodesic Dijkstras algorithm to find the shortest path to establish the set point between the geodesic distance. Laplace Eigenmaps (LE) is also a non-linear dimension reduction method that is mainly used to construct data graph

the relationship between the figure adjacent points in the low dimensional space so the neighboring points can be as close as possible.

Many previous NMF revisions did not explicitly take the potential data space geometry into account, and this happens to be some clustering and classification issues that needs some special attention, so Cai *et al.* [32] presented in Fig formal Non-negative matrix factorization (Graph Regularized Non-negative Matrix Factorization, GNMF) to resolve this issue. Figure formal non-negative matrix factorization (GNMF) is based on manifold learning (manifold learning), the concept of the proposed algorithms, so that the original spatial data near the adjacent points will be projected after neighboring points. The results confirm that, NMF and GNMF clustering results are indeed beyond the principal component analysis (PCA), which is better than the effectiveness of GNMF and NMF. Cai *et al.* later made revisions to the original GNMF, adding the concept of semi-supervised learning. This has became a new approach: the semi-supervised non-negative matrix factorization formal diagram (SGNMF). SGNMF mainly modifies the weight matrix structure, so that the label (label) will be introduced when the information becomes available for graphical structure. Specifically, when the two data points is the same label (label), through the weight matrix can be adjacent data points, but if the two data points with different labels (label), the corresponding weight will be set to 0 . Combining all mentioned above, although the principal component analysis (PCA), independent component analysis (ICA), vector quantization (VQ) and isometric mapping (Isomap) have their advantages as linear and nonlinear dimensionality reduction method, but because these methods are not restricted within the matrix of positive and negative, it may result in cancellation of the positive and negative. This study attempts to use the SGNMF improved from NMF as a method to reduce the dimensions of the data, and compare it with the past several dimensionality reduction method constructed.

Asset allocation refers to the way in which you weight investments in your portfolio in order to try to meet a specific objective. For instance, if your goal is to pursue growth (and you're willing to take on market risk in order to do so), The asset classes you choose, and how you weight your investment in each stock, will probably hinge on your investment time frame and how that matches with the risks and rewards of each asset class. Grey relational analysis is adopted in our study for portfolio selection.

Based on the above intention of our study, this paper illustrates that rough set theory (RS), allied with the use of Grey Prediction, SGNMF and Grey relational Analysis of Grey System Theory, K-mean clustering, and Buffet investment rules together to construct a prediction model and applied it for selecting stock market portfolio. A trading system in Taiwan with good performance based on a forecasting model mentioned above is presented in this paper.

The structure of the rest of the paper is as follows: In Section 2, fundamental notions of the RS theory, the semi-supervised non-negative matrix factorization and grey theory are presented. Section 3 discusses application of this forecasting model to analyze stock data from the Taiwan Stock Exchange. In Section 4, the empirical results of this analysis are shown. Finally, the paper concludes with the discussion.

2. Methodologies Review

2.1. Rough Sets

Pawlak (1982) first introduced rough set theory. Rough set theory is a powerful mathematical tool to handle vagueness and uncertainty inherent in making decisions. The concept of that theory is founded on the assumption that every object of the universe of discourse is associated with some information. Objects characterized by the same information are indiscernible (similar)in view of the available information. The indiscernibility relation generated in this way is the mathematical basis for the rough set

theory. The most important problems that can be solved by Rough Sets Theory are: finding description of sets of objects in terms of attribute values, checking dependencies (full or partial) between attributes, reducing attributes, analyzing the significance of attributes, and generating decision rules [33].

(1) Information System

In the rough set theory, information systems are used to represent knowledge. An information system $S=(U,\Omega,V_q,f_q)$ consist of:

U -a nonempty finite domain set;

Ω -a nonempty, finite set of attributes;

$\Omega=C\bigcup D$, in which C is a finite set of condition attributes and D is a finite set of decision-making attributes; For each $q\in R$, V_q is definition domain of q ； f_q is information function； $f:U\to V_q$

Elements(X)can be interpreted as cases, states, processes, patients and observations. Attributes($C\&D$) can be interpreted as features, variables and characteristic conditions. A particular case in information system called decision-making table or attribute value table, in which the columns represent elements and rows describe attributes.

(2) Approximation of Sets

There are often some inexact, unclear parts in real world data; therefore the inconsistent among elements(X) appears to generate ($IND(R)$). This is it is hard to discern the relation or differences of two or above elements with various decision-making attributes in the condition set, herein the decision-making table is called indiscernibility table. However, the major function of approximate sets in rough set theory is to handle with this indiscernibility of elements. This decision table is called inconsistent decision table. In the rough set theory, the approximations of sets are introduced to deal with inconsistency. If $S=(U,\Omega,V_q,f_q)$ is a decision table, $R\subseteq\Omega$ and $X\subseteq U$, then R^- and $R\_$ are the upper and lower approximate of X ,are defined, respectively, as follows:

$$R^-(X)=\bigcup\{Y\in U/IND(R):Y\bigcap X\neq\varphi\} \tag{1}$$

$$R\_(X)=\bigcup\{Y\in U/IND(R):Y\subseteq X\} \tag{2}$$

Herein $U/IND(R)$ expresses the equivalence of R ； $IND(R)$ is called indiscernibility of R, they are defined as：

$$IND(R)=\{(x,y)\in U^2:\ for\ every\ a\in R, a(x)=a(y)\} \tag{3}$$

When using attribute set R, the lower approximate set R_X describes the set of the completely same rank elements(X) under Y decision-making attribute, R^-X represents the set of the possible same rank elements (X) under Y decision-making attribute. The set $BN_R(X)=R^-(X)-R\_(X)$ is called boundary set of X .

2.2. Grey Prediction Modeling(GM(1,1))

The grey prediction model has three operations: (a) accumulated generation, (b) inverse accumulated generation, and (c) grey modeling. The grey prediction model uses the operation of accumulated generation to build a differential equation. Intrinsically speaking, it has the attributes of requiring less data to construct the model.

The GM (1, 1) model constructing process is described below:

Denote the original data sequence by

$$x^{(0)} = \left(x^{(0)}(1), x^{(0)}(2), x^{(0)}(3), \cdots, x^{(0)}(n) \right) = (x^{(0)}(k); k = 1, 2, ..., n) \quad (4)$$

where is the number of years observed.

The AGO formation of $x^{(0)}$ is defined as:

$$x^{(1)} = \left(x^{(1)}(1), x^{(1)}(2), x^{(1)}(3), \cdots, x^{(1)}(n) \right) \quad (5)$$

Where n is

$$x^{(1)}(1) = x^{(0)}(1), \text{ and } \quad x^{(1)}(k) = \sum_{m=1}^{k} x^{(0)}(m), \quad k = 2, 3, \cdots, n \quad (6)$$

The GM (1,1) model can be constructed by establishing a first order differential equation for $x^{(1)}(k)$ as:

$$dx^{(1)}(k)/dk + ax^{(1)}(k) = b \quad (7)$$

Therefore, the solution of Equation (7) can be obtained by using the least square method. That is,

$$\hat{x}^{(1)}(k) = \left(x^{(0)}(1) - \frac{\hat{b}}{\hat{a}} \right) e^{-\hat{a}(k-1)} + \frac{\hat{b}}{\hat{a}} \quad (8)$$

where

$$\left[\hat{a}, \hat{b} \right] = \left(B^T B \right)^{-1} B^T X_n \quad (9)$$

and

$$B = \begin{bmatrix} -0.5\left(x^{(1)}(1) + x^{(1)}(2) \right) & 1 \\ -0.5\left(x^{(1)}(2) + x^{(1)}(3) \right) & 1 \\ \vdots & \vdots \\ -0.5\left(x^{(1)}(n-1) + x^{(1)}(n) \right) & 1 \end{bmatrix}, \quad (10)$$

$$X_n = \left[\; x^{(0)}(2), x^{(0)}(3), x^{(0)}(4), \cdots, x^{(0)}(n) \; \right]^T \quad (11)$$

We obtained $\hat{x}^{(1)}$ from Equation (8). Let $\hat{x}^{(0)}$ be the fitted and predicted series,

$\hat{x}^{(0)} = (\hat{x}^{(0)}(1), \hat{x}^{(0)}(2), \hat{x}^{(0)}(3), ... \hat{x}^{(0)}(n))$,

Where $\hat{x}^{(0)}(1) = x^{(0)}(1)$.

Applying the inverse AGO, we then have

$$\hat{X}^{(0)}(k) = (X^{(0)}(1) - \frac{\hat{b}}{\hat{a}}) e^{-\hat{a}(k-1)} (1 - e^{\hat{a}}), k = 2, 3, ..., n \quad (12)$$

where $\hat{x}^{(0)}(1), \hat{x}^{(0)}(2), \cdots, \hat{x}^{(0)}(n)$ are called the GM(1,1) fitted sequence, while $\hat{x}^{(0)}(n+1), \hat{x}^{(0)}(n+2), \cdots$ are called the GM(1,1) forecast values.

2.3. Semi-supervised Graph Regularized Non-negative Matrix Factorization (SGNMF)

NMF and GNMF are both unsupervised algorithms, so when labeling (label) the information cannot be directly applied. To solve this problem, Cai *et al.* combine the GNMF semi-supervised learning concept, the new matrix decomposition method - semi-supervised non-negative matrix factorization formal diagram (SGNMF). This method can map the same labeled data points to low-dimensional space so that they are divided in the same category. Assuming that all raw data points would constitute a diagram for each data point v_i we can find the nearest neighbors between v_i and its nearest neighbor p, in which we will make some changes to the original weight matrix: Suppose there are c kinds of data classification, then

$$c_{ij} = \begin{cases} 1, if\ v_i v_j\ belongs\ to\ same\ classification \\ 0, \qquad\qquad others \end{cases} \tag{13}$$

$N_p v_i$ is defined as the data points v_i and p neighbors collection:

$$P_{ij} = \begin{cases} e^{\frac{\|v_j - v_i\|^2}{\sigma}} & if\ v_i\ EN_p(v_j) \quad or \quad v_j\ EN_p\ v_i \\ 0 & others \end{cases} \tag{14}$$

The above Equation (13, 14) combined into a new matrix S is the weight calculation map (15):

$$S = CP \tag{15}$$

Other parts are same with GNMF, they will map the lower dimension v_i to the new basis as $z_i = [h_{i1,}, ..., h_{i1,}]^T$. Euclidean distance can also be used in this to measure the dissimilarity of the two data points in the low dimensional map to the new group.

$$D(z_i, z_j) = \| z_i, z_j \|^2 \tag{16}$$

As defined above, the weight matrix S can be measured for the smoothness of the low-dimensional

$$R = \frac{1}{2} \sum_{i,j=1}^{N} \| z_i, z_j \|^2 S_{ij} \tag{17}$$

Where $tr(.)$ is the trace of the matrix, the diagonal matrix E is the weight matrix S is sum of column or row of elements, and (since S is a symmetric matrix), *i.e.*

$E_{ii} = \sum_l S_{il}$. Matrix $L = E - S$.

By minimizing R, we can expect if the two data points v_i, v_j are similar, then z_i, z_j, will be similar, and this combined with the NMF.

Given a data matrix $V = [V_1, V_2, ..., V_n] \in R^{N \times K}$, each row vector V for the sample, and finding the two non-negative matrices $W = [w_{ik}] \in R^{N \times K}$, $H = [h_{ik}] \in R^{N \times K}$, so that the two matrices will be approximate the original matrix V:

$$v \approx WH^T W \geq 0, H \geq 0 \tag{18}$$

In this approximation using the Euclidean distance will solve the problems:

$$D = \| V - WH^T \|^2 + \lambda tr(H^T L H) \tag{19}$$

Which λ is greater than or equal to 0 normalization parameter. Formula (19) above can be obtained when minimized as follows:

$$W_{ik} \leftarrow W_{ik} \frac{(VH)_{ik}}{(VH^T H)_{ik}} \tag{20}$$

$$h_{jk} \leftarrow h_{jk} \frac{(V^T + \lambda SH)_{jk}}{(HW^T W + \lambda EH)_{jk}} \tag{21}$$

2.4. Grey Relational Analysis

In the grey relational analysis (GRA), the data that contain the same features are regarded as a series. The relationship between two series is determined by the difference of the two series, and the difference measure refers to a value of background for generating a grey relational grade. Compared with the usual distance measurement, the GRA combines with the concept of wholeness and can express the relationship of the two objects more exactly and objectively.

The grey relational coefficient is defined as follows:

$$\gamma(x_0(k), x_i(k)) = \frac{\Delta_{\min} + \zeta \Delta_{\max}}{\Delta_{0i}(k) + \zeta \Delta_{\max}} \tag{22}$$

where $i = 1, 2, ..., m, k = 1, 2, ..., n,$ $x_0(k)$ is the reference value, and $x_i(k)$ is the comparative value. where $\Delta_{0i}(k) = \left\| x_0(k) - x_i(k) \right\|$, $\Delta_{\min} = \min_{\forall i} \min_{\forall k} \left\| x_0(k) - x_i(k) \right\|$, and $\Delta_{\max} = \max_{\forall i} \max_{\forall k} \left\| x_0(k) - x_i(k) \right\|$.

Define a distinguishing coefficient ξ, between 0 and 1. It is often set to 0.5. Professor Deng thought that 0.5 is the middle value so that the error could be reduced to the lowest.

The grey relational grade represents the relational measure of the respective elements. The grey relational grade is usually defined as the average of the grey relational coefficients.

$$\Gamma_{0i} = \gamma(x_0, x_i) = \frac{1}{n} \sum_{k=1}^{n} \gamma(x_0(k), x_i(k)) \tag{23}$$

After the grey relational grade is calculated, according the value, we can rank the sequence, and this procedure is called grey relational rank. If $\gamma(x_0, x_i) \geq \gamma(x_0, x_j)$, then we found that under the reference sequence $x_0(k)$, the grey relational rank of $x_i(k)$ is greater than grey relational rank of $x_j(k)$, the rank is $x_i \succ x_j$.

3. Research Model Development

3.1. Establishes Data Set for Rough Set By Grey Prediction Model

The mathematical formula for rough set theory is as follows where U is the domain and R is the equivalences of U,

Rough Set $X \subset U$ is : $(R\_(X), R^{-}(X)), BN_R(X)$

Here in : X is set of elements ; $U / IND(R)$ is the equivalent of R, $IND(R)$ is the indiscernibility of R ; φ is zero set ; R is the attribute set of X, which includes condition set （C） and decision-making set (D) ; $R\_(X)$ is lower approximate of X, $R^{-}(X)$ is upper approximate of X, $BN_R(X)$ is boundary of X. Every element in domain U ($X \subseteq U$) has its attribute set （R）, which describes the particular value of X.

We hypothesize in this study that if we give every X ($X \subseteq U$) of U a trend prediction attribute value $R = (C_1, C_2, ..., C_n, D_1, D_2, ..., D_m)$, then let them go through rough set system, The typical prediction models need many complicated calculations and tests to form the correct and practical model for unclear and fuzzy system information forecast before the independent and dependent relation between the factors is determined. However, the grey system theory referred by professor Deng can be used to find out the data regularity and the fit rule modeling and prediction through its data adding up generation, subtracting out generation, Grey Relation, *etc.* It is showed that its prediction effect is much better than the typical ones. It is proved that it can solve the problems of fuzzy, uncertain information system controlling and decision-making in many research projects covering various aspects.

For the rough set model, the particular attribute values $C_1 \sim C_n, D_1 \sim D_m$ of every element (X_i) in U were assigned to their trend prediction value through Grey Generation Modeling, then, were integrated to trend of U. We set the adding times (n) equal to 4 and regarded four continuous time-serial actual value as generation point of Grey Generation Modeling to predict the value of next time point because it showed in the relative Grey Generation times and Grey Prediction error [34] that the error is minimum and the prediction value was accurate relatively when $n = 4$.

3.2. Decision-Making Attributes Selecting: Combining the Taiwan Industry Features with Buffet's Rules

The decision-making attribute is the key part of the whole model selecting process in rough set theory. Its preciseness determined the qualification of the selected companies. We combined the valuable Buffet's investment rules with the industry features of Taiwan and the past relative research to deduce seven major decision-making attributes for our model and expected to make it approach completely through screening level by level.

In Enterprise and Market Principle of his four investment principles, Buffet believed that an enterprise with good outlook should hold the lion share of the consumption market of its product line so that it should own attributes such as "the lowest cost", " high profit margin", "high inventory turnover". He thought the lower cost it owned the bigger competition capacity it had and the more possible it could rival the competitors in price strategy. In addition, high profit margin plus high inventory turnover described the real characteristic of a profit-making company. Only the companies with all these three attributes could improve manufacture process, develop new products, or benefit from merging and acquiring other enterprises so that it could run on-going and earn profit for shareholders.

We deduced the following seven standards of decision-making attributes of rough set theory:

D1: return on asset (after tax) >0
D2: return on equity >0
D3: gross profit ratio >0
D4: equity growth rate >0

D5: quick ratio > median of all industry
D6: inventory turnover rate > median of all industry
D7: constant EPS >0

Since the beginning, we expected the rough set model like a net which might catch all good fishes but not one just catching big fishes, that means, we did not want to omit the enterprises with some potential because of the too strict stock selecting standard. Therefore, we defined the threshold values of part decision-making attributes (D1, D2, D3, D4, and D7) just > 0.

3.3. Modeling Flow of the Stock Portfolio Model

Basing on those theories mentioned above, we developed a standard stock selection process to filter investment targets in Taiwan Stock Market in order to find out the best portfolio. In the following, we explained data processing course step by step from the first to the final. Details were showed in Figure 1, as following:

(a) Flow Chart of Model Establishment (Figure 1)

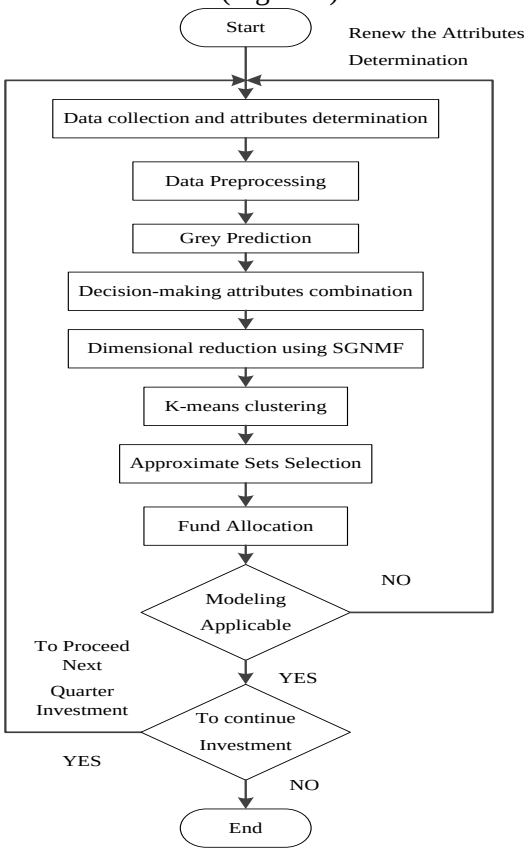

Figure 1. Flow Chart of Stock Portfolio Modeling

(b) Steps of the Modeling Flow

Step1：Data collection and attributes determination

Before selecting stock, we needed to set up range and object to be explored and make decision about conditional attributes and decision-making attributes. After attributes determined, we chose proper database and then collected relative financial ratio datum within the target study period, whose frequency should be quarterly.

Since conditional attributes selection was focusing on evaluation the quality of a company, we suggested using seven classes to conduct selection, such as profitability,

capitalized cost ratio, individual share ratio, growth rate, debt ratio, operation leverage, statutory ratio *etc*. And the function of decision-making attribute was to test whether the profitability of company was good or not from viewpoint of investors so that they could make investment target selection.

Step 2：Data Preprocessing

After relative datum collected, we conducted basic process in order to improve their usefulness. First, we eliminated the columns of omission ones because they could not be operation in the system if some values left out.

In practical operation, some very big or very small values often appeared in the study materials, which were so-called outliers. In order to handle such datum, we use the Box Plots method to set up an inter-quartile range in our study: if the data exceeded the range, it would be seemed as an outlier and eliminated for the interval.

Step 3：Grey Prediction

We adopted GM (1,1) rolling model in grey prediction to conduct rolling trend dynamic prediction in the study, which was first constructing a GM (1,1) model by using the very previous four datum in the same series to predict value of the next data; and next building another GM (1,1) model with the next four datum and predicting another next data; then continuing the same process again and again to the last data.

Step 4：Decision-making attributes combination

A big shortcoming of traditional rough set theory was it had to set threshold when decision-making attributes were conducted to transfer datum, which might generate personal subjective judgments and lose objective that a stock selection system should hold. We tried to combine many decision-making attributes in information system into a most important one by adaptable neuro-fuzzy inference system (ANFIS) [35] in order to overcome the deficiency of setting threshold in traditional rough set theory. In this step, it mainly let many inputs to be a single output through a series of learning and drilling. The advantage of using ANFIS was it did not need to set complicate rules like fuzzy theory, which caused disturbances for system operation.

Step 5：Dimensional reduction using SGNMF

The basis for deleting conditional attributes was the reliability how decision-making attributes depend on the conditional attribute. In order to find out the significance of every attribute, we used the SGNMF method to find out the top ten rank between influencing sequences(conditional attributes) and the major sequence(decision-making attribute).

Step 6：K-means clustering

We adopted K-means clustering method to transfer dynamic trend datum predicted in order to find out useful information within the disorder and unsystematic datum. While conducting K-means clustering, we set three ($K = 3$)groups as. Every quarterly conditional attributes ($C_1 \sim C_n$)were divided into three groups by using K-means clustering tool.

Step 7: Approximate Sets Selection

We modified conditional attributes and combined decision-making attributes. Then datum in the table was inputted into operation system and the rough set method was applied to select low approximate set. The generalized rules extracted by the rough sets model were all recognized rules or relationships in the investment industry; this indicated that the rough set analysis was useful in determining the contributions of attributes to identify top stock performers, allowing also for a construction of the decision rules which might be applied to the evaluation of new stocks.

What we did let investors shake off the investment methods of following blindly or dashing darts, which most of them took in the past, and create a set of stock selection strategy which could be tested and proved by scientific method.

Step 8: Fund Allocation

Fund allocation plays a key role on the whole portfolio performance so how to determine the weight in portfolio in order to reach the premium rate of return become an important work in our study. Since great achievements had been made in many aspects under the study of Grey Relation Analysis of Grey System Theory, we expected to use Grey Relation Sequence tool to find out the best Fund allocation proportion to improve investment benefit effectively. Meanwhile, we picked another Fund allocation method (average weighted allocation method) as its comparative groups.

The formulas are:

(1)Fund allocation by Grey Relation Sequence：

$$stock\ weight(i) = \frac{n-i+1}{\sum_{i=n} i}$$

Where i : grey relation order of each stock, n : the number of stock investing

$$stock\ weight \quad = \frac{1}{number\ of \quad stock \quad investing}$$

(2)Average weighted fund allocation：

After all of above steps, two decisions must be carried out. Firstly, whether the stock portfolio modeling is applicable for investment should be checked by the rate of return. If not, we renew the attributes determination. Then, make the subsequent stage depending on whether conducting next quarter investment.

4. The Empirical Results of Stock Portfolio Model in Taiwan Stock Market

4.1. Material Reasoning

This study focuses on electron sector stock to sift the optimal stock portfolio out applying the financial statement datum from the New Taiwan Economy database(TEJ).The data period was from the first quarter in 2009 to the fourth quarter in 2012,totally 16 quarters financial ratio datum; and the period forecasted was from the first quarter in 2010 to the first quarter in 2013,totally 13 quarters. This study used GM(1,1) to predict dynamic trend, to forecast the fifth quarter datum by the previous four quarters datum.

Herein, it should be noticed that the time when financial statements were published was pretty late: annual report would be published after 4 months; half-year report would be 2 months late and even the first and third quarterly report(without notarization) would have to be waited for 1 month. And the finishing times when the file of financial statement datum in New Taiwan Economy database(TEJ) was build were:

(1) Annual report(the sending period Security Superintendence Commission required was within 4 months after the closing balance day.)---listed companies in past years(TSE and OTC) should finish filing before 5/31.

(2) Half-year report(the sending period Security Superintendence Commission required was within 2 months after the closing balance day.) ---listed companies in past years(TSE and OTC) should finish filing before 9/21.

(3) First quarter report(the sending period Security Superintendence Commission required was within 1 months after the closing balance day.) ---listed companies in past years(TSE and OTC) should finish filing before 5/31.

(4) Third quarter report(the sending period Security Superintendence Commission required was within 1 months after the closing balance day.) ---listed companies in past years(TSE and OTC) should finish filing before 11/15.

Since this study used GM(1,1) rolling prediction method to forecast data dynamic trend, it had to take four-quarter financial ratio datum to predict the next quarter ones. However, the last quarter datum every year could be gained until 31st May in next year, it could not predict the datum of first quarter in the next. So, we could conduct forecast only three times(5/31~09/22,9/22~11/15,11/15~05/31 next year) every year.

4.2. Difference of Fund Allocation and Their Investment Results

(1) Between the two fund Allocation methods-- Grey Relation Sequence(GRS), Average Weight(AV)—Grey Relation Sequence was much different from Average Weight; as shown in Table 1(a) and Table 1(b).

Table 1. (a) The Returnrate in 2nd Quarter of 2011 Using AW Fund Allocation

Company	Investment (ten thousands)	Call (5/31)	number	Put (9/22)	returnrate
6131	20.00 For each company	38.19	5	35.21	-1.60%
2391		41.63	4	51.62	4.31%
3035		38.13	5	33.67	-2.40%
6141		27.91	7	21.86	-4.56%
2460		16.79	11	16.69	-0.12%
Total returnrate： -4.38%					

Table 1. (b) The Returnrate in 2nd Quarter of 2011 Using GRS Fund Allocation

Company	Investment (ten thousands)	Call (5/31)	number	Put (9/22)	returnrate
6131	33.33	38.19	8	35.21	-2.63%
2391	26.67	41.63	6	51.62	6.60%
3035	20.00	38.13	5	33.67	-2.45%
6141	10.33	27.91	4	21.86	-2.67%
2460	6.67	16.79	3	16.69	-0.03%
Total returnrate： -1.18%					

(2) Under rough set allied with the use of dynamic grey method long-term mechanic investment rules, GRS fund Allocation method is better on accumulated yearly return and average accumulated return;

(3) Through comparing the results of two methods, Grey Relation Sequence had less loss when the negative rate of return happened in the second and third quarter of 2011.

Summed up the above three opinions, we discovered GRS method owned the best performance between them, so we regarded its result as the final result of implementation in Taiwan market (Table 2).

Table 2. Fund Allocation Methods and Rates of Return

Investment period		Fund Allocation Method			
		Grey Relation Sequence		Average Weight	
		Quarter	Yearly	Quarter	Yearly
		Rate of return			
2010	Second quarter	-1.18%	18.31%	-4.38%	9.46%
	Third quarter	-0.83%		-2.49%	
	Fourth quarter	20.32%		16.33%	
2011	Second quarter	5.93 %	30.03%	3.47%	27.13%
	Third quarter	-0.73 %		0.13%	
	Fourth quarter	24.83%		23.53%	
2012	Second quarter	-6.57%	12.88%	-2.20%	20.28%
	Third quarter	-5.29%		-5.08%	
	Fourth quarter	24.74%		27.57%	
Average year rate of return		20.41%		18.96%	
Accumulated 3 years rate of return		61.22%		53.66%	

5. Conclusions and Discussions

The system $U = X\{x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$ is the set of general financial information of the companies. The financial ratio subsets $R = (C_1, C_2, ...C_n ; D_1, D_2...D_m)$ may have some relativity and independence. Herein, we regarded the subsets as independently individual data to conduct summing up generation modeling and predicting such that we acquired the subset independent trend system $\hat{U} = (\hat{x}_0, \hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4, \hat{x}_5, \hat{x}_6, \hat{x}_7)$ and element attribute subset $R = (\hat{C}_1, \hat{C}_2, ..., \hat{C}_n, \hat{D}_1, \hat{D}_2, ..., \hat{D}_m)$. Then, we classified them into the trend upper approximated set and the trend lower approximated set through rough set model attributes discerning.

This stock portfolio model has the following attributes:

(1)Because this model applied rough set model allied with the use of grey prediction theory, It has both functions of rough set model, including attribute deleting and discerning, and grey theory, such as grey prediction and SGNMF, *etc*.

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

(3) It can be used in a very various sphere, like investment risk-avoid, dynamic enterprise sound inventory modeling, ecological environment evaluation, particular illness outbreak prediction and evaluation, *etc*.

(4) The decision rules obtained from the rough sets model are based on facts of eliminating any redundancy or unimportant typical of the original data. The eliminating attributes process is called information reduction,. In the study, we use SGNMF method to get rid of the conditional attributes that have not important relationships with decision-making attributes.

(5) Its rolling share selecting method working in with the decision-making attribute screening institute can correct the prediction values quickly according to the financial statues of last quarter so its forecast error is small and prediction trend is precise. Both the decision-making quality and performance are good so that it is pretty close to the practice.

(6) Our experimental showed that its portfolio rates of return that use Grey Relation Sequence Fund Allocation method is better than Average Weight one, so we regarded its result as the final result of implementation in Taiwan stock market.

The main purpose of model creation and improvement is to close to the real world and benefit human beings. However, the key factor of whether a prediction tool is used correctly or not or a model is constructed properly or not is the matching degree of the data feathers and the prediction tool because it determines the accuracy and error range of the forecast values. To build a new model successfully needs many tests and continuous adjusting at the right moment.

In the above study of rough set model for stock portfolio, it has been demonstrated that rough set model is a promising alternative method to conventional methods. It got great achievements on company evaluation. We expect that it can bring about a broad range of derivatives in other research fields.

Acknowledgements

This research is supported by the Science and Technology Project of Education Department of Fujian Province under contract Nos. JA14331, the recruiting high level talent program of Ningde Normal University under contract Nos.2014Y005 , School Innovation Team of Ningde Normal University under contract Nos.2015T001, the Vital Construction Project of Ningde Normal University under contract Nos. 2012H311.

References

[1] R. Brachman, T. Khabaza, W.Kloesgen, G.Piatetsky-Shapiro and E. Simoudis, “Mining business databases”, Communications of ACM, vol. 39, no. 11, **(1996)**, pp.42-48.
[2] J. Bauer Jr., “Genetic Algorithms and Investment Strategies”,Wiley, New York, **(1994)**.
[3] G. Mani, K. K. Quah, S. Mahfoud and D. Barr, “An analysis of neural-network forecasts from a large-scale, real world stock selection system”, Proceedings of the IEEE/IAFE, Conference on Computational Intelligence for Financial Engineering (CIFER95), IEEE, New Jersey, pp. 72–78, **(1995)**.
[4] C. Skalko, “Rough sets help time the OEX”, Journal of Computational Intelligence in Finance, vol. 4, no. 6, **(1996)**, pp. 20–27.
[5] L. Lin and J. Piesse, “Identification of corporate distress in UK industrials: A conditional probability analysis approach”, Applied Financial Economics, vol. 14, **(2004)**, pp. 73–820.
[6] Z. Pawlak, “Rough sets”, International Journal of Information and Computer Sciences, vol. 11, no. 5, **(1982)**, pp. 341-356.

[7] S. Tsumato, S. Slowinski, J. Komorowsk and J. W. Grzymala-Busse, "The fourth international conference on rough sets and current trends in computing. RSCTC'2004, lecture notes in artificial intelligence, **(2004)**.
[8] R. Li and Z.-O. Wang, "Mining classification rules using rough setsand neural networks", European Journal of Operational Research, vol. 157, **(2004)**, pp. 439–448.
[9] R. Jensen, "Combining rough and fuzzy sets for feature selection: Ph.D. Dissertation", School of Informatics, University of Edinburgh,UK, **(2004)**.
[10] J. W. Grzymala-Busse, "LERS-A system for learning from examples based on rough sets", In: Slowinski, R. (Ed.), Intelligent Decision Support – Handbook of Applications and Advances of the Rough Sets Theory. Kluwer Academic Publisher, Dordrecht, no. 1, **(1992)**, pp. 3-18.
[11] J. W. Grzymala-Busse, "LERS-A knowledge discovery system", In: L. Polkowski and A. Skowron, (Eds.), Rough Sets in Knowledge Discovery, Physica-Verlag, Wurzburg, vol. 2, **(1998)**, pp. 562–565.
[12] V. Stefanowski, "The rough set based rule induction techniques for classification problems", In: Proceedings of the 6th European Congress on Intelligent Techniques and Soft Computing 1, Aachen, September 7–10, **(1998)**, pp. 109– 113.
[13] J. G. Bazan, A. Skowron and P. Synak, "Market data analysis: A rough set approach", ICS Research Reports 6/94, Warsaw University of Technology, **(1994).**
[14] J. K. Baltzersen, "An attempt to predict stock market data: a rough sets approach", Diploma Thesis, Knowledge Systems Group, Department of Computer Systems and Telematics, The Norwegian Institute of Technology, University of Trondheim, **(1996)**.
[15] J. G. Bazan and M. Szczuka, "RSES and RSESlib – A collection of tools for rough set computations", In: Ziarko, W., Yao, Y.Y. (Eds.), Proceedings of the Second International Conference on Rough Sets and Current Trends in Computing (RSCTC'2000), Banff, Canada, **(2000)**, pp. 74-81.
[16] J. L. Deng, "Introduction to Grey System Theory", The Journal of Grey System, vol. 1, no. 1, **(1989)**, pp.1-24.
[17] J. L.Deng, "Grey Differential Equation", The Journal of Grey System, vol. 5, no. 1, **(1993)**, pp.1-14.
[18] W. Ziarko, "Variable precision rough set model", Journal of Computer and System Sciences, vol. 46, no. 3, **(1993)**, pp. 39–59.
[19] A. Lenarcik and Z. Piasta, "Discretization of condition attributes space", In: Slowinski, R. (Ed.), Intelligent Decision Support – Handbook of Applications and Advances of the Rough Sets Theory. Kluwer Academic Publishers, Dordrecht, **(1992)**, pp. 373–389.
[20] A. Lenarcik and Z. Piasta, "Rule induction with probabilistic rough classifiers. ICS Research Report 24/96, Warsaw University of Technology, **(1996)**.
[21] A. Lenarcik and Z. Piasta, "Learning rough classifiers from large databases with missing values. In: L Polkowski and A. Skowron, (Eds.), Rough Sets in Knowledge Discovery 1, vol. 1. Physica-Verlag, Wurzburz, **(1998)**, pp. 483–499.
[22] A. Lenarcik and Z. Piasta, "ProbRough – A system for probabilistic rough classifiers generation. In: Polkowski, L., Skowron, A. (Eds.), Rough Sets in Knowledge Discovery, Physica-Verlag, Wurzburg, vol. 2, **(1998)**, pp. 569–571.
[23] W. Kowalczyk, "TRANCE: A tool for rough data analysis, classification, and clustering. In: Tsumoto, S., Kobayashi, S., Tanaka and H., Nakamura, A. (Eds.), Proceedings of the Fourth International Workshop on Rough Sets, Fuzzy Sets and Machine Discovery, RSDF'96, Tokyo, **(1996)**, pp. 269–275.
[24] W. Kowalczyk, "1998b, TRANCE: A tool for rough data analysis, classification, and clustering. In: Polkowski, L., Skowron, A. (Eds.), Rough Sets in Knowledge Discovery Physica-Verlag, Wurzburg,, vol. 2. **(1998)**, pp. 566–568.
[25] R. Slowinski, "Rough set learning of preferential attitude in multi-criteria decision making. In: Komorowski, J., Ras, Z.W. (Eds.), Methodologies for Intelligent System. Lecture Notes in Artificial Intelligence, Springer, Berlin, vol. 689, **(1993)**, pp. 642–651.
[26] R. Slowinski and J. Stefanowski, "Rough classification with valued closeness relation", In: Diday, (Eds.), New Approaches in Classification and Data Analysis. Springer, Berlin, **(1994)**, pp. 482–488.
[27] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization". Nature, vol. 401, no. 6755, **(1999)**, pp. 788–791.
[28] I. Buciu and I. Pitas, "Application of non-negative and local non negative matrix factorization to facial expression recognition", In Proc. of Intl. Conf. Pattern Recogn. (ICPR), **(2004)**, pp. 288–291.
[29] M. Heiler and C. Schn¨orr, "Reverse-convex programming for sparse image codes", In Proc. of Energy Minim. Methods in Comp. Vision and Pattern Recog. (EMMCVPR), Springer, vol. 3757 of LNCS, **(2005)**, pp. 600–616.
[30] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints", J. of Mach. Learning Res., no. 5, **(2004)**, pp. 1457–1469.
[31] J. B. Tenenbaum, V. de Silva and J. C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction", Science , vol. 290, no. 5500, **(2000)** December 22, pp. 2319-2323.
[32] X. H. DengCai, J. Han and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 8, **(2011)**, pp. 1548–1560.

[33] Z. Pawlak, "Rough sets. In:Lin,T.Y.,Cercone, N. (Eds.)", Rough Sets and Data Mining. Kluwer Academic Publisher,Dordrecht, **(1997)**, pp. 3–8.
[34] J. L. Deng, "Grey Differential Equation", The Journal of Grey System, vol. 5, no. 1, **(1993)**, pp.1-14**.**
[35] J.-S. R. Jang, "ANFIS: adaptive-network-based fuzzy inference system", IEEE Transactions on Systems, Man and Cybernetics, vol. 23, no. 8, **(1992)**, pp.665-685.

Authors

Shiliang Zhang, he graduated from Department of Physics of Fujian Normal University in 1996 and received Master degrees from Information Engineering College of Jiangxi University in 2008. He is now a vice Professor of the Department of Computer science, Ningde Normal University and as a visiting scholar in Peking University. He has presented more than a dozen of papers and managed participated more than a dozen of projects. He has a lot of experience about project development, especially has researched a lot in electronic map and 3D simulation and gets many achievements.

Tingcheng Chang, he received the M. S. and Ph. D. degrees in Process Control and Mechanical Engineering in 1992 and 1996, respectively, from University of Houston and University of Texas at Arlington, Texas, USA. He is now a Professor of Computer and Information Engineering, Ningde Normal University, Fujian, China. His research interests lie in the field of Grey Theory, Data Mining, Industry Product Design and Optimal Theory.

