Applied Rough Set Logics for Multi-criteria Decision Analysis in Stock Market Prediction

Wen-Rong Jerry Ho*

* Dept of Banking & Finance, Chinese Culture University, Yang-Ming-Shan, Taipei 111, Taiwan

Corresponding author: **Wen-Rong Jerry Ho**

Department of Banking & Finance, Chinese Culture University,

55, Hwa-Kang Road, Yang-Ming-Shan, Taipei 111, Taiwan

wrho@ms26.hinet.net. Tel.: +886 921368989

Applied Rough Set Logics for Multi-criteria Decision Analysis in Stock Market Prediction

Abstract

The purpose of this study is to propose a rule based forecasting method for predicting future stock market fluctuations. This research intends to establish a stock market indicators prediction framework using a multiple criteria decision making model consisting of the cluster analysis technique and Rough Set Theory to select the important attributes and forecast TSEC Capitalization Weighted Stock Index. The proposed prediction model was leveraged to predict the index in first half year of 2009 with an accuracy of 66.67%. The results indicate the decision rules were authenticated to employ in predicting the stock market fluctuations appropriately.

Keywords: Stock market, Cluster Analysis (CA), Rough Set Theory (RST), Multiple Criteria Decision Making (MCDM)

1. Introduction

The predictions of stock market fluctuations are regarded as one of the most important task for investors. However, investors always have problems correctly predicting the trends of future stock market development and making precise investment decisions due to the nonlinear nature of the stock market behavior which is extremely hard to predict without experienced or expert knowledge (McMillan, 2007). An appropriate forecasting of stock market fluctuations can assist investors in avoiding investment risks and enhancing profitability. Albeit difficult, scholars and investors continue to seek possible means to model stock market behaviors correctly. However, most stock market forecast models are either hard to be manipulated or difficult to understand due to their mathematical formulations. Therefore, a simple rule-based method which can generate "if … then" rules for stock market predictions will be very helpful. Thus, the purpose of this research is to propose a rule-based forecasting (RBF) method for predicting future stock market fluctuations.

To verify the feasibility and effectiveness of this proposed rule-based research framework, the historical data of TSEC Capitalization Weighted Stock Index (TAIEX) is presented as the decision attribute while fluctuations in the derived macroeconomic indicators from 1999 to 2008 are put on as the condition attributes. To verify the feasibility of the derived results, the first half year of 2009 historical data of the above mentioned stock market

index and macroeconomic indicators is set up to verify the decision rules being derived by the proposed multiple criteria decision making (MCDM) framework.

The remainder of this paper is organized as follows. The related literature is reviewed in Section 2. The novel MCDM prediction framework consisting of the Delphi method, Cluster analysis and Rough Set Theory based rule derivation method are introduced in Section 3. An empirical study based on historical data of TAIEX is brought in to verify the feasibility of the proposed framework in Section 4. Some implications and discussion were presented in section 5. Finally, we present some concluding remarks in Section 6.

2. Applied stock market Rough Set Theory forecast model

The purpose of this Section is to identify the stock market forecast and Rough Set Theory applications based on past literatures and discuss the areas of scarcity in these studies. To make up for such a lull, this study conducted a literature review of the predictions of stock market fluctuations for the sake of more accurately predicting the trends of future stock market development.

Roh (2007) combined a neural network and time series model for forecasting the volatility of the stock price index. Huarng and Yu (2006) applied the backpropagation neural network to establish fuzzy relationships in fuzzy time series for forecasting stock price. Tseng et al. (2001) developed a fuzzy ARIMA model for forecasting the exchange market. Nikolopoulos and Fellrath (1994) combined genetic algorithms (G.A.) and a neural network to develop a hybrid expert system for investment advising. Kimoto et al. (1990) proposed a stock market prediction system using a neural network. The distributions of stock data do not really follow statistical assumptions in practical stock markets about data distributions for the conventional financial time series models (Cheng and Wei, 2009). Box and Jenkins (1976) provided the autoregressive moving average (ARMA) model to perform forecasting at the linear stationary condition. The ARIMA model was introduced under the non-stationary condition to describe such homogeneous non-stationary behavior (Box

and Jenkins, 1976). Engle (1982) purported the ARCH(p) (Autoregressive Conditional Heteroscedasticity) model that has been used by many financial analysts to forecast stock market. Bollerslev (1986) suggested the GARCH (Generalized ARCH) model to refine the ARCH model. Nelson (1991) presented the EGARCH (Exponential GARCH) model to overcome the drawbacks of the GARCH model, and leverage effects. J. P. Morgan (1996) proposed the exponentially weighted moving average (EWMA) model to strengthen the initial GARCH model -- non-stationary GARCH (1,1) model.

Some macroeconomic indicators that were discovered to be noteworthy of expected stock return included money supply (Bilson et al., 2001; Kwon and Shin, 1999; Mandelker and Tandon, 1985; Rogalski and Vinso, 1977; Robichek and Cohn, 1974), inflation (Kim and In, 2005; Park and Ratti, 2000 ; Balduzzi, 1995; Gultekin, 1983; Fama, 1981), interest rates (Abugri, 2008; Domian et al., 1996; Geske and Roll, 1983; Kim and Wu, 1987), and industrial production (Ferson and Harvey, 1998; Fama, 1990; Chen et al., 1986). Bilson et al. (2001) found that money supply have explanatory power over stock returns in six emerging markets. Kwon and Shin (1999) concluded that money supply provide a direct long-run equilibrium relation with each stock price index. On their study, Kim and In (2005) demonstrated that there is a positive relationship between stock returns and inflation at the shortest and the longest

period of the time. Ferson and Harvey (1998) as well as Fama (1990) discovered industrial productivity has positively related to stock returns.

According to Walczak andMassart (1999), Rough Set Theory does not require any hypothesis or external information, and it can deal with vagueness and uncertainty of information, hence, Rough Set Theory offers an alternative toolset for financial and business analysis. Teoh et al. (2008) employed Rough Set Theory to fuzzy logical relationship from time series and an adaptive expectation model to adjust forecasting results, to improve forecasting accuracy. Shen and Loh (2004) investigated and forecasted stock market movements by retrieving knowledge that could lead investors on when to buy and sell. Susmaga et al. (1997) utilized Rough Set Theory to identify and pick top stock performers by arguing Rough Set Theory can offer powerful tools for constructing decision rules in evaluating the performance of new stocks. Yao and Herbert (2009) utilized time-series data analysis with Rough Set Theory to create rough rules from the New Zealand stock exchange.

3. Concepts of Rough Set Theory, Delphi algorithm and Cluster Analysis Method for Decision Analysis

In this section we briefly introduce Rough Set Theory and its use in analyzing the attributes of combination values for making insurance marketing decisions. In Section 3.1 Delphi algorithm is described to summarize the macroeconomic indicators. In Section 3.2 Concepts of Cluster Analysis are demonstrated to group the numbers of the same indicator with the same fluctuation tendencies for more precise predictions. Then, in Section 3.3 Rough Set Theory for decision-making is presented. At last, the derived rules are validated by the first half year of 2009 TAIEX dataset.

3.1. Delphi algorithm

The Delphi algorithm is a procedure to ‘‘obtain the most reliable consensus of opinion of a group of experts . . . by a series of intensive questionnaires interspersed with controlled opinion feedback’’ (Dalkey & Helmer, 1963). The method was first originated during the 1950s at the RAND Corporation as an alternative to develop information systems models of effects of Soviet weapons systems (Turoff, &. Linstone, 1975).

The Delphi algorithm is “a methodology for efficiently obtaining concensus from a panel of evaluators on questions which are shrouded in uncertainty and can not be

measured or evaluated in the classical sense" as described by Pill (1971). According to Pill (1971) and Whitman (1990), the elements and the route of Delphi algorithm include some degree of anonymity for the individual responses, multiple iterations with controlled feedback, and statistical summaries. Carl et al. (2008) deduced that the Delphi algorithm is well suited for use on the Internet because of the rapid communication of Delphi rounds resulting in reduced cost and time of collection for data analysis.

On the whole, the Delphi algorithm accepts for participation without restrictions of geographical location and time. It also allows for anonymous involvement, which is deemed to decrease partiality from defense of more prestigious or multiple members of a panel (Fink et al., 1984; Milholland et al., 1973).

3.2. Concepts of Cluster Analysis

Cluster analysis (CA) is a multivariate method for data reduction that sorts observations into similar sets or groups. Cluster Analysis sets notices on finding subsets, called clusters, which are standardized and/or well alienated. There are several types of clustering analysis methods, *k*-means cluster analysis has a time complication and is thought to generate poorer clusters, and hierarchical agglomerative cluster analysis (HACA) is often described as the better property

clustering approach. When the sample size is larger than 100, *k*-means cluster analysis should be used. Therefore, because it is difficult to find the proper *k* value before clustering, hierarchical agglomerative cluster analysis should be used because of no need to find *k* value (Hartigan, 1975; Massart and Kaufman, 1992). The procedure of hierarchical agglomerative cluster analysis starts with each case in a separate cluster and forms in a stepwise conduct progress to reduce clusters until only one is left. However, the algorithm of the *k*-means cluster analysis is based on the nearest centroid sorting to determine group membership, that is, a case uses the fact that the progress of the cluster groups (centroids) is the smallest for consecutive iterations.

An important step in most clustering is to select and to work with a correct distance measure (Johnson, 1967). One of the most common distance measure is the squared Euclidean distance, which served as the proximity measure, whereas the cluster formation was achieved through Ward's criterion (Ward, 1963; Massart et al, 1983). For the hierarchical agglomerative cluster analysis, according to Huang, Qiu, and Guo (2009), the expression of the squared Euclidean distance between two cases (D_{ab}^2) can be expressed as:

$$D_{ab}^2 = \sum_{j=1}^{n} \left(x_{aj} - x_{bj} \right)^2 \quad (1)$$

Where *n* is the number of the variables. For the *k*-means CA, the squared Euclidean distance of each cluster (D_k^2) can be calculated as:

$$D_k^2 = \sum_{j=1}^{n} \sum_{i=1}^{m_k} \left(x_{ijk} - \overline{x_{jk}} \right)^2 \quad (2)$$

where x_{ijk} represents the value of variable *j* for case *i* in cluster *k*, $\overline{x_{jk}}$ is the center of the cluster *k*, and m_k is the number of the cases in cluster *k*.

Using the following formulas (Johnson and Wichern, 2007), the *i*th coordinate, *i*=1,2,..., *p*, of the centroid is easy to be updated:

$$\bar{x}_{i,new} = (n\bar{x}_i + x_{ji}) / (n+1) \quad \text{if the } j\text{th item is added to a group} \quad (3)$$

$$\bar{x}_{i,new} = (n\bar{x}_i - x_{ji}) / (n-1) \quad \text{if the } j\text{th item is removed from a group} \quad (4)$$

where *n* is the number of items in the group, before items added or removed, with centroid $\bar{x}' = (\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_p)$.

Finally, according to Huang, Qiu, and Guo (2009), in order to validate the partition of the clusters, analysis of variance (ANOVA) is executed with the significance level less than 0.001.

The two-step cluster combines hierarchical agglomerative cluster analysis and *k*-means cluster analysis. Hierarchical agglomerative cluster analysis is employed to find the initial estimation of clusters. *K*-means cluster analysis is utilized to split objects into clusters. In this study (SPSS, 2001), we first use a sequential cluster method to the large dataset to compress the dense regions and form sub-clusters. We then operate a cluster method to the sub-clusters to obtain the desired amount of clusters.

3.3. Rough Set Theory for decision-making

Rough Set Theory was proposed by Pawlak (1982, 1984, 2004), which is an effective rule-based decision-making method for extracting information from data tables. The technique can manage crisp datasets and fuzzy datasets with no necessity for a pre-assumption membership function, which fuzzy theory needs. According to Pawlak (2002), Rough Set Theory philosophy is originated on the assumption that is related to discourse about every element of the universe and our information (knowledge). For example, if objects are patients suffering from a particular disease, symptoms of the disease form information about patients. Objects illustrated by the similar information are unobvious in view of the available information about them. The framework for discovering facts from imperfect information in this way is the mathematical basis of Rough Set Theory (Slowinski, 1992).

Rough Set Theory is considered to be important to artificial intelligence, particularly in the fields of knowledge acquisition, decision science, machine learning, expert systems, inductive reasoning and pattern recognition. Rough Set Theory has been successfully applied in many real-life problems in finance and banking, marketing, medicine, pharmacology, engineering, environment management and others (Pawlak, 2002).

Rough Set Theory is used with the indiscernibility relation and perceptible knowledge and it can handle inexact, uncertain, and vague datasets (Walczak & Massart, 1999). This study uses the definitions of Rough Set Theory presented by Walczak and Massart (1999) as follow.

Information systems: Given an information system IS, $\mathrm{IS} = (U, A)$, where U is the universal object sets of IS; A is the model attribute sets, consisting of attributes $\{a_1, a_2, ..., a_n\}$. Each attribute $a \in A$ defines an information function $f_a : U \to V_a$, where V_a represents the domain (value sets) of attribute *a*.

Indiscernibility relation and classification: The most significant method in the ability of classification is indiscernability, which is a proper instrument for extracting and discovering facts from imperfect data. B is a subset of A, Ind(*B*) is an indiscernibility, and can be defined as: two objects, x_i and x_j, are indiscernible by the set of attributes *B* in *A*, if $b(x_i) = b(x_j)$ for every $b \subset B$. The equivalence class of Ind(*B*) is called elementary set in IS. Any x_i of U can be induced so that the value sets of attributes represented in B are in the same class. A set stands for the smallest discernible groups of objects.

Approximation accuracy and Attribute dependence: Let X be U's subset, that is, $X \subset U$. Let B be a subset of V_a, $B \subseteq V_a$. On equation (1), $\underline{BX}$ represents the lower approximation of *B*, object x_i belongs to the elementary sets contained in X.

$$\underline{BX} = \left\{ x_i \in U \,\middle|\, [x_i]_{Ind(B)} \subset X \right\} \quad (1)$$

On equation (2), BX represents the upper approximation of *B*, object x_i may, or may not, belong to the elementary sets contained in *X* that have non-empty intersections.

$$BX = \left\{ x_i \in U \,\middle|\, [x_i]_{Ind(B)} \cap X \neq \varnothing \right\} \quad (2)$$

On equation (3), BNX is called the boundary region of X, demonstrating that the objects are inconsistent or vague.

$$BNX = BX - \underline{BX} \quad (3)$$

According to Pawlak (1984), the attribute dependence can be defined as: the set of attributes is said to be independent if for every attribute, a_i, its removal increases the number of elementary sets in the IS.

Therefore, we can induce $Ind(A) = Ind(A - a_i)$, where a_i is a superfluous attribute. Otherwise, the attribute a_i is indispensable in *A*.

Reduct and core attribute sets: Two fundamental concepts of the Rough Set Theory are the concepts of core and reduct. Reducts are the minimum subsets, which are the clear-cut way of discerning object classes, provided that the object classification is the same number of elementary sets as the whole set of attributes. The core is the common part of all reducts. The reduct is the necessary part of an IS, which can distinguish all objects discernible by the original IS. The reduct attribute

sets are created to get rid of the superfluous attributes, so that the set of attributes is dependent. Although it is possible having more than one reduct attribute set in an IS, crisscrossing a number of reduct attribute sets generates a core attribute set. The process of decision-making is affected by the reduct attribute set, but the most important attribute in decision-making is the core attribute.

$$\mathrm{RED}(B) \subseteq A \quad (4)$$

$$\mathrm{COR}(C) = \cap \mathrm{RED}(B) \quad (5)$$

Equation (4) and equation (5) show the creation of the reduct attribute sets and the core attribute set based on the approximation method. On equation (4), B is the reduct attribute set, and A is the attribute set of U, so that B is included in, or equal to, A. On equation (5), the core attribute set is the intersection of all reduct attribute sets.

Decision rules: Decision rules can also be regarded as a set of decision (classification) rules of the form (Walczak and Massart, 1999): $a_{k_i} = d_j$

where a_{k_i} denotes attribute a_k has value *i*, d_j denotes the decision attributes and the symbol '$\Rightarrow$' denotes the propositional implication. In decision rule $\Phi \Rightarrow \Psi$, formulas Φ and Ψ are called the condition and decision, respectively (Pawlak, 2002). We can minimize the set of attributes, and reduct the superfluous attributes and group elements into different groups by way of the decision rules. There may have many decision rules, but a stronger rule will cover more objects as well as has shorter

restriction sets and less justifications.

The $\mathrm{supp}_S(\Phi,\Psi)$ is called the support of the rule $\Phi \to \Psi$ in IS and $\mathrm{card}(U)$ is the cardinal set that is the number of objects contained in the U (Pawlak, 2002).

$$\sigma_S(\Phi,\Psi) = \mathrm{supp}_S(\Phi,\Psi) / \mathrm{card}(U) \quad (6)$$

is the strength of the decision rule $\Phi \to \Psi$ in S.

$$cov_S(\Phi,\Psi) = \mathrm{supp}_S(\Phi,\Psi) / \mathrm{card}(\Psi_S) \quad (7)$$

is the coverage factor of the decision rule $\Phi \to \Psi$ in S.

4. TAIEX case of Stock Market Indicators Prediction Framework

In this section, the empirical process is displayed to illustrate the application of the proposed a rule based forecasting (RBF) method for predicting future stock market fluctuations.

4.1. Problem descriptions

This research intends to establish a stock market indicators prediction framework using a multiple criteria decision making (MCDM) model consisting of the cluster analysis (CA) technique and Rough Set Theory (RST) to select the important attributes and forecast TSEC Capitalization Weighted Stock Index (TAIEX). The proposed method not only can provide decision-making rules, but also can offer alternative strategies for better stock market indicators prediction model.

4.2. Research data collection

This study utilized the Delphi algorithm obtain the consensus of opinion of a group of experts with professional knowledge of finance and specialty of investment including 3 scholars of finance, 4 managers of financial institutions and 3 financial experts from investment banks to refine the nine selected macroeconomic indicators. The indicators refined by the Delphi are exploited as the condition attributes for the

Rough Set Theory based MCDM prediction framework. The ten derived indicators are: (1) Rediscount Rate, (2) Interest Rate (One Year Interest Rate), (3) Interbank Call Loan Rates, (4) Exchange Rates between NTD and USD, (5) Wholesale Price Index (WPI), (6) Crude Oil Price, (7) Monetary Aggregate (M1B), (8) Unemployment Rates, (9) Leading Indicator.

This research simultaneously used the macroeconomic indicators of January 1999 as condition attributes and the TAIEX of February 1999 as the decision attribute to form an information system. From 1999 to 2008, there are 120 monthly data sets in the decision table. Before the dataset is fed into the proposed novel MCDM framework, the fluctuations in all the datasets will be transformed into month over month growth rates. To verify the accuracy of the prediction results, i.e. the prediction rules, the historic data of the first half of fiscal 2009 was selected. All the data sets were extracted from the database being maintained by the Taiwan Economic Journal (2009).

4.3. Cluster analysis algorithms for data reduction and classification

The historical data of the price index and economic indicators discretised from two kinds of cluster analysis, the SPSS two-step cluster analysis and *k*-means cluster analysis which is a scalable cluster analysis algorithm designed to handle very large

data sets, to classify the numbers of every attribute (SPSS, 2008). To enhance the hit rate of the Rough Set Theory based MCDM forecast mechanism, two cluster analysis techniques were introduced to first classify 9 condition attributes. The condition attributes were automatically classified into two, three or four clusters according to their data characteristics (SPSS, 2001).

Then, the *k*-means cluster analysis was proposed to determine the number of clusters and to classify the decision attributes. Cluster analysis algorithms for data reduction and classification results of this study include: 1. Attributes that are classified into two clusters consist of the Interest Rate (One year Interest Rate), the Interbank Call Loan Rates, the Crude Oil Price, the Unemployment Rates and the TAIEX. 2. Attributes that are classified into three clusters consist of the Exchange Rates (NTD/USD), the Wholesale Price Index (WPI), the Rediscount Rate, the Monetary Aggregate (M1B) and the Leading Indicator.

4.4. The Analytical Procedure of Using Rough Set Theory to Establish a Stock Market Indicators Prediction Framework

The data discretised from the cluster analysis was further analyzed using the four-step analytical procedure provided by the Rough Set Data Explorer (ROSE) system, a software that implements basic elements of RST and rule discovery

techniques which was developed by Predki et al. (1998). This procedure generated several decision rules for establishing a stock market indicators prediction framework based on Rough Set Theory.

1. Decision table formation. The attributes domain and variables' value were defined as Table 1 based on the historical data collected from 1999 to 2008. The decision table of each object and condition variables is shown in Table 2.
2. Approximation computation. The approximations of the decision classes were calculated and the lower and upper approximations are shown in Table 3. The accuracy of approximation for the decision class 1 was 0.7059. The accuracy of approximation for the decision class 2 was 0.7222. The accuracy of approximation for the overall classification is 0.7143 while the overall quality of approximation for the overall classification is 0.8333. According to Shuai and Li (2005), if the accuracy is higher, the classification is less ambiguous; if the quality is better, the classification is better. Therefore, the classification is unambiguous and acceptable.
3. The reducts of attributes and the core of attributes finding. The reduct process of condition attributes utilizes the discernibility matrix to establish the superfluous attributes and to produce the reduct attribute sets in the decision table. There are no superfluous attributes and only one reduct attribute and

eight cores of attributes. The core set is the same as the reduct set, which is $\{c_1, c_3, c_4, c_5, c_6, c_7, c_8, c_9\}$.

4. The decision rules construction. After the reducts of attributes and the core of attributes finding, the minimal covering method, which is employed to produce a set of decision rules based on minimal covering algorithm (ROSE2, 1999), is utilized to discover the minimum number of attribute values for a decision rule. Consequently, the reduct with 53 rules are derived and demonstrated in Table 4.

4.5. The Stock Market Indicators Prediction Framework Verification

In order to exhibit that this empirical study is practical, the first half of fiscal 2009 of the validation sample data were put in to manage the decision rule hit test following the study validate the viability of the decision rules in this study. The decision table based on the first half of fiscal year 2009 is demonstrated in Table 5 to verify the effectiveness of the proposed stock market indicators prediction framework. Using the reduct to establish the stock market indicators prediction framework, 6 objects hit these rules. The results in Table 6 show that the decision rules can correctly predict 4 out of 6 objects from the decision rules. Therefore, the hit rate of the reduct is 66.67%.

4.6. Discussion

In the above empirical process, the two-step cluster analysis were employed to classify the condition attributes and the decision attribute. Before using the framework with these two kinds of cluster analysis, this study tried to particularly divide the dataset of each condition as well as decision attributes into three clusters by the *k*-means cluster analysis technique. Nevertheless, the actual estimate results obtained by the MCDM prediction model consisting of the *k*-means cluster analysis technique as well as Rough Set Theory were not good enough. As a result, this study decided to exploit the SPSS two-step cluster analysis methods in addition for the improvement of hit rate.

The framework with the two-step cluster analysis technique accomplished a better hit rate in this study, for that reason, the two-step cluster analysis-based MCDM prediction model could be more appropriate than the *k*-means cluster analysis-based MCDM forecast mechanism. The k-means cluster analysis necessitates specification of the number of clusters in advance, and it does not change during the iteration. However, the two-step cluster analysis possibly provides the ability to robotically find the optimal number of clusters (SPSS, 2001).

This study selected 8 macroeconomic indicators from 9 attributes: Rediscount Rate, Interbank Call Loan Rates, Exchange Rates (NTD/USD), Wholesale Price Index (WPI), Crude Oil Price, Monetary Aggregate (M1B), Unemployment Rates, and Leading Indicator.

The superfluous attributes, Interest Rate (One Year Interest Rate) were eliminated.

The total accuracy of the prediction model is 71.43% while the total quality is 83.33%. Both high accuracy and high quality correspond to low vagueness in addition to the high correctness of the generated rules. For the generated rules, rule 1 to rule 20 can be used to predict the growth of the TAIEX; rule 21 to 46 can be used to predict the decline of TAIEX; and rule 47 to rule 53 represents the approximate rule which implies the rule overlaps more than one decision class.

For the expediency of calculation and rule deductions by Rough Set Theory, decision makers can comprehend the proper timing of buying or selling shares of stock without difficulty. Taking the decision rule 1(see Table 4) as an example, the decision rule 1 means if the NT Dollar (NTD/USD) is appreciated ($c4=3$), the Crude Oil Price doesn't change or falls ($c6=2$), the Unemployment Rates rises ($c8=1$), then TAIEX probably rises ($d=1$) in this month. If the set of macroeconomic indicators in the prior month hit only decision rule 1, then, decision makers could count on the obvious and comprehensible information to increase their shares of stock.

In this paper, the most sustained rules include rule 25 (strength = 5.83 %), rule 40 (strength = 5.83 %) and rule 44 (strength = 5.83 %). As of rule 25, when the NT Dollar (NTD/USD) doesn't change, the Unemployment Rates rises, and the Leading Indicator falls, then the TAIEX will fall. Based on rule 40, when the Wholesale Price Index (WPI) rises, and

the Monetary Aggregate (M1B) doesn't change or falls, then the TAIEX will fall. Anchored in rule 44, when the Crude Oil Price doesn't change or falls, the Unemployment Rates rises, and the Leading Indicator falls, then the TAIEX will fall.

The rule with the highest strength is only 5.83 %. In addition, many rules maintain just one or two objects, implying the highly uncertain nature of the rule. As a result, all rules have a low strength rate. Nonetheless, Rough Set Theory can provide rules which cover only subsets of the basic objects or data records available (Curry, 2003). Given that, Rough Set Theory is certainly right and proper to obtain the stock market prediction rules. Therefore, we utilize the proposed stock market indicators prediction model to predict data in the first six months of year 2009, and the hit rate is 66.67%. The high hit rate (greater than 60%) means that the stock market prediction rules and indicators provided in this study are practical to predict stock price index fluctuations.

5. Conclusions and Remarks

This study demonstrates a stock market indicators prediction framework using a multiple criteria decision making model (MCDM) consisting of the cluster analysis technique and Rough Set Theory to select the important attributes and forecast TSEC Capitalization Weighted Stock Index (TAIEX). Nine indicators were introduced while eight indicators were selected as the condition attributes to be fed into the Rough Set Theory-based MCDM prediction framework. From the research results, fifty three rules were finally derived by Rough Set Theory for the TAIEX fluctuation predictions. The method not only can provide decision-making rules, but also can offer alternative strategies for better stock market indicators prediction model. The proposed prediction model was leveraged to predict the index in first half year of 2009 with an accuracy of 66.67%. The results indicate the decision rules were authenticated to employ in predicting the stock market fluctuations appropriately.

Tables

Table 1. Definition of Variables

Attribute/Descriptions	Variables Val
C1: Rediscount Rate	Code 1 if rate rises; code 2 if rate doesn't change; code 3 if rate f
C2: Interest Rate (One Year Interest Rate)	Code 1 if rate rises or doesn't change; code 2 if rate high falls
C3: Interbank Call Loan Rates	Code 1 if rate high rises or high falls; code 2 if rate low rises or (
C4: Exchange Rates (NTD/USD)	Code 1 if NT Dollar depreciation; code 2 if NT Dollar doesn't c
C5: Wholesale Price Index (WPI)	Code 1 if price rises; code 2 if price doesn't change; code 3 if pri
C6: Crude Oil Price	Code 1 if price rises; code 2 if price doesn't change or falls
C7: Monetary Aggregate (M1B)	Code 1 if M1B high rises; code 2 if M1B low rises ; code 3 if M
C8: Unemployment Rates	Code 1 if rate rises; code 2 if rate falls
C9: Leading Indicator	Code 1 if indicator high rises; code 2 if indicator low rises or do
D: TSEC Capitalization Weighted Stock Index	Code 1 if index rises ; code 2 if index falls

Table 2. The Decision Table

	Condition Variables									
Objects	C1	C2	C3	C4	C5	C6	C7	C8	C9	DV
1	2	1	1	2	2	2	3	2	1	2
2	3	2	1	1	2	1	2	2	1	1
3	2	1	1	2	2	2	2	1	1	1
4	2	1	1	3	2	2	2	2	1	1
5	2	1	1	2	2	1	2	1	1	2
6	2	1	1	3	2	2	1	1	2	1
7	2	1	1	2	2	2	1	1	2	2
8	2	1	1	3	2	2	1	1	2	1
9	2	1	1	2	1	2	1	2	2	2
10	2	1	1	2	1	1	1	2	2	1
11	2	1	1	2	1	2	1	2	2	2
12	2	1	1	3	2	2	1	2	2	1
13	2	1	1	3	2	1	1	2	1	1
14	2	1	1	2	2	2	1	1	2	2
15	2	1	1	3	2	1	1	2	2	1
16	2	1	1	2	2	1	1	2	2	2
17	2	1	1	2	1	2	1	1	2	1
18	2	1	1	2	2	2	2	1	2	2
19	2	1	1	1	2	1	2	1	3	2
20	2	1	1	2	2	2	2	1	3	2
21	2	1	1	1	2	2	2	2	3	2
22	2	1	1	1	1	1	2	1	3	2
23	2	1	1	1	1	2	2	1	3	2
24	2	1	1	2	2	1	3	1	3	2
25	2	1	1	3	2	1	3	1	3	1
26	3	1	1	2	2	2	3	1	3	2
27	3	1	1	1	2	1	3	1	3	1
28	2	2	1	2	2	2	3	1	3	2
29	3	1	1	1	2	2	3	1	3	2
30	3	1	1	1	1	1	3	1	3	2
31	2	1	1	1	2	1	3	1	2	2
32	3	2	1	2	2	2	3	1	2	1
33	3	2	1	2	2	1	2	1	2	2
34	3	2	2	2	2	1	2	1	1	1
35	3	2	1	2	2	1	2	2	1	1
36	3	1	1	1	2	1	1	2	1	1
37	2	1	1	2	1	2	2	2	1	1
38	2	1	1	2	1	2	1	2	1	2
39	2	1	1	2	1	2	1	1	1	1
40	2	1	1	3	1	2	1	2	1	2
41	2	1	1	3	2	1	1	1	2	2
42	3	1	1	3	2	1	1	1	2	2
43	2	2	1	2	2	2	1	1	2	2
44	2	1	1	1	2	2	1	1	2	2
45	2	1	1	1	1	2	1	2	2	2
46	2	1	1	2	1	1	1	2	2	1
47	3	2	1	2	1	1	1	2	2	1
48	2	1	2	2	1	2	2	2	2	2
49	2	2	2	2	2	2	1	2	2	1
50	2	1	1	2	1	2	2	1	2	2
51	2	1	1	2	2	1	2	2	2	2
52	2	1	1	2	3	1	2	2	2	2
53	2	1	1	2	2	2	2	1	1	1
54	3	1	1	2	2	2	2	1	1	1
55	2	2	2	2	2	2	1	1	1	1
56	2	1	1	3	2	2	1	1	1	1
57	2	1	1	3	2	1	1	2	1	2
58	2	1	1	2	2	2	1	2	1	1
59	2	1	1	2	1	1	1	2	1	2
60	2	1	1	2	1	2	1	2	1	1

	Condition Variables									
Objects	C1	C2	C3	C4	C5	C6	C7	C8	C9	DV
61	2	1	1	3	1	2	1	2	2	1
62	2	1	1	2	1	1	1	1	2	1
63	2	1	1	3	1	2	1	2	2	2
64	2	1	1	1	2	1	1	2	2	2
65	2	1	1	2	1	2	1	1	2	2
66	2	1	1	1	2	1	1	1	2	2
67	2	1	1	1	1	2	1	1	2	2
68	2	1	1	2	1	2	1	1	2	1
69	2	1	1	2	2	2	1	2	2	1
70	1	1	1	3	2	2	1	2	2	2
71	2	1	1	3	2	1	1	2	2	1
72	1	1	1	3	2	1	1	2	2	1
73	2	1	1	2	2	2	2	2	2	2
74	2	1	1	3	2	2	2	1	2	1
75	1	1	1	1	2	2	2	2	2	2
76	2	1	1	3	2	1	2	2	2	2
77	2	1	1	2	3	1	2	1	2	1
78	2	1	1	1	2	2	2	1	2	1
79	1	1	1	1	2	2	2	1	2	1
80	2	1	1	1	2	2	2	1	1	2
81	1	1	1	1	1	1	2	2	2	1
82	2	1	1	1	1	1	2	2	2	2
83	2	1	1	2	2	1	2	2	2	1
84	1	1	1	3	2	2	2	2	2	1
85	2	1	1	3	2	2	2	2	2	2
86	2	1	1	1	1	1	2	1	2	1
87	1	1	1	2	2	2	2	2	2	1
88	2	1	1	3	1	2	2	2	2	1
89	2	1	1	2	1	1	2	1	2	2
90	1	1	1	1	1	1	2	1	2	2
91	2	1	1	1	1	2	3	1	2	2
92	2	1	1	2	2	1	2	1	2	1
93	1	1	1	2	2	1	2	2	2	1
94	2	1	1	2	2	1	2	2	2	1
95	2	1	1	3	1	2	2	2	2	1
96	1	1	1	2	1	2	2	2	2	1
97	2	1	1	1	2	1	2	2	2	2
98	2	1	1	2	1	2	2	2	2	1
99	1	1	1	2	1	2	2	1	2	2
100	2	1	1	2	1	2	2	2	2	2
101	2	1	2	3	2	1	2	1	2	1
102	1	1	2	3	2	2	2	1	2	1
103	2	1	2	2	2	2	2	1	2	1
104	2	1	1	2	2	1	2	1	2	2
105	1	1	1	3	1	2	2	2	2	1
106	2	1	1	2	1	2	2	2	2	1
107	2	1	1	2	1	2	2	2	2	2
108	1	1	1	2	1	1	3	2	2	2
109	2	1	1	3	1	2	3	2	2	2
110	2	1	1	3	2	2	3	1	2	1
111	1	1	1	3	2	2	3	2	3	1
112	2	1	1	2	2	2	3	2	3	1
113	2	1	1	2	1	2	3	1	3	2
114	1	1	1	2	1	2	3	1	3	2
115	2	1	1	2	1	1	3	1	3	2
116	2	1	1	1	3	1	3	1	3	1
117	2	1	1	1	3	1	3	1	3	2
118	3	2	1	1	3	1	3	1	3	2
119	3	2	2	1	3	1	3	1	3	2
120	3	2	2	3	3	1	3	1	3	1

Remark: Decision variable is abbreviated as DV

Table 3. The Lower and Upper Approximations

Class Number	Number of Objects	Lower Approximatio	Upper Approximatio	Accuracy
1	59	48	68	0.7059
2	61	55	72	0.7222

Table 4. Rules of the Reduct

Rule	C1	C3	C4	C5	C6	C7	C8	C9	DV	Strength(%)
1			3		2		1		1	5
2			3		1	1	2	2	1	2.5
3	1		3			2			1	2.5
4				1	1	1		2	1	3.33
5			2		2	2		1	1	3.33
6	2			2	2	1	2		1	3.33
7	1		2			2	2		1	2.50
8	3							1	1	4.17
9						3	2	3	1	1.67
10			1		2	2	1	2	1	1.67
11				3			1	2	1	0.83
12		2		2					1	5.00
13			3	1		2			1	2.50
14	3		1	2	1				1	2.50
15				1			1	1	1	0.83
16	1			1		2	2		1	2.50
17			3					3	1	2.50
18			3			2		1	1	0.83
19	3					3		2	1	0.83
20	2		1			2	1	2	1	1.67
21	2	1	2	2	2		1	2	2	3.33
22	2		1				2		2	4.17
23	2		1	2	1				2	4.17
24			2	1	2	1	2	2	2	1.67
25			2				1	3	2	5.83
26			2	1		2	1		2	2.50
27	2			2	2	2	2	2	2	1.67
28	3		1	3					2	1.67
29				2	1	1	1		2	2.50
30			2		1	1		1	2	0.83
31	2		1					1	2	0.83
32				3			2		2	0.83
33			1		2		2		2	2.50
34	1			1			1		2	2.50
35	1				2	1			2	0.83
36			3	1				1	2	0.83
37	3	1	2		1		1		2	0.83
38	2				1	2		1	2	0.83
39	2		3	2		2	2	2	2	1.67
40				1		3			2	5.83
41			1			1		2	2	4.17
42						3		1	2	0.83
43		2		1					2	0.83
44					2		1	3	2	5.83
45				1				3	2	5.00
46			2	2	1	1			2	0.83
47	2	1	2	1		2	2	2	1 or 2	3.33
48	2		2	2	1	2		2	1 or 2	4.17
49			2	1	2	1	1	2	1 or 2	2.50
50			3		1			1	1 or 2	1.67
51			2	1	2	1	2	1	1 or 2	1.67
52	2			3				3	1 or 2	1.67
53			3	1		1		2	1 or 2	1.67

Table 5. The Decision Table (First Half of Fiscal Year 2009)

	Condition Variables									
Objects	C1	C2	C3	C4	C5	C6	C7	C8	C9	DV
1	3	2	2	1	2	2	2	1	2	2
2	3	2	2	1	1	1	2	1	1	1
3	2	1	1	3	2	2	2	1	1	1
4	2	1	1	3	2	2	2	2	1	1
5	2	1	2	3	2	2	1	1	1	1
6	2	1	1	2	1	2	1	1	1	2

Remark: Decision variable is abbreviated as DV

Table 6. The Reduct Verified Decision Table

	Condition Variables									
Objects	C1	C2	C3	C4	C5	C7	C8	C9	DV	Hit
1	2	2	2	2	1	2	3	3	2	X
2	2	2	2	3	4	2	2	3	2	V
3	1	2	2	3	4	2	3	3	2	V
4	2	2	2	2	2	2	3	3	3	V
5	2	2	2	2	1	2	2	4	3	V
6	1	2	2	2	1	2	2	4	3	X

Remark: (1) Decision variable is abbreviated as DV

(2) V: hit; X: not hit

References

Balduzzi, P. (1995), "Stock returns, inflation, and the 'proxy hypothesis': A new look at the data", Economics Letters, 48(1), 47-53.

Bilson, C. M., T. J. Brailsford, and V. J. Hooper (2001), "Selecting macroeconomic variables as explanatory factors of emerging stock market returns", Pacific-Basin Finance Journal, 9(4), 401-426.

Bollerslev T. (1986), "Generalized autoregressive conditional heteroscedasticity", Journal of Econometrics 31 (3) 307–327.

Box, G. and G. Jenkins (1976), "Time series analysis: Forecasting and control" (Holden-Day, San Francisco,).

Carl, K.C., D.K. Gary, B.E. Joel, and B.S. Samuel (2008), "The Delphi process in dental research", The Journal of Evidence-based Dental Practice, 8 (4) Carl, 211-220.

Chen, N. F., R. Roll, and S. A. Ross (1986), "Economic forces and the stock market", Journal of Business, 59(3), 383-403.

Cheng, C.H. and L.Y. Wei (2009), "Volatility model based on multi-stock index for TAIEX forecasting", Expert Systems with Applications 36 (3) 6187–6191.

Curry, B. (2003), "Rough sets: current and future developments", Expert Systems 20 (5)

247-250.

Dalkey, N. C., & Helmer, O. (1963), “An experimental application of the Delphi method to the use of experts”, Management Science, 9, 458–467.

Engle, R.F. (1982), “Autoregressive conditional heteroscedasticity with estimator of the variance of United Kingdom inflation”, Econometrica 50 (4) 987-1008.

Fama, E. F. (1981), “Stock returns, real activity, inflation, and money”, American Economic Review, 71(4), 545-565.

Fama, E. F. (1990), “Stock returns, expected returns and real activity”, Journal of Finance, 45(4), 1089-1108.

Ferson, W. E. and C. R. Harvey (1998), “Fundamental determinants of national equity market returns: A perspective on conditional asset pricing”, Journal of Banking and Finance, 21(11-12), 1625-1665.

Fink, A., J. Kosecoff, M. Chassin, and R. Brook, “Consensus methods: characteristics and guidelines for use”, American Journal of Public Health, 74 (9) (1984) 979-983.

Gultekin, N. B. (1983), ”Stock market returns and inflation: Evidence from other countries”, Journal of Finance, 38(1), 49-65.

Hartigan, J.(1975), “Clustering Algorithms”, John Wiley & Sons, New York.

Huang, J.Y., Y.B. Qiu, X.P. Guo, Cluster and discriminant analysis of electrochemical noise

statistical parameters, Electrochimica Acta 54 (8) (2009) 2218-2223.

Huarng, K. and H.K. Yu (2006), The application of neural networks to forecast fuzzy time series, Physica A 363 (2) 481-491.

Johnson,S.C.(1967), “Hierarchical clustering schemes”, Psychometrika, 32 (3) 241-254.

Johnson,R.A. and D. Wichern (2007), “Applied Multivariate Statistical Analysis”, Person Education, Inc..

Kim, M. K. and C. Wu (1987), “Macro-economic factors and stock returns”, *Journal of Financial Research*, 10(2), 87-98.

Kimoto, T., K. Asakawa, M. Yoda, and M. Takeoka (1990), “Stock market prediction system with modular neural network”, Proceedings of the international joint conference on neural networks, San Diego, California, 1-6.

Kwon, C. S. and T. S. Shin (1999), “Cointegration and causality between macroeconomic variables and stock market returns”, Global Finance Journal, 10(1), 71-81.

Mandelker, G. and Tandon, K. (1985), “Common stock returns, real activity, money, and inflation: Some international evidence”. Journal of International Money and Finance, 4(2), 267-286.

Massart, D.L., L. Kaufman, P.J. Elving, and J.D. Winefordner (1983), Chemical Analysis

Series, 65,John Wiley & Sons, New York.

McMillan, D.G. (2007), "Non-linear forecasting of stock returns: does volume help?" International Journal of Forecasting 23 (1) 115-126.

Milholland, A., S.G. Wheeler, and J.J. Heieck (1973), "Medical assessment by a Delphi group opinion technique", The New England Journal of Medicine, 188 (24) 1272-1275.

J. P. Morgan, Reuters (1996), RiskMetrics – Technical Document (New York).

Nelson, D.B. (1991), "Conditional heterosdasticity in asset returns: a new approach", Econometrica 59 (2) 347-370.

Nikolopoulos, C. and P. Fellrath (1994), "A hybrid expert system for investment advising", Expert Systems 11 (4) 245-250.

Pawlak, Z. (2004). Decision networks. Rough Sets and Current Trends in Computing by Shusaku Tsumoto, Roman Slowinski, Jan Komoroski, Jerzy W. Grzymala-Busse (Eds.), Lecture Notes in Artificial Intelligence (LNAI), 3066(1), 1–7.

Pawlak Z. (2002), Rough sets and intelligent data analysis, Information Sciences 147 (1-4) 1-12.

Pawlak, Z. (1984). Rough classification. International Journal of Man–Machine Studies, 20(5), 469–483.

Pawlak Z. (1982), "Rough sets", International Journal of Computer and Information

Sciences 11 (5) 341-356.

Pill, J. (1971), "The Delphi method: substance, context, a critique and an annotated bibliography", Socio-Economic Planning Sciences, 5 (1) 57-71.

Predki, B., R. Slowinski, J. Stefanowski, R. Susmaga, and S. Wilk, ROSE (1998), "software implementation of the rough set theory", Lecture Notes in Computer Science 1424 605-608.

Robichek, A. A. and Cohn, R. A. (1974). "The economic determinants of systematic risk", *Journal of Finance*, 29(2), 439-447.

Rogalski, R. J. and Vinso, J. D. (1977), "Stock returns, money supply and the direction of causality", *Journal of Finance*, 32(4), 1017-1030.

Roh,T.H. (2007), "Forecasting the volatility of stock price index", Expert Systems with Applications 33 (4) 916-922.

Roll, R. (1992), "Industrial structure and the comparative behavior of international stock market indices", *Journal of Finance*, 47(1), 3-41.

ProSoft, ROSE (Rough Set Data Explorer) Version 2.0 User's Guide, 1999, http://wwwidss.cs.put.poznan.pl/software/rose/main.html.

Shen, L. and H.T. Loh (2004), "Applying rough sets to market timing decisions", Decision Support Systems 37 (4) 583–597.

Shuai, J.J. and H.L. Li (2005), Using rough set and worst practice DEA in business

failure prediction, Lecture Notes in Computer Science 3462 503-510.

Slowinski R. (1992), “Intelligent Decision Support. Handbook of Applications and Advances of the Rough Sets Theory”, Kluwer Academic Publishers, Dordrecht.

SPSS Inc. (2008), SPSS (Statistical Package for the Social Sciences) version 16.0.

SPSS Inc. (2001), “The SPSS Two Step Cluster Component”.

Sumsion, T. (1998), “The Delphi technique: an adaptive research tool”, The British Journal of Occupational Therapy, 61 (4) 153-156.

Susmaga, R., W. Michalowski, and R. Slowinski (1997), “Identifying regularities in stock portfolio tilting”, International Institute for Applied Systems Analysis Interim Report IR 97-066 1-21.

Taiwan Economic Journal (TEJ), Taiwan Economic Journal Profile database (2009). http://www.tej.com.tw/twsite/.

Teoh, H.J., T.L. Chen, C.H. Cheng, and H.H. Chu (2008), “A hybrid multi-order fuzzy time series for forecasting stock markets”, Expert Systems with Applications 36 (4) 7888-7897.

Tseng, F.M., G.H. Tzeng, H.C. Yu, and J.C. Yuan (2001), “Fuzzy ARIMA model for forecasting the foreign exchange market”, Fuzzy Sets and Systems 118 (1) 9–19.

Walczak, B. and D.L. Massart (1999), “Rough set theory”, Chemometrics and

Intelligent Laboratory Systems 47 (1) 1-16.

Ward, J. H. (1963), “Hierarchical grouping to optimize an objective function”, Journal of the American Statistical Association, 58, 236-244.

Whitman, N. (1990), “The committee meeting alternative using the Delphi technique”, Journal of Nursing Administration, 20 (7-8) 30-36.

Yao, J.T. and J.P. Herbert (2009), “Financial time-series analysis with rough sets”, Applied Soft Computing 9 (3) 1000-1007.