Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 35 (2008) 19–29
Expert Systems with Applications
www.elsevier.com/locate/eswa

Mining stock category association and cluster on Taiwan stock market

Shu-Hsien Liao *, Hsu-hui Ho, Hui-wen Lin

Department of Management Sciences and Decision Making, Tamkang University, No. 151, Yingjuan Road, Danshuei Jen, Taipei 251, Taiwan, ROC

Abstract

One of the most important problems in modern finance is finding efficient ways to summarize and visualize the stock market data to give individuals or institutions useful information about the market behavior for investment decisions. The enormous amount of valuable data generated by the stock market has attracted researchers to explore this problem domain using different methodologies. This paper investigates stock market investment issues on Taiwan stock market using a two-stage data mining approach. The first stage Apriori algorithm is a methodology of association rules, which is implemented to mine knowledge and illustrate knowledge patterns and rules in order to propose stock category association and possible stock category investment collections. Then the *K*-means algorithm is a methodology of cluster analysis implemented to explore the stock cluster in order to mine stock category clusters for investment information. By doing so, this paper proposes several possible Taiwan stock market portfolio alternatives under different circumstances.

Keywords: Data mining; Association rule; Cluster analysis; Stock market analysis; Stock portfolio

1. Introduction

The stock market is one of the most popular forms of investment due to its high-expected profit. However, higher expected profit, also imply higher risk. Thus, numerous studies have proposed different analysis methods to assist investors in analysis and decision-making. On the other hand, many individual investors, stockbrokers, and financial analysts attempt to predict stock market price activities and their potential development. This mass behavior runs counter to the counsel of the many academic studies, which contend that the prediction of stock market development is ineffective. This point of view is codified as the generally called efficient markets hypothesis (Fama, 1991; Haugen, 1997).

There are three degrees of market efficiency. The first degree is the strong form of the efficient markets hypothesis, which states that all information that is knowable is immediately factored into the market's price for a security. If this is true, then all of those price predictors are definitely wasting their time, even if they have access to private information. The second degree is the semi-strong form of the efficient markets hypothesis, that all public information is considered to have been possessors of private information, which can use that information for profit. The third degree is the weak form, which holds only that any information gained from examining the security's past trading history is reflected in price. Indeed, the past trading history is public information implying that the weak form is a specialization of the semi-strong form, which itself is a specialization of the strong form of the efficient market hypothesis.

Due to the different degrees of market efficiency, academic researchers investigate the efficient market hypothesis by exploring the unknown and valuable knowledge from historical data, using techniques such as data mining. Enke and Thawornwong (2005) introduces an information gain technique used in machine learning for data mining to evaluate the predictive relationships of numerous financial and economic variables. Neural network models for level estimation and classification are then examined for their ability to provide an effective forecast of future values. Boginski, Butenko, and Pardalos (2006) propose a network representation of the stock market data referred to as the

[*] Corresponding author.
E-mail address: michael@mail.tku.edu.tw (S.-H. Liao).

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.06.001

market graph, which is constructed by calculating cross-correlations between pairs of stocks based on the opening price data over a certain period of time. Chun and Park (2005) proposes a learning technique, which extracts new case vectors using Dynamic Adaptive Ensemble CBR (DAE CBR). The main idea of DAE CBR originates from finding combinations of parameter and updating and applying an optimal CBR model to an application or domain area. These concepts are investigated against the backdrop of a practical application involving the prediction of a stock market index. In addition, Rapach and Wohar (2006) implement an analysis of in-sample and out-of-sample tests of stock return predictability in an effort to better understand the nature of the empirical evidence on return predictability. That study finds that certain financial variables display significant in-sample and out-of-sample predictive ability with respect to stock returns. Overall, most articles consider stock market analysis as a time series problem, and there have been few studies using stock market efficiency to explore the possible cause-and-effect relationships among different stock categories or the influence of outside factors.

This paper investigates stock market investment issues in the Taiwan stock market by implementing a two-stage data mining approach. First, the Apriori algorithm is a methodology of association rules that mines knowledge from historical data and this knowledge is illustrated as knowledge patterns and rules in order to propose stock category association and possible stock investment collections. Next, the *K*-means algorithm is a methodology of cluster analysis that explores the clustering of stock in order to mine this information for investment. Thus, using two different data mining approaches, this paper provides two aspects of data mining results in terms of presenting possible investment portfolio with stock market association and cluster knowledge. The rest of this paper is organized as follows. In Section 2, we describe the Taiwan stock market. Section 3 presents the research design. Section 4 introduces the proposed data mining system, which includes system framework, relational database design, and physical database design. Section 4 presents the data mining approach, including the Apriori and *K*-means algorithm. Section 5 describes the data mining results. Research findings and discussions are presented in Sections 6 and 7 presents a brief conclusion.

2. Taiwan stock market

TSEC, Taiwan Stock Exchange Corporation, maintains stock price indices to allow investors to grasp both overall market movement and different industrial sectors' performances conveniently. The indices may be grouped into market value indices and price average indices. The former are similar to the Standard and Poor's Index, weighted by the number of outstanding shares, and the latter are similar to the Dow Jones Industrial Average and the Nikkei Stock Average. The Taiwan Stock Exchange Capitalization Weighted Stock Index ("TAIEX") is the most widely quoted of all TSEC indices. The base year value as of 1966 was set at 100. TAIEX is adjusted in the event of new listings, de-listings and new share offerings to offset the influence on TAIEX owing to non-trading activities. TAIEX covers all of the listed stocks excluding preferred stocks, full-delivery stocks and newly listed stocks that have listed for less than one calendar month. The other market value indices are calculated and adjusted similarly to that of the TAIEX, but with different groupings of stocks included for calculation. Out of the TAIEX Component Stocks, the non-Finance Sub-Index, Non-Electronics Sub-Index, and Non-Finance Non-Electronics Sub-Index include stocks not in the financial sector, not in the electronics sector, and not in either sector. Similarly, the Industrial Sub-Indices are calculated for different industrial sectors. In 1986, eight Industrial Sub-Indices were introduced, i.e. Cement/Glass/Ceramics, Textiles, Foods, Plastics/Chemicals/Rubber, Electric Machinery/Electric Appliance/Cable/Electronics, Paper/Pulp, Construction, Finance. In 1995, the TSEC introduced additional 14 Industrial Sub-Indices, i.e. Cement, Plastics, Electric machinery, Electric appliance/cable, Automobile, Chemicals, Glass/ceramics, Iron/steel, Rubber, Electronics, Transportation, Tourism, Retail and others. This expansion was to give a broader perspective of industrial performance and a more comprehensive comparison with overall market trends. Total Return Indices add back cash dividends to the index calculations, and are published at the end of each trading day. This expansion can serve as a better indicator to measure the performance of funds.

The Industrial Price Average Index and the Composite Price Average Index contain 20 and 30 issues, respectively. The samples are chosen based on their representation in the market as a whole and are adjusted every year by taking considering the profitability, operational efficiency and trading liquidity of the shares, so that the indices can mirror the market trend. All of the TSEC indices (excluding Total Return Indices) are constantly computed and broadcast every minute during the trading hours through the TSEC MIS system and information vendors' networks. This information can be easily accessed on the systems of local and international information vendors, such as Reuters, Bridge, Quick, Bloomberg, Primark, etc. Monthly summaries of all the TSEC indices data are also available on TSEC website (http://www.tse.com.tw/en/).

3. Research design

3.1. Research procedure

In this article, we use the TAIEX indices, including 19 index categories and international stock indices such as the NIKKEI 225, KOSPI, Dow Jones, etc. and construct a database. This database provides the basis for data mining. Moreover, the database can be mined out of portfolio investment suggestions by using of association rule and

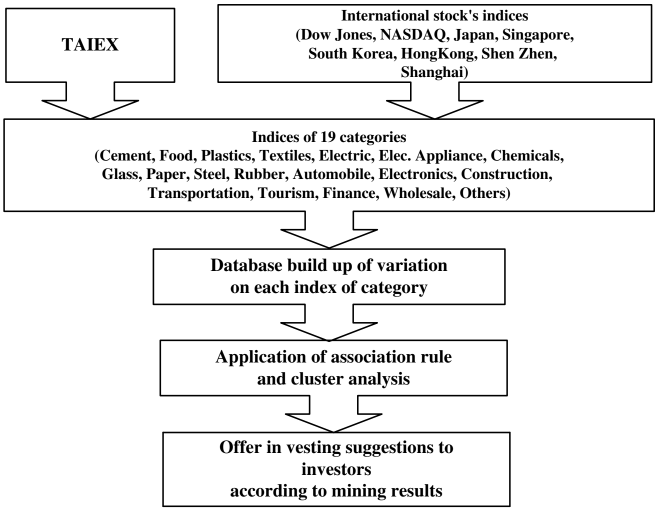
TAIEX
International stock's indices
(Dow Jones, NASDAQ, Japan, Singapore,
South Korea, HongKong, Shen Zhen,
Shanghai)
Indices of 19 categories
(Cement, Food, Plastics, Textiles, Electric, Elec. Appliance, Chemicals,
Glass, Paper, Steel, Rubber, Automobile, Electronics, Construction,
Transportation, Tourism, Finance, Wholesale, Others)
Database build up of variation
on each index of category
Application of association rule
and cluster analysis
Offer in vesting suggestions to
investors
according to mining results

Fig. 1. Research procedure.

cluster analysis in order to make suggestions to the investors. The research procedure is shown in Fig. 1.

3.2. Data

According to the classification of TSEC in the present day, there are totally 19 indices categories of stocks which involve cement, food, plastics, textiles, electric and machinery, electric appliance and cable, chemicals, glass and ceramics, paper and pulp, steel and iron, rubber, automobile, electronics, construction, transportation, tourism, finance and insurance, wholesale and retail, and others. In this study, we select the top 6 countries by their trading amount with Taiwan in 2005 and adopt their stocks' indices as international indices. The selected countries and their trading amounts are listed in Table 1.

These indices were gathered from the Taiwan Stock Exchange Council (2005) for the period from January 2000 to December 2005, a total of 1509 trading days. In addition, this article uses the subtraction of indices of two successive trading days to represent the rise or fall of the index. When the index of later trading day subtracted from the former trading day is positive (or negative), we use 1, 0 (or 0, 1) to represent this situation.

3.3. Database – the star schema

A star schema is a simple database design (particularly suited to ad hoc queries) in which dimensional data (describing how data are commonly aggregated) are separated from fact or event data (Devlin, 1997). A star schema consists of two types of tables: fact tables and dimension tables. Fact tables contain factual or quantitative data and dimension tables have descriptive data. Each dimension table has a one-to-many relationship to the central fact table. Each dimension table generally has a simple primary key, as well as several non-key attributes. In this study, the star schema is used to design the database. To build up the normalized relation is an important target of database design. Generally speaking, with higher orders of normalization, more joint operations are needed to produce a specific output. Increasing the normalization will increase the tables, which thus increases the input/output actions of the hard drive and also slows the operation speed as well. Therefore, the database allows certain degree of de-normalization in order to speed up the operation of data, but it will do so at the cost of producing duplicate data. Because the data of stock price indices is non-duplicate, this article adopts the star schema for database design and to speed up the data operation. In this article, the star schema contains the four elements of fact table, dimension table, attributes and attribute hierarchies (Fig. 2). Each element can be described as follows.

Table 1
Ranking of trading partners with Taiwan

Country	Total amount of trading	Ratio (%)	Rank
China	60,806,739,930	16.390	1
Japan	60,421,119,699	16.286	2
US	49,498,010,632	13.342	3
Hong Kong	32,607,320,243	8.789	4
South Korea	18,778,408,285	5.062	5
Singapore	12,597,268,828	3.396	6

(1) *Fact table*: In general, facts are stored in a fact table, which linked to *n*-dimension entities. The primary key of the fact table must be comprised by foreign keys, which link to the relative dimension table. In this study, the fact tables include indices of Taiwan's stock market and international stock markets.
(2) *Dimension table*: A dimension table can provide additional view points for given facts. The data of decision support system are always checked for their associations with other data. In this study, the dimensions include the TAIEX, indices of international stock markets, date and the format of data appearance.
(3) *Attribute*: Attributes are used to search, filter or classify the facts and attributes are always contained in each dimension. Dimensions acquire the descriptive characteristics of facts via attributes. This study uses 19 categories of stock indices from Taiwan and eight international stock indices as attributes.
(4) *Attribute hierarchies*: Attributes in the dimension table can be ordered by using attribute hierarchies that have been defined specifically. Attribute hierarchies can be used to operate the drill down or roll up in data analysis.

4. Data mining

4.1. Association rules

As stated in Agrawal, Imilienski, and Swami (1993), discovering association rules is an important data mining problem, and there has been considerable research on using

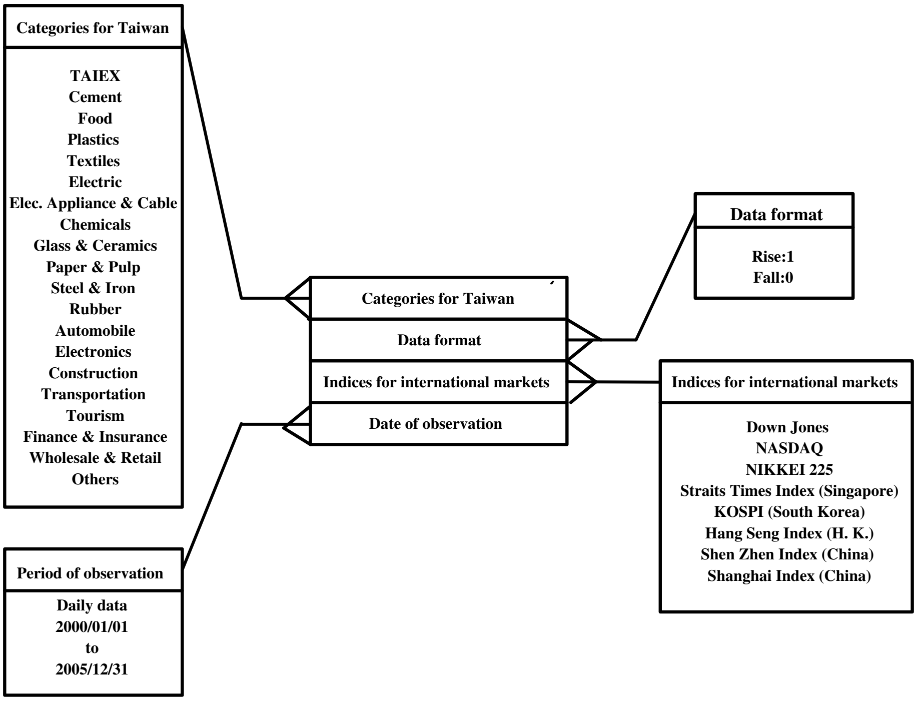

Fig. 2. Database – the star schema.

association rules in the field of data mining problems. The association rules algorithm is used mainly to determine the relationships between items or features that occur synchronously in the database. For instance, if people who buy item X also buy item Y, there is a relationship between item X and item Y, and this information is useful for decision makers. Therefore, the main purpose of implementing the association rules algorithm is to find synchronous relationships by analyzing the random data and to use these relationships as a reference during decision-making. The association rules are defined as follows (Wang, Chuang, Hsu, & Keh, 2004):

Make $I = \{i_1, i_2, \ldots, i_m\}$ as the item set, in which each item represents a specific literal. D stands for a set of transactions in a database in which each transaction T represents an item set such that $T \subseteq I$. That is, each item set T is a non-empty sub-item set of I. The *association rules* are an implication of the form $X \to Y$, where $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$. The rule $X \to Y$ holds in the transaction set D according to two measure standards – *support* and *confidence*. Support (denoted as $Sup(X,D)$) to represent the rate of transactions in D containing the item set X. *Support* is used to evaluate the statistical importance of D, and the higher its value, the more important the transaction set D is. Therefore, the rule $X \to Y$ has *support* $Sup(X \cup Y, D)$ represents the rate of transactions in D containing $X \cup Y$. Each rule $X \to Y$ also has the other measuring standard called Confidence (denoted as $Conf(X \to Y)$), representing the rate of transactions in D that contain X and also Y. That is, $Conf(X \to Y) = Sup(X \cap Y)/Sup(X,D)$.

In this case, $Conf(X \to Y)$ denotes that if the transaction includes X, the chance that transaction also contains Y is relatively high. The measure confidence is then used to evaluate the level of confidence about the association rules $X \to Y$. Given a set of transactions D, the problem of mining association rules is to generate all transaction rules that have certain user-specified minimum support (called *Min*sup) and confidence (called *Minconf*) (Kouris, Makris, & Tsakalidis, 2005; Padmanabhan & Tuzhilin, 2002). According to Agrawal and Shafer (1996), the problem of mining association rules can be decomposed into two steps. The first step is to detect a large item set whose support is greater than *Min*sup and the second step is to generate association rules, using the large item set. Such rules must satisfy two conditions:

1. $Sup(X \cup Y, D) \geqslant Min\text{sup}$
2. $Conf(X \to Y) \geqslant Minconf$

To explore the association rules, many researchers use the Apriori algorithm (Agrawal et al., 1993). In order to reduce the possible biases incurred when using these measure standards, the simplest way to judge the standard is to use the *lift* judgment. *Lift* is defined as: *Lift* = *Confidence*$(X \to Y)/Sup(Y)$ (Wang et al., 2004).

4.2. Cluster analysis and K-means algorithm

The process of partitioning a large set of patterns into disjoint and homogeneous clusters is fundamental in knowledge acquisition. It is called *Clustering* in the literature and it is applied in various fields, including data mining, statistical data analysis, compression and vector quantization. The *k-means* is a very popular algorithm and is one of the best for implementing the clustering process. K-means clustering proceeds in the following order. Firstly, K number of observations is randomly selected from all N number of observations, according to the number of clusters, and these become centers of the initial clusters. Secondly, for each of the remaining $N-K$ observations, find the nearest cluster in terms of the Euclidean distance with respect to $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip}, \ldots, x_{iP})$ is found. After each observation is assigned the nearest cluster, the center of the cluster is re-computed. Lastly, after the allocation of all observation, the Euclidean distance between each observation and cluster's center point is calculated to confirm whether or not it is allocated to the nearest cluster. In addition, implementation of the K-means algorithm implementing for cluster analysis as data mining approach has been discussed in several studies (Kuo, Liao, & Tu, 2005; Ture, Kurt, Turhan, & Ozdamar, 2005; Vrahatis, Boutsinas, Alevizos, & Pavlides, 2002).

4.3. Data mining tool – SPSS Clementine

In this study, SPSS Clementine is employed as data mining tool for analysis. The difference between Clementine and other software is that its data processing is through the use of nodes, which are then connected together to form a stream frame. In addition, data visualization can present to users after mining process has been done. All of the nodes can be divided into six categories: the source node, record options node, field options node, graphs node, modeling node, and output node.

Source node: The source selection not only includes those that can undergo the data connection node via the Open Database Connectivity (ODBC) and relational database management system. It also includes nodes that can input all sorts of common file contents.

Record options node: This is for the recording and correction of data. These operations are very important in the data investigation stage of data interpretation and data preparation stage, because these operations enable the data to fulfill specific business needs.

Field options node: The field option node can help the user do modeling and data preparation for the logical data design stage.

Graphs node: This is one of the stages in the data mining process that uses graphs for exploratory data analysis. Another purpose is to examine the new record option's distribution and relationships.

Modeling node: Modeling is the core of the data investigation process. The modeling method of this node enables the user to retrieve new information from the data and to form a forecast model. These modeling methods are derived Machine Learning (ML), Artificial Intelligence (AI), and Statistics, etc. All these methods have their own advantages and are suitable for specific types of problem. The algorithm includes:

(1) Decision Tree (C5.0, CART).
(2) Neural Net and RBF Function.
(3) Association Rule (Apriori, GRI).
(4) Sequence Detection.
(5) Clustering Analysis (K-means, Two-step and Kohonen).
(6) Regression (Linear Regression, Logistic Regression).
(7) Factor Analysis and PCA.

Output node: This offers one method to achieve data that is related to the users and the model. It can output all types of data in different forms to other software interfaces.

5. Data mining results

5.1. Mining results for the Taiwan stock market

In this study, the initial support and confidence are set to be 10% and 70% respectively. In addition, the lift value should be greater than 1. After exploring the decision variables of TAIEX and 19 index categories of stocks, the minimum support and confidence are 29% and 90% respectively. The mined association rules are shown in Tables 2 and 3. In addition, Fig. 3 shows the association map of TAIEX and index categories.

Table 2
Association rules of TAIEX and index categories (min sup = 29%; min conf = 90%)

Rule	Lift	Sup (%)	Conf (%)	Consequent	Antecedent		
R_{A1}	2.035	32.8	98.4	TAIEX	Electronics	Financial and insurance	
R_{A2}	2.021	32.0	97.7	TAIEX	Electronics	Textiles	
R_{A3}	2.019	30.2	97.6	TAIEX	Electronics	Rubber	
R_{A4}	2.015	31.0	97.4	TAIEX	Electronics	Foods	
R_{A5}	2.009	32.4	97.1	TAIEX	Electronics	Chemicals	
R_{A6}	1.998	29.1	96.6	TAIEX	Electronics	Steel and iron	
R_{A7}	1.915	29.4	92.6	TAIEX	Financial and insurance	Chemicals	Electric appliance and cable
R_{A8}	1.906	30.6	92.2	TAIEX	Financial and insurance	Electric and machinery	Chemicals
R_{A9}	1.902	29.6	91.9	TAIEX	Financial and insurance	Electric and machinery	Electric appliance and cable
R_{A10}	1.894	29.9	91.6	TAIEX	Financial and insurance	Chemicals	Textiles
R_{A11}	1.892	29.6	91.5	TAIEX	Financial and insurance	Electric appliance and cable	Textiles
R_{A12}	1.869	30.5	91.1	TAIEX	Financial and insurance	Electric and machinery	Textiles

Table 3
Association rules on each index category (min sup = 28%; min conf = 90%)

Rule	Lift	Sup (%)	Conf (%)	Consequent	Antecedent			
R_{B1}	1.865	28.6	91.0	Electric appliance and cable	Others	Electric and machinery	Textiles	Foods
R_{B2}	1.861	29.2	90.7	Electric appliance and cable	Others	Electric and machinery	Chemicals	Textiles
R_{B3}	1.857	28.4	91.4	Textiles	Others	Electric and machinery	Electric appliance and cable	Foods
R_{B4}	1.854	28.2	90.4	Electric appliance and cable	Others	Electric and machinery	Textiles	Rubber
R_{B5}	1.853	28.1	90.3	Electric appliance and cable	Electric and machinery	Chemicals	textiles	Foods
R_{B6}	1.85	28.8	91.0	Textiles	Others	Electric and machinery	Chemicals	Foods
R_{B7}	1.849	28.2	90.1	Electric appliance and cable	Others	Chemicals	Textiles	Foods

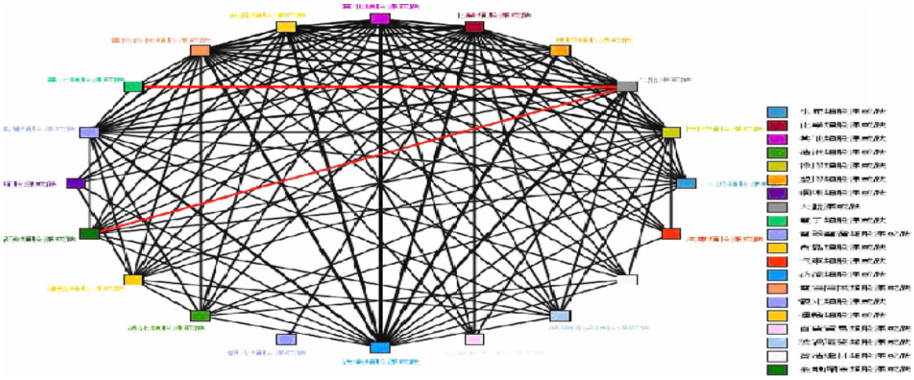

Fig. 3. Association map of TAIEX and index categories.

In the rules of Table 3, the rule R_{B1} has the maximum lift of 1.865, represents the index categories of food, textiles, electric and machinery; while the others are mostly associations to the index category of electrical appliance and cable. This rule can suggest that investors favor the stocks in these categories for their portfolio in order to diversify risk.

On the other hand, Fig. 4, which stems from Tables 2 and 3, and Fig. 3 shows that the index categories of electronics and finance and insurance have the strongest

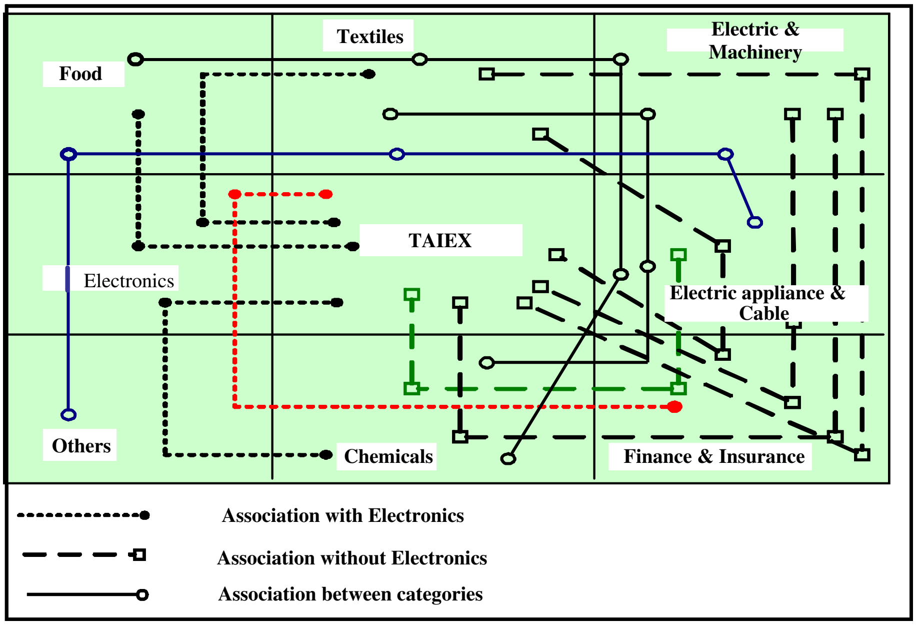
Textiles
Electric & Machinery
Food
TAIEX
Electronics
Electric appliance & Cable
Others
Chemicals
Finance & Insurance
Association with Electronics
Association without Electronics
Association between categories

Fig. 4. Association map of index categories on Taiwan stock market.

association to TAIEX in R_{A1}. Rule R_{A1} is in the rule set with the index category of electronics and R_{A1} has the maximum lift value up to 2.035 with maximum confidence of 98.4%. At the same time, we can find that the index category of electronics with other index categories of stock also have a very strong association to TAIEX. These results show that TAIEX is considerably affected by the index category of electronics. On the other hand, in the rule set without the index category of electronics, rule R_{A7} shows that the index categories of finance and insurance, electric appliance and cable and chemicals have the strongest association to TAIEX with the lift up to 1.915 and confidence of 92.6%. Hence, we can conclude that the rising (or falling) of the index category of electronics will lead to the rising (or falling) of the TAIEX. It also indicates that the probability of the TAIEX rising will increase when the index category of electronics is also rising. This result can be justified by the fact that the lift is almost over 2 and the confidence greater than 97% for the rules which contain the index category of electronics.

Other results can be found from Fig. 4. If we substitute the index category of electronics with the index category of finance to pull up the TAIEX, its pulling power is apparently less than the category of electronics. Whereas the index category of electronics with the other index categories of stock will pull up TAIEX to a certain extent, the index category of finance with the other 2 index categories of stock hardly pull up TAIEX to the same extent. Table 2 showed that the lifts of those rules without the index category of electronics are greater than one but less than two and confidences levels are approximately 91%. These results are far less then those rules which include the index category of electronics. Consequently, we can conclude that the index category of electronics is more vigorous than other index categories of stocks.

In addition, it can be found that the index categories of electronics, appliance and cable, others, electric and machinery, textiles and food have the highest associations, with lift value of 1.865 and confidence of 91%. Hence, we may propose the investment suggestions from Tables 2 and 3 and Fig. 4 that the index category of others should be added to a portfolio. This is because the index category of others is the most frequent in Table 3, indicating that the index category of others has a high association to with other index categories of stocks.

Table 4
Set of association rules between TAIEX and international indices (min sup = 17%; min conf = 45%)

Rule	Lift	Sup (%)	Conf (%)	Consequent	Antecedent
R_{C1}	1.297	52.8	62.8	TAIEX	South Korea
R_{C2}	1.254	49.0	60.8	TAIEX	Tokyo
R_{C3}	1.246	49.1	60.4	TAIEX	Hong Kong
R_{C4}	1.222	50.0	59.2	TAIEX	Singapore
R_{C5}	1.095	50.5	53.1	TAIEX	NASDAQ
R_{C6}	1.039	50.0	50.3	TAIEX	Shen Zhen
R_{C7}	1.033	49.3	50.1	TAIEX	Shanghai
R_{C8}	1.025	50.7	49.7	TAIEX	Dow Jones

5.2. Mining results for the international and Taiwan's stock market

We selected the stocks' indices of Taiwan's dominant trading countries in 2005 as decision variables. After several testing sequences, we set the minimum support as 17% and minimum confidence as 45%. The association rules were mined and are shown in Table 4, which shows that the confidence of TAIEX with South Korea, Tokyo and Hong Kong all exceed 60% and the lifts are also higher than the others. These results reveal that TAIEX has the strongest linkages with these three indices and they will have the same trends as TAIEX.

If we exploit the categories of stocks of Taiwan to mine the correlations with these three countries' stock indices, the mining results have been listed in Tables 5–7 and Fig. 5 which is composed by Tables 5–7.

The index categories of food, finance and insurance, and electronics have the strongest correlation with NIKKEI 225 Index. Their lift and confidence value are 1.436 and 69.7% respectively. Furthermore, the Hang Seng Index of

Table 5
Set of association rules between NIKKEI 225 and categories (min sup = 24%; min conf = 65%)

Rule	Lift	Sup (%)	Conf (%)	Consequent	Antecedent		
R_{D1}	1.436	25.0	69.7	NIKKEI 225	Foods	Financial and insurance	Electronics
R_{D2}	1.391	27.0	67.5	NIKKEI 225	Others	Financial and insurance	Electronics
R_{D3}	1.373	28.0	66.6	NIKKEI 225	Others	Electric appliance and cable	Electronics
R_{D4}	1.351	27.4	65.5	NIKKEI 225	Financial and insurance	Foods	Plastic

Table 6
Set of association rules between Hang Seng and categories (min sup = 27%; min conf = 65%)

Rule	Lift	Sup (%)	Conf (%)	Consequent	Antecedent		
R_{E1}	1.418	27.1	69.6	Hang Seng	Electric appliance and cable	Electronics	Financial and insurance
R_{E2}	1.396	27.1	68.6	Hang Seng	Electric appliance and cable	Electronics	Textiles
R_{E3}	1.375	27.8	67.5	Hang Seng	Others	Electronics	Textiles
R_{E4}	1.37	28.4	67.2	Hang Seng	Electric appliance and cable	Others	Electronics

Table 7
Set of association rules between KOSPI and categories (min sup = 31%; min conf = 70%)

Rule	Lift (%)	Sup (%)	Conf	Consequent	Antecedent		
R_{F1}	1.406	33.9	74.8	KOSPI	Electronics	Other categories	
R_{F2}	1.393	33.4	74.1	KOSPI	Electronics	Electric appliance and cable	
R_{F3}	1.335	32.4	71.0	KOSPI	Others	Electric appliance and cable	Chemicals
R_{F4}	1.329	31.2	70.7	KOSPI	Others	Electric appliance and cable	Foods

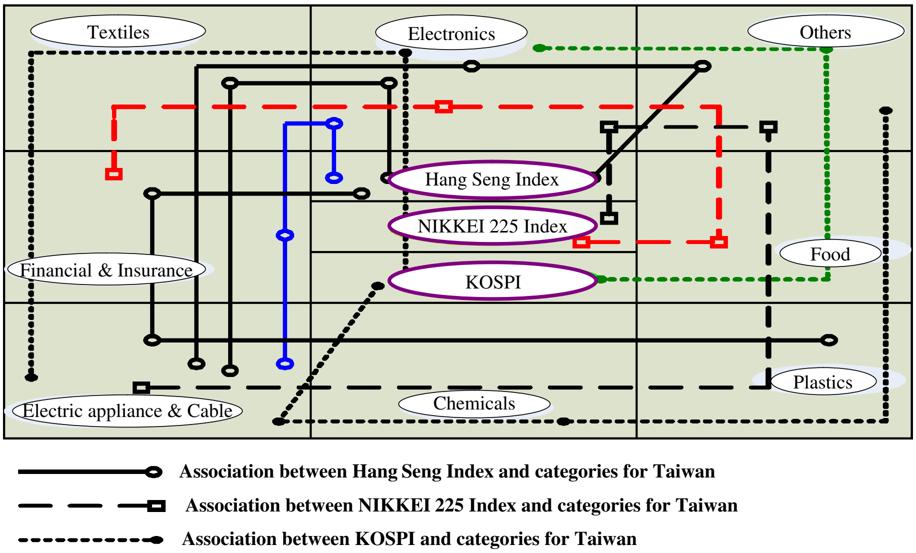
Textiles
Electronics
Others
Hang Seng Index
NIKKEI 225 Index
Financial & Insurance
Food
KOSPI
Plastics
Electric appliance & Cable
Chemicals
Association between Hang Seng Index and categories for Taiwan
Association between NIKKEI 225 Index and categories for Taiwan
Association between KOSPI and categories for Taiwan

Fig. 5. Association map of categories with international markets.

Hong Kong is correlated with the categories of electric appliance and cable, electronics and finance and insurance with lift of 1.418 and confidence of 69.6%. In addition, KOSPI is highly correlated to the categories of electronics and others with lift of 1.406 and confidence of 74.8%. Hence, we can conclude that these three countries stocks' indices are most correlated to the categories of electronics, electric appliance and cable and others.

Comparing the NIKKEI 225 index and the Hang Seng index from Table 4, KOSPI has highest confidence and support with TAIEX. This reveals that TAIEX is highly correlated with KOSPI. Therefore, they both will have the same trend of variation. In this study, we find that TAIEX has higher correlations with Asian stock markets, because Taiwan is in this region. Taiwan's top 5 dominant trading countries are all located in Asia which give the TAIEX correlations with these three indices. Finally, in contrast to previous studies, we find that in addition to the category of electronics, the categories of electric appliance and cable and others have high correlation with international stock markets, especially the Asian markets.

5.3. Clustering results for the whole market

To provide overall statements between Taiwan and global markets, we use cluster analysis to separate the markets and use association rule for these separated markets. The clustering results are shown in Table 8, which indicates that there are two clusters, cluster-1 and cluster-2 of these, cluster-1 has less possibility to increase. From each cluster, we select 4 categories that have the greater possibility to increase for cross-comparison and the comparison results are shown in Table 9. Although the lifts are all greater than 1 in both clusters, the lifts in cluster-2 are greater than in cluster-1. This reveals that the categories in cluster-2 are more able to pull up TAIEX, which is similar to the association analysis. In Table 3, we can easily find that the association rules comprised by the categories in cluster-2 have higher lifts. Consequently, the results of association analysis and cluster analysis are not conflicting, but are identical. Moreover, the lifts in cluster-1 are smaller than the lifts between cluster-1 and cluster-2 shown in Table 9 lead to insufficient power to pull up a bear market. Therefore, cluster analysis could not provide the information of cross-effect between cluster-1 and cluster-2.

Fig. 6 is the bar chart composed of the 1s in two clusters. Fig. 8 also reveals that the situations in two clusters are identical to TAIEX. For instance, the confidence values in Tables 2 and 3 are all above 90%. Similar results emerge for cluster-2, since the rising possibilities are all above 70%. Furthermore, we could investigate the linkages between Taiwan and international stock markets. The stock markets of Tokyo, Hong Kong and South Korea are most linked to Taiwan. Although the Shanghai synthesis index has more than 80% rising possibility with TAIEX in cluster-2, the degree of correlation between three countries'

Table 8
Percentages in two clusters of stock market

Category	Cluster-1, 574 records		Cluster-2, 607 records	
Fall:0/rise:1	0	1	0	1
Others	83.97%	16.03%	17.13%	82.87%
Chemicals	81.01%	18.99%	18.78%	81.22%
Plastics	78.22%	21.78%	24.71%	75.29%
Rubber	81.71%	18.29%	22.24%	77.76%
Cement	80.14%	19.86%	25.21%	74.79%
Automobile	77%	23%	27.84%	72.16%
Construction	81.36%	18.64%	29.82%	70.18%
Glass and ceramics	79.62%	20.38%	24.88%	75.12%
Wholesale and retail	77%	23%	28.83%	71.17%
Textiles	83.80%	16.20%	19.44%	80.56%
Tourism	75.61%	24.39%	32.45%	74.30%
Paper and pulp	82.06%	17.94%	25.70%	74.30%
Transportation	77.70%	22.30%	26.52%	73.48%
Finance and insurance	82.23%	17.77%	23.23%	76.77%
Steel and iron	76.31%	23.69%	29.98%	70.02%
Electric appliance and cable	82.58%	17.42%	22.90%	77.10%
Electronics	77.70%	22.30%	29.49%	70.51%
Electric and machinery	81.18%	18.82%	18.78%	81.22%
Food	80.66%	19.34%	21.91%	78.09%
TAIEX	86.24%	13.76%	18.78%	81.22%
South Korea	58.89%	41.11%	36.08%	63.92%
Hong Kong	60.28%	39.72%	42.17%	57.83%
Tokyo	61.15%	38.85%	41.35%	58.65%
Shanghai	51.05%	48.95%	18.78%	81.22%
Singapore	59.93%	40.07%	40.36%	59.64%
Shenzhen	51.39%	48.61%	49.26%	50.74%
Dow Jones	49.13%	50.87%	49.26%	50.74%
NASDAQ	54.01%	45.99%	45.14%	54.86%

indices and TAIEX are higher than the Shanghai synthesis index in cluster-1. From the point of view of linkage, Taiwan is most synchronous with Tokyo, Hong Kong and South Korea. We can prove this result from the association rules R_{c1}, R_{c2} and R_{c3} in Table 4, which have confidence value above 60%. The results are identical with in Table 8 and Fig. 6.

6. Research findings

6.1. In the regard of Taiwan's stock market

According to statistics of the Taiwan Stock Exchange Corporation (TSEC) published in 2004, there is a total of 538,085,368,762 stock shares issued to the public market. In these issued stocks, the category of electronics comprise 42.2%, while the category of finance and insurance comprise 23.99%. Both of them occupied a proportion of 66.19% to the whole issued stocks. Because TAIEX is computed based on the weighting values (proportions), the category of stock which has larger weighting value will become the primary affected factor of TAIEX. Because the category of electronics has the largest proportion of the market, it has the greatest primary ability to pull up TAIEX, while the category of finance and insurance was a secondary ability to pull up TAIEX. Therefore, investors in Taiwan's stock market are suggested to favor these two categories of stocks for their portfolios.

Table 9
Lift value among categories that have most rising possibility on two clusters

	Tourism[a]	Steel[a]	Automobile[a]	Wholesale[a]	Others[b]	Chemicals[b]	Electric[b]	Textiles[b]
Tourism[a]	1	1.276	1.31	1.368	1.351	1.41	1.376	1.37
Steel[a]	–	1	1.361	1.274	1.381	1.382	1.416	1.408
Automobile[a]	–	–	1	1.318	1.454	1.379	1.416	1.412
Wholesale[a]	–	–	–	1	1.423	1.396	1.391	1.394
Others[b]	–	–	–	–	1	1.495	1.557	1.509
Chemicals[b]	–	–	–	–	–	1	1.469	1.515
Electric[b]	–	–	–	–	–	–	1	1.514
Textiles[b]	–	–	–	–	–	–	–	1

[a] Represent higher rising possibility of stock category on cluster-1.
[b] Represent higher rising possibility of stock category on cluster-2.

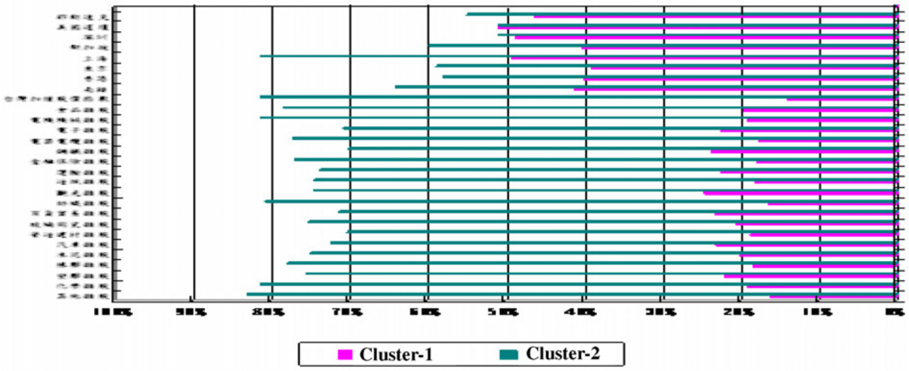

Fig. 6. Bar chart of two cluster comparison based on lift value 1.

In addition, according to the statistics of transaction percentage published by TSEC in 2005, domestic legal entifies occupied 13.3%, and the foreign legal entifies occupied 15.5% of transactions on the market. Although together they comprise less 30% of the total transactions, they both have the significant abilities to affect Taiwan's stock market due to their superior information and capital. In particular, the foreign entifies have the most impact on Taiwan's stock market. The investors may refer to the categories of stocks which the domestic legal and foreign entifies invested, but they should be cautioned to avert the trap that the both two entifies deliberately manipulate stock prices and then put their stocks back on the market. In this case, we would propose the following suggestion to diversify investment risk. Due to the correlations between the categories of electric machinery, chemicals, textiles, others and electric appliance and cable, investors may allocate these categories of stocks with categories of electronics and finance and insurance in their portfolios to diversify investment risk (Fig. 7).

6.2. In the regard of Taiwan with international stock market

TAIEX has higher correlations with NIKKEI, Hang Seng index, and KOSPI than other Asian indices. This shows that TAIEX will rise easily, when other Asian indices are rising. It revealed that the linkages between nations' indices in the same area are higher than linkages with indices from other area. Thus investors can use these regional indices as a reference for their investment decisions. In these three indices which have the most correlation to TAIEX, the category of electronics also has an important role. Furthermore, in addition to the category of electronics, the categories of electric appliance and cable as well as others could be better choices for investment. Therefore, we suggest that investors investigate the trends of these three indices in order to select the investing targets most which link to these 3 regional indices in Taiwan's market (Fig. 8).

6.3. In the regard of global market

Emerging markets are more vulnerable to local information than mature markets. Taiwan's stock market is an emerging market which is easily affected by local information and this will be reflected in the stock's index. Thus, investors should allocate their portfolio in a specific area and diversify their portfolio in this area's market. For example, stock index futures are a hedging tool to diversify the risk investment of stocks. Due to highly correlations between TAIEX and the indices of Tokyo, Hong Kong, and South Korea, investors could refer these 3 regional

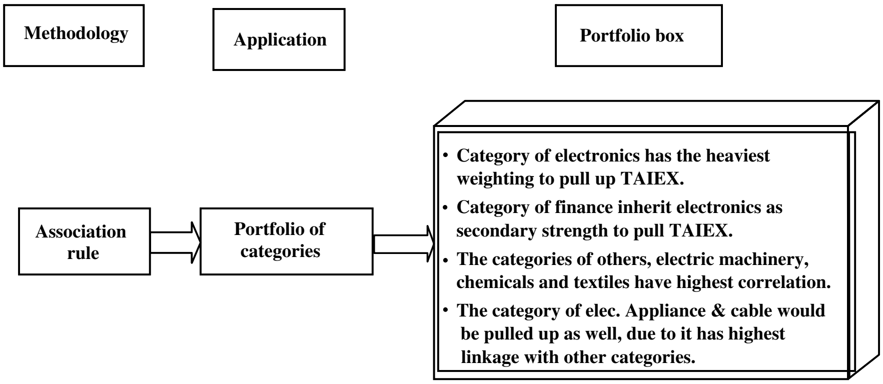
Methodology
Application
Portfolio box
Association rule
Portfolio of categories
• Category of electronics has the heaviest weighting to pull up TAIEX.
• Category of finance inherit electronics as secondary strength to pull TAIEX.
• The categories of others, electric machinery, chemicals and textiles have highest correlation.
• The category of elec. Appliance & cable would be pulled up as well, due to it has highest linkage with other categories.

Fig. 7. Possible portfolio of stock index categories in Taiwan market.

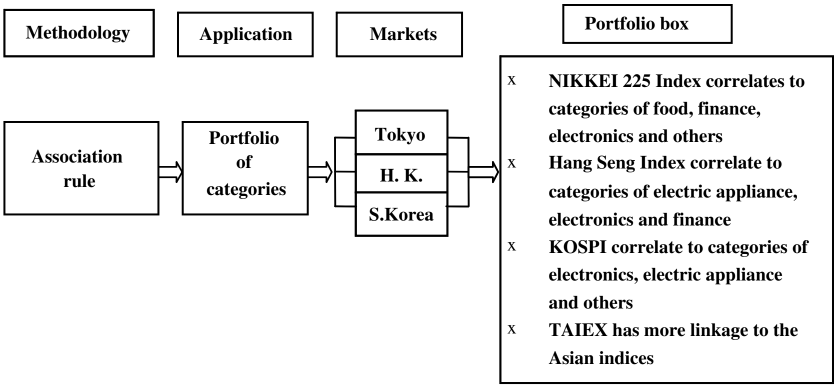
Methodology
Application
Markets
Portfolio box
Association rule
Portfolio of categories
Tokyo
H. K.
S.Korea
x NIKKEI 225 Index correlates to categories of food, finance, electronics and others
x Hang Seng Index correlate to categories of electric appliance, electronics and finance
x KOSPI correlate to categories of electronics, electric appliance and others
x TAIEX has more linkage to the Asian indices

Fig. 8. Possible portfolio of Taiwan and international markets.

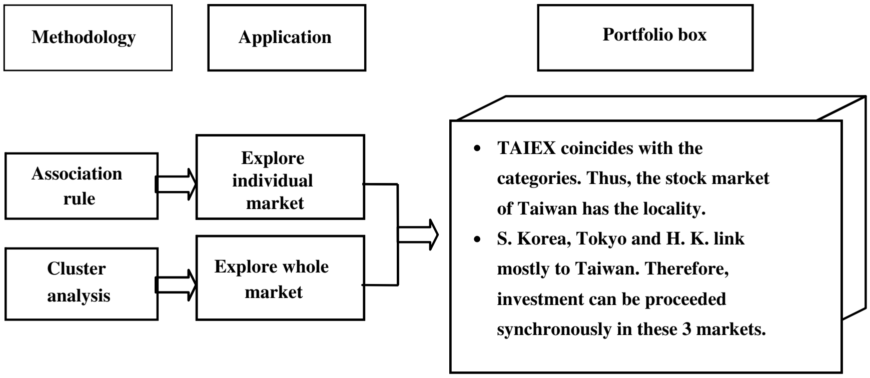
Methodology
Application
Portfolio box
Association rule
Explore individual market
Cluster analysis
Explore whole market
• TAIEX coincides with the categories. Thus, the stock market of Taiwan has the locality.
• S. Korea, Tokyo and H. K. link mostly to Taiwan. Therefore, investment can be proceeded synchronously in these 3 markets.

Fig. 9. Possible portfolio for a global market.

indices to anticipate the variation trends of TAIEX. Investors should diversify their investment in a specific regional market rather than diversified their capital over several regional markets. The Taiwan 50 index was issued by TSEC in October 2002 and is composed of the top 50 stocks in market value on the Taiwan public market. Because the Taiwan 50 index is closely linked to TAIEX with a correlation of up to 98%, investors could select it as an investment target for Taiwan market when they refer to those 3 regional indices (Fig. 9).

7. Conclusion

This paper considers that a stock market strong associations with both inside and outside factors. Some stock index categories of stock rise or drop together at the same time or are influenced by domestic or foreign economic, social, and political situations. For individual or institutional investors, finding indications for the trend of stock market association is a valuable task. Data mining of the stock market analysis and interpretation of the properties of the data mining results gives new insights into possible associations in the stock markets. In this paper, we use two data mining approaches, Apriori algorithm and K-means, for association rule and clustering analysis. By doing so, this research finds that different possible portfolio of stock categories investment can be implemented in the Taiwan stock market. Thus, this case study of implementing data mining approaches and integrating them into stock market research on Taiwan stock market is an example for future research and implementation.

References

Agrawal, R., Imilienski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on management of data* (pp. 207–216).

Agrawal, R., & Shafer, J. (1996). Parallel mining of association rules. *IEEE Transactions on Knowledge and Data Engineering, 8*(6), 962–969.

Boginski, V., Butenko, S., & Pardalos, P. M. (2006). Mining market data: A network approach. *Computers & Operations Research, 33*, 3171–3184.

Chun, S. H., & Park, Y. J. (2005). Dynamic adaptive ensemble case-based reasoning: Application to stock market prediction. *Expert Systems with Applications, 28*, 435–443.

Devlin, B. (1997). *Data warehouse: From architecture to implementation*. Reading, MA: Addison Wesley Longman.

Enke, D., & Thawornwong, S. (2005). The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with Applications, 29*, 927–940.

Fama, E. (1991). Efficient capital markets. *Journal of Finance, XLVI*, 1575–1617.

Haugen, R. (1997). *Modern investment theory*. Upper Saddle River, NJ: Prentice-Hall.

Kouris, I. N., Makris, C. H., & Tsakalidis, A. K. (2005). Using information retrieval techniques for supporting data mining. *Data & Knowledge Engineering, 52*, 353–383.

Kuo, R. J., Liao, J. L., & Tu, C. (2005). Integration of ART2 neural network and genetic K-means algorithm for analyzing Web browsing paths in electronic commerce. *Decision Support Systems, 40*(2), 355–374.

Padmanabhan, B., & Tuzhilin, A. (2002). Knowledge refinement based on the discovery of unexpected patterns in data mining. *Decision Support Systems, 33*, 309–321.

Rapach, D. E., & Wohar, M. E. (2006). In-sample vs. out-of-sample tests of stock return predictability in the context of data mining. *Journal of Empirical Finance, 13*, 231–247.

Taiwan Stock Exchange Council (2005). Annual Report of Taiwan Stock Market Exchange Statistics. http://www.tse.com.tw/ch/products/indices/ftse/taiwan50.php.

Ture, M., Kurt, I., Turhan, K. A., & Ozdamar, K. (2005). Comparing classification techniques for predicting essential hypertension. *Expert Systems With Applications, 16*(4), 379–384.

Vrahatis, M. N., Boutsinas, B., Alevizos, P., & Pavlides, G. (2002). The new k-windows algorithm for improving the k-means clustering algorithm. *Journal of Complexity, 18*(1), 375–391.

Wang, Y. F., Chuang, Y. L., Hsu, M. H., & Keh, H. C. (2004). A personalized recommender system for the cosmetic business. *Expert Systems with Applications, 26*, 427–434.