Concept Hierarchy-Based Pattern Discovery in Time Series Database: A Case Study on Financial Database

Yan-Ping Huang
Department of Information Management, Chin Min Institute of Techndogy

Chung-Chian Hsu
Department of Information Management, National Yunlin University of Science and Technology

Abstract

Data Mining is the process of automatically searching large volumes of data for patterns and it is also a fairly recent and contemporary topic in computing. Nowadays, pattern discovery is a field within the area of data mining. In general, large volumes of time series data are contained in financial database and these data have some useful but not easy finding patterns in it and many financial studies in time series data analysis use linear regression model to estimate the variation and trend of the data. However, traditional methods of time series analysis used special types or linear models to describe the data. Linear models can achieve high accuracy when linear variation of the data is small, however, if the variation range exceeds a certain limit, the linear models has a lower performance in estimated accuracy. SOM is a famous non-linear model and traditional method to extract pattern with numeric data. Many researches extract pattern from numeric data attributes rather than categorical or mixed data. It does not extract the major values from pattern rules, either.

The purpose of this study is to provide a novel architecture in mining patterns from mixed data that uses a systematic approach in the financial database information mining, and try to find the patterns for estimate the trend or for special event's occurrence. This study uses ESA algorithm to discover the pattern in the Concept Hierarchy based Pattern Discovery (CHPD) architecture. Specifically, this architecture facilitates the direct handling of mixed data, including categorical and numeric values. This mining architecture can simulate human intelligence and discover patterns automatically, and it also demonstrates knowledge pattern discovery and rule extraction.

Key words: Data mining, Cluster analysis, Pattern discovery, Time series analysis.

以概念階層為導向之時間序列模式資料探勘——以財務資料庫為例

黃燕萍
親民技術學院資訊管理學系

許中川
雲林科技大學資訊管理學系

摘要

資料探勘是從大量資料中擷取隱藏、未知與潛在，但具有實用性的資訊分析方法。在資料探勘領域中，知識探勘的相關研究已有長足的進步。時間序列資料，包含大量未知與潛在的資訊。財務類型的資料庫中，通常存有大量的時間序列資料。過去時間序列相關研究以迴歸分析為主，傳統迴歸分析模型的統計性質，多半建立在線性模型的基礎上；然而，線性模型對於變動幅度不大的非線性模型，尚可作較高準確度的估計，但是，若變動幅度超過某一限度，估計的準確性就會降低，因而減少其應用上的價值。自組映射圖類神經網路，目前是時間序列資料研究中經常使用的分析方法之一。然而，自組映射圖類神經網路為一種高度非線性模式，只能處理數值型資料，無法有效處理混合型的資料。

因此，本研究提出以概念階層為導向之樣版資料探勘模式，利用物以類聚的原理，從分析過去的樣本資料中，針對財務類型資料庫之時間序列資料，學習樣版辨識，利用同類相聚的特性以達分群之目的；更進一步在模式中找出未知、潛在但具有實用性的樣版資訊，及精簡且具代表性的規則，以此協助預估財務資料的變動。

關鍵字：資料探勘、分群演算法、樣板探勘、時間序列分析

1. Introduction

Knowledge discovery in database is a nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in the data [Fayyad et al., 1991; Fayyad et al., 1996]. Data mining is a step of knowledge discovery process consisting of particular algorithms to produce patterns [Chen & Han, 1996; Imielinski & Mannila, 1996]. Data mining, which is also referred to as knowledge discovery in databases, has been recognized as the process of extracting non-trivial, implicit, previously unknown, and potentially useful information from data in databases [Agrawal el al., 1993; Han & Kamber, 2001].

Data mining is the process of automatically searching large volumes of data for patterns. Data mining is a fairly recent and contemporary topic in computing. Pattern discovery is a field within the area of machine learning. In the real world, there are thousands of time series data that coexist with others. Time series dataset arises in medical, economic and scientific applications. Traditional methods of time series analysis always use special types or linear models to describe the analysis data. Linear models can achieve high accuracy when linear variation of the data is small, but if the variation range exceeds a certain limit, the linear models give a lower performance in estimated accuracy. The time series data mining architecture concerns two general questions. First, it defines the patterns with appropriate data mining tools. Second, it shows the patterns derived as profitable or informative [Huang et al., 2007]. How to find a pattern from the time series datasets and how to prove the pattern be useful become more and more important.

However, many researches face pattern discovery in numeric data attributes. It also does not extract the major values from pattern rules. It has little human intelligence. Therefore, the present article provides a brief architecture to find the pattern, which is defined by the user's request returns and prove the pattern be profitable or informative. This study addresses the issue by proposing a framework for time series datasets.

The remainder of the paper is organized as follows. Section 2 introduces the related literature. The time series pattern mining architecture is developed in Section 3. Section 4 presents the performance study. Section 5 discusses the issues and points out some future research plans.

2. Related Work

Artificial neural networks (ANNs) are motivated by biological neural networks. Clustering is the unsupervised classification of patterns into groups [Jain et al., 1999]. In

competitive learning, similar patterns are grouped by the network and represented by a single neuron. This grouping is done automatically based on data correlations. Well-known examples of ANNs used for clustering include Kohonen's learning vector quantization (LVQ) and self-organizing map [Kohonen, 1984], and adaptive resonance theory models [Carpenter & Grossberg, 1990].

The SOM gives an intuitively appealing two-dimensional map of the multidimensional, data set, and it has been successfully used for vector quantization and speech recognition. SOM is suitable for detecting only hyper spherical clusters [Hertz et al., 1991]. A two-layer network that employs reg5larized Mahalanobis distance to extract hyper ellipsoidal clusters was proposed in Jain and Mao [Jain & Mao, 1994].

In recent years, many researchers 4ried to use SOM algorithm or other related techniques to discover the patterns from a huge financial database to support investors to make decisions, financial forecasting and management [Chen & He, 2003; Chen & Tsao, 2003; Deboeck & Kohonen, 1998; Deboeck & Alfred, 2000; Kohonen, 1996], medical diagnosis [Vesanto et al., 1999; Chen et al., 2000], image object classification [Kramer et al., 2000], image retrieval [Becanovi, 2000], and image processing [Laaksonen et al., 2000; Toivanen et al., 2003; Gunter & Bunke, 2002; Wu & Chow, 2004].

The ViSOM, extends from SOM, is a non-linear multi-dimensional projection method. In order to faithfully preserve the structure of the training data to the map, the ViSOM takes account of the distances between neurons in the data space and on the map, respectively. The objective of ViSOM is to preserve the data structure and the topology as faithfully as possible. The ViSOM considers the distance between two neurons (winner and neighborhood) in the data space and on the map, respectively, and uses a resolution parameter that controlled the inter-neuron distance on the map [Yin, 2002a; Yin, 2002b]. The EViSOM algorithm integrates the concept hierarchies such that the extended system properly handles the mixed data [Hsu & Wang, 2005].

Attribute-oriented induction is a method for knowledge discovery in databases that has recently been described and widely applied by Han. [Cai et al., 1991; Han et al., 1992; Han & Cercone, 1993]. AOI has been applied to many fields for data analysis, including GIS data for discovering association rules between geographic data and non-geographic data [Ester et al., 1997; Koperski et al., 1998], analyzing patterns from multimedia data [Zaiane et al., 1998], and integrating with clustering algorithm for analyzing web access patterns in access logs [Fu et al., 1999]. However, AOI still has problems. If a few of an attribute's values take up a major portion of the attribute, the traditional approach might overly generalize that attribute and thus cannot reveal the fact that there are major values in that attribute. EAOI resolves the problem. It utilizes concept hierarchies, which are associated with attributes, to generalize data and then output general, concise patterns of

the original data. It generalizes specific data using concept hierarchies and produces concise patterns from a large amount of raw data [Hsu, 2004].

Time series analysis is an important course of financial management. To approach idealization and simplification, traditional methods of time series analysis always use special types or linear models to describe the analysis data. In the time-domain models, Engle（1982）described ARIMA models (Autoregressive Integrated Moving Average Process) [Fama, 1992], and Bollerslev (1986) described GARCH model (Generalized Autoregressive Conditional Heteroscedasticity) [Bollerslev, 1986] for the extrapolation of past values into the immediate future. It was based on the correlations among lagged observations and error terms. It does not match the fact condition. Linear models can achieve high accuracy when linear variation of the data is small, but if the variation range exceeds a certain limit, the linear models give a lower performance in estimated accuracy.

Artificial neural network is a tool of information technique that rapidly raises in these few years, especially using in financial area, the performance is very outstanding. SOM theory is one of artificial neural network. It is the new science in these years, has a good ability to explain the nature phenomena. Identifying fractal structure can explore the behaviors behind the time series data. Some researches propose the self-organizing maps great clustering function and human visualization to discover the patterns from a huge financial database to support investors to make decisions, financial forecasting and management [Li & Kuo, 2005; Kuo et al., 2004; Li, 2001; Tsai et al., 2005]. The SOM networks have been successfully applied to decision support systems, including group technology [Kiang et al., 1995; Kulkarni et al.,1995; Kiang et al., 1995; Ong et al., 2005], financial news map [Smith & Ng, 2003], telecommunication [Kiang et al., 2006]. These researches deal with the general numerical data. However, they do not deal with mixed data. Therefore, this study proposes ESA algorithm and the conceptual hierarchy tree to solve similar degrees of mixed data.

About clustering mixed data attributes, there are two approaches for mixed data. One is resorted to a pre-process, which transferred the data to the same type, either all numeric or all categorical. For transferring continuous data to categorical data, some metric function is employed. The function is based on simple matching in which two distinct values result in distance 1, with identical values of distance 0 [Guha et al., 1999]. The other is to use a metric function, which can handle mixed data [Wilson & Martinez, 1997]. Overlap metric is for nominal attributes and normalized Euclidean distance is for continuous attributes.

Among problems with simple matching and binary encoding, a common approach for handling categorical data is simple matching, in which comparing two identical categorical values result in distance 0, while two distinct values result in distance 1 [Wilson & Martinez, 1997; Ester et al., 1998]. In this case, the distance between patterns of YUFO

and Foxconn in the previous example becomes d(YUFO, Foxconn) =1, which is the same as d(Foxconn, Leadtek) = d(YUFO, Leadtek) = 1. Obviously, the simple matching approach disregards the similarity information embedded in categorical values. Another typical approach to handle categorical attributes is to employ binary encoding that transforms each categorical attribute to a set of binary attributes and a categorical value is then encoded to a set of binary values. As a result, the new relation contains all numeric data, and the clustering is therefore conducted on the new dataset. For example, as the domain of the categorical attribute: Product_Attribute. The set of it is {PSP, iPOD, GPS}. Product_Attribute is transformed to three binary attributes: PSP, iPOD and GPS in the new relation. The value PSP of Product_Attribute in a pattern is transformed to a set of three binary values in the new relation, i.e. {PSP=1, iPOD=0, GPS=0}. The Euclidean distance of patterns YUFO and Foxconn is d(YUFO, Foxconn) = $\sqrt{2}$, which is the same as d(YUFO, Leadtek) and d(Foxconn, Leadtek), according to the new relation. Traditional clustering algorithm transfers Product_attribute categorical attributes into a binary numerical attribute type as shown in table 1.

Table 1: Traditional clustering algorithm transfers Product_Attribute categorical attributes into binary numerical attribute type.

ID	Product_Attribute	Price.
YUFO	PSP	76
Foxconn	iPOD	200
Leadtek	G PS	35

➡

ID	PSP	iPOD	GPS	Price.
YUFO	1	0	0	76
Foxconn	0	1	0	200
Leadtek	0	0	1	35

After transformation, each original categorical attribute handles by the binary encoding approach contributes as twice as that by the simple matching approach, as shown in the above example of distance (YUFO, Foxconn). Consequently, when the binary encoding approach is adopted by a `clustering algorithm, categorical attributes have larger influence on clustering data than those adopting the simple matching approach.

The time series data mining architecture concerns two general questions. First, it defines the patterns with appropriate data mining tools. Second, it shows the patterns derived as profitable or informative. ESA algorithm combines Extended Visualization-induced Self-Organizing Map algorithm (EViSOM) [Hsu, 2004; Wang & Hsu, 2005; Huang et al., 2007] and Extended Attribute-Oriented Induction algorithm (EAOI) [Hsu, 2004] to automatically discover the patterns in the financial database. EViSOM algorithm calculates the distance between the categorical and numeric data for pattern finding and EAOI algorithm generalizes major values using concept hierarchies for pattern major values extraction in financial database. This study uses ESA algorithm to discover

the pattern in the CHPD architecture. We provide a brief architecture to find the pattern, which is defined by the user's request returns and prove that the pattern is profitable.

3. Methodology

3.1 Pattern mining architecture for Time series dataset

The concept hierarchy based pattern discovery architecture provides four steps of stock information mining. These steps included: preprocess, pattern discovery analysis and extraction, pattern evaluation analysis and comparison evaluation. Figure 1 shows the CHPD architecture.

1. Preprocess: This step included removes noise and inconsistent data. It retrieves relevant data from the database and transforms the data format for pattern discovery.
2. Pattern discovery and extraction: This step has two phases. First, the intelligent methods are employed to extract knowledge patterns, and tried to identify representing knowledge of the patterns which is using the ESA algorithm. Second, major values are extracted from the pattern features and pattern rules are described. The target pattern vision is defined by the investor. The ESA algorithm is described in Section 3.3.
3. Pattern evaluation: It evaluates the target pattern features or groups from source data. The source class is specified by the ESA algorithm.
4. Comparison evaluation: Patterns are referred as actionable. Patterns can be used in strategy. Measuring the interest in patterns is essential for the efficient discovery of the value of patterns.

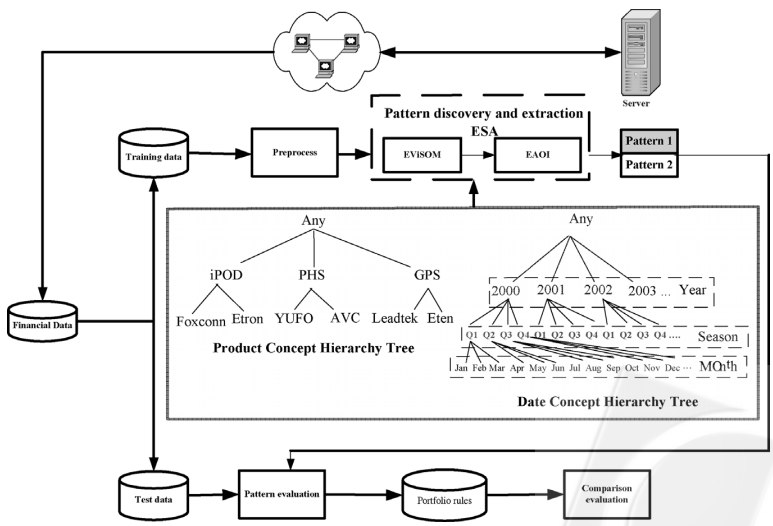

Figure 1: The concept hierarchy based pattern discovery architecture.

3.2 Example

Assuming a set of data records is stored in a database table T defined by a set of attributes $A_1, A_2, ..., A_m$, where m is the number of attributes. Each attributes A_j describes values with domain value $DOM(A_j) = \{a_{m,1}, a_{m,2}, ..., a_{m,nm}\}$, where n_m is the number of distinct values in attributes A_j. For example, DOM (Industry) = {iPOD, GPS, PSP}. The time series datasets based on the financial dataset are implemented in table2. Here i is the time interval and j is the dataset in each data interval. If i = 2 and j = 5, it means that there are two intervals. In each interval, there are five days. A typical time series dataset has the following format: $x_{ij} = \{x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{21}, x_{22}, x_{23}, x_{24}, x_{25}\}$.

Table 2: A portion of the time series dataset .

No	A_1	A_2	A_3	A_4	A_5	No	A_1	A_2	A_3	A_4	A_5
X_{11}	174.4	17243	4.53	1.5	1146	X_{21}	170.37	7926	-2.3	0.69	1146
X_{12}	170.4	7926	-2.31	0.69	1146	X_{22}	169.87	8497	-0.3	0.74	1146
X_{13}	169.9	8497	-0.3	0.74	1146	X_{23}	166.34	6184	-2.1	0.54	1146
X_{14}	166.3	6184	-2.08	0.54	1146	X_{24}	174.91	15577	5.15	1.36	1146
X_{15}	174.9	15577	5.15	1.36	1146	X_{25}	172.39	18476	-1.4	1.61	1146

Table 2 shows an example, which has two interval cycles and each interval cycle has datasets of five days, each record has five attributes. Following is an example in the CHPD architecture. Table 3 shows an example of original financial dataset.

Table 3: A portion of the financial dataset.

Id	Product	Stocks name	Date	Price	Return	TV	TR	SIC	BME
5412	iPOD	Foxconn	2000/02/09	108.72	0.81	16006	1.46	1100	1.90
5413	iPOD	Foxconn	2000/02/10	110.92	2.02	9863	0.90	1100	1.93
5414	iPOD	Foxconn	2000/02/11	118.37	6.72	18152	1.65	1100	2.05
5415	iPOD	Foxconn	2000/02/14	118.37	0	21421	1.95	1100	2.08
5416	iPOD	Foxconn	2000/02/15	124.95	5.56	12693	1.15	1100	2.20

Note: TV (Trading Volume), TR (Turnover Rate), SIC (Size In Circulation), BME (Book to Market Equity)

About attributes selection, this study chooses the attributes like return, price, trading volume, turnover rate, size in circulation, and the book to market equity. About the attributes of book to market and the book to market equity attributes, Fama and French (1992,1995) described the portfolio strategies and found out that size in circulation and book-to-market equity factors could affect stock returns[Fama & French, 1992; Fama & French, 1995]. Jegadeesh and Titman (1993) described that momentum strategies could affect the stock returns [Jegadeesh & Titman, 1993]. Brennan, Chordia and Subrahmanyam (1996) described the trading volume factor and it is the cross-section of the expected stock

returns [Brennan et al., 1998]. Therefore, this research uses these core stock indices like price, trading volume, turnover rate, size in circulation, and the book to market equity as dataset attributes.

3.3 Extended visualization-induced Self-organizing map and Attribute-oriented induction algorithm

ESA algorithm combines the EViSOM algorithm [Hsu & Wang, 2005; Huang et al., 2007] and EAOI algorithm [Hsu, 2004] according to the concept hierarchy for pattern discovery and pattern rule extraction. The ESA algorithm is a clustering algorithm for finding pattern from mixed data. It can generalize the attributes and extract the major values and rules from the pattern attributes. The ESA algorithm is outlined in the following steps:

Input: Time-series database, a relation W with an attribute set A; a set of concept hierarchies; sliding-window threshold s; the learning rate $\alpha(t)$; generalization threshold θ, and majority threshold β.

Output: Prototype vectors and generalized relation P.

Method:

Step 1. It preprocesses task-relevant records in the dataset. The sliding-window calculates the return for days exceeding the given threshold. The threshold can be defined by the user or the investor. The investor can define which performance pattern is preferred by the user. Such that $f(x_i) = \sum_{i=1}^{n}(x_i)$, i is the time interval and j is the dataset in each interval data. If $f(x_i)$ > threshold, then it uses these datasets for prediction using pattern discovery analysis.

Step 2. Update the weights of the numeric attributes of neighboring neurons according to Equation (1)

$$\begin{aligned} F_{kx} &\equiv x(t) - w_k(t) = [x(t) - w_v(t)] + [w_v(t) - w_k(t)] \\ &\equiv F_{vx} + F_{kv} \end{aligned} \tag{1}$$

In which, the updating force $[x(t) - w_k(t)]$, can be rearranged and decomposed into two forces. w_k is a neighborhood neuron.

Update and the weights of the categorical attributes, the distance between two points, X and Y, in the concept hierarchy is defined according to Equation (2) and (3)

$$| X - Y | = d_X + d_Y - 2 \times d_{LA} \tag{2}$$

$$d_{LA} = \min(d_X, d_Y, d_{LCA(N_X, N_Y)}) \tag{3}$$

Where the least ancestor (LA) is a point that is closest to the root among point X, point Y, and the LCA of point X and point Y. The d_{LA} is the distance from

point LA to root. It represents the part of duplicate distance and must be discarded.

Step 3. It initializes the map or weights either to the principal components or to small random values. Each of the neurons i in the 2-D map is assigned a weight vector. At each training step t, a training data $x(t) \in R^n$ is randomly drawn from the dataset and the Euclidean distances between $x(t)$ and all neurons are calculated. A winning neuron w_v can be found according to the minimum distance to $x(t)$. Find the winner neuron according to Equation (4)

$$v = \arg\min_i \| x(t) - w_i(t) \|, i \in \{1,...,M\} \tag{4}$$

Step 4. The SOM adjusts the weight of the winner neuron and neighborhood neurons. It moves closer to the input vector in the input space, and updates the weights of the winner neuron according to Equation (5)

$$w_i(t+1) = w_i(t) + \alpha(t) \times h_{vi}(t) \times [x(t) - w_i(t)] \tag{5}$$

Where $\alpha(t)$ is the learning rate and $h_{vi}(t)$ is the neighborhood kernel at time t. Both $\alpha(t)$ and $h_{vi}(t)$ decrease monotonically with time within 0 and 1. The neighborhood kernel $h_{vi}(t)$ is a function defined over the lattice points according to Equation (6). It represents the adapted range on the map and decreases with $||w_v(t)-w_i(t)||$. Finally, the winner is only adapted at the end of the training process. A widely applied neighborhood kernel can be written in terms of the Gaussian function, where r_v and r_i are the position of winner neuron and neighborhood neuron on the map, respectively. $\sigma(t)$ is the kernel width and decreases with time.

$$h_{vi}(t) = \exp\left(- \frac{\| r_v - r_i \|^2}{2\sigma^2(t)} \right) \tag{6}$$

Step 5. Update the weights of neighborhood neurons according to Equation (7)

$$w_k(t+1) = w_k(t) + \alpha(t) \times h_{vk}(t) \times \begin{cases} \left([x(t) - w_v(t)] + [w_v(t) - w_k(t)] \left(\frac{d_{vk}}{\Delta_{vk} \lambda} - 1 \right) \right), & \text{if } w_v(t) \text{ between } x(t) \text{ and } w_k(t) \\ \left([x(t) - w_v(t)] - [w_v(t) - w_k(t)] \left(\frac{d_{vk}}{\Delta_{vk} \lambda} - 1 \right) \right), & \text{if } w_k(t) \text{ between } x(t) \text{ and } w_v(t) \\ \left([x(t) - p] + [p - w_k(t)] \left(\frac{d_{vk}}{\Delta_{vk} \lambda} - 1 \right) \right), & \text{otherwise} \end{cases} \tag{7}$$

Where d_{vk} and Δ_{vk} are the distances between neurons v and k in the data space on the map, respectively, and λ is a positive pre-specified resolution parameter. It represents the desired inter-neuron distance that is reflected in the input

space and varies with the size of the map and data variance, and requires resolution of the map.

Step 6. Refresh the map randomly and choose the neuron weight, which is the input at a small percentage of updating time.

Step 7. Repeat Steps 2-6 until the map converges.

Step 8. Determine whether to generalize numeric attributes. If numeric attributes are not to be generalized, their averages and deviations will be computed in Step 10.

Step 9. Determine whether to generalize categorical attributes. For each categorical attribute A_i to be generalized in W, determine first whether A_i should be removed; and if not, determine its minimum desired generalization level L_i in its concept hierarchy. Construct its major-value set Mi according to θ and β. For $v \in \text{Dom}(A_i)$, if $v \notin M_i$, construct the mapping pair as (v, v_{Li}, M_{Li}); otherwise, as (v, v).

Step 10. Derive the generalized relation P by replacing each value v with its mapping value and computing other aggregate values.

4. Experiments

The architecture is developed by using Borland C++ Builder 6, Access database. In the experiments, it presents the results of the CHPD mining architecture in the financial time series database. The database is segmented to the empirical stock indices, which is the Taiwan Stock Exchange Corporation (TSEC). These original datasets cover the daily closing prices from 1/1/2000 to 12/31/2003. The training datasets are from 1/1/2000 to 12/31/2001. Index-based investment alternatives have surfaced recently.

Among the index tracking stocks, various types of iPOD, PSP and GPS stocks are the most popular. For stock indices, each index can be limited to the types of these stocks. These companies in Taiwan include iPOD, PSP and GPS companies. The GPS companies include Atech (亞元), Eten (倚天), MiTAC (神達), Leadtek (麗臺)；the PSP companies include MiTAC (神達), ASUS (華碩), I-SHENG (鎰勝), AVC (奇鋐), YUFO (育富), Cyber TAN (建漢)；the iPOD companies include PowerTech (力成), JI-HAW (今皓), Abo (友旺), Foxlink (正崴), Mustang (同協電子), AVID (合邦), Porolific (旺玖), ENIght (英誌), TRIPOD (健鼎), ACON (連展), Transcend (創見), GENESYS (創惟), ASUS(華碩), Etron(鈺創), Milestones (銘異), Foxconn (鴻海), APCB (競國).

This study uses ESA to find patterns in the time series database. It sets the sliding window within 5 days and the accumulated returns are more than 10%. The number of patterns with support≥0.1 extracted from each cluster is shown in Table 4. The results

obtained by the ESA indicate that all the clusters, clusters 1 and 2, have major values in product class, date, trading volume, turnover rate, size in circulation and ratio of book to market as dataset attributes. For example, iPOD is the major value of product class in clustering. Between 2000 and 2001, the result of cluster 1 shows that iPOD has lower price stock like MiTAC. The turnover rate is the highest, meaning that investors buy it more frequently. The size in circulation and the ratio of book to market are the smallest. The result of cluster 2 shows that iPOD has the highest price stock like Foxconn and ASUS. The size in circulation and the ratio of book to market are the highest because these stocks are controlled by someone. It seems more stable in this situation.

Table 4: ESA rule extraction in iPOD, PSP and GPS industries.

Date 2000 - 2001, θ=3, Gen Rela =20, β=1, sliding window within 5 days and the accumulated returns more than 10%							
Name	Date	Price (u, σ)	TV (u, σ)	TR (u, σ)	SIC (u, σ)	RBM (u, σ)	Support
IF C1	2 rules						
iPOD	2000	(36.7;23.7)	(3070.7;3377.4)	(2.4;2.1)	(117.8;61.5)	(0.3;0.3)	0.731
GPS	2000	(31.3;20.9)	(2446.8;2649.0)	(2.9;2.8)	(82.4;35.0)	(0.2;0.2)	0.238
IF C2	2 rules						
鴻海	2000	(115.2;22.3)	(8882.3;4093.0)	(0.7;0.3)	(1403.7;277.6)	(2.8;0.4)	0.56
華碩	2000	(114.9;26.5)	(11905.0;4733.3)	(0.8;0.3)	(1554.7;242.9)	(2.7;0.5)	0.44

Note: TV (Trading Volume), TR (Turnover rate), SIC (Size in circulation), RBM (ratio of book to market), u (mean), σ (standard error).

In time series data, we find a particular problem is plagues by multi-collinearity. There are several other remedies that solve multicollinearity problems like Factor analysis (FA) or Principal Component Analysis (PCA) and ridge regression. In this study, we use the PCA to solve the multicollinearity [Hair et al., 1998; Gujarati, 1999]. The principal component analysis model is a popular tool for exploratory data analysis or, more precisely, for assessing the dimensionality of sets of items.

PCA model was used in each case to identify the underlying components, which were then rotated to obtain the final solution. An oblique rotation was used because the underlying components were expected to be correlated. The rotated matrix is principal component analysis. Correlation matrices for price, trading volume, turnover rate, size in circulation and book to market equity are presented in table 5, and their loadings are listed in table 6. A common rule of thumb for assessing construct validity is that individual items should have a highest loading. For attribute independence, the PCA models are Rising_Prin1, Rising_Prin2, Falling_Prin1 and Falling_Prin2. Rising_Prin1 and Rising_Prin2 are the models of sliding window within 5 days and the accumulated returns

more than 10%. Falling_Prin1 and Falling_Prin2 are the models of sliding window within 5 days and the accumulated returns more than -10%.

Rising_Prin 1 = 0.352* price + 0.102 * trading volume + -0.110 * turnover rate + 0.343 * size in circulation + 0.376 * book to market equity

Rising_Prin 2 = -0.070 * price + 0.578 * trading volume + 0.577 * turnover rate + 0.116 * size in circulation + -0.079 * book to market equity

Falling_Prin 1 = 0.305 * price + 0.155 * trading volume + -0.093 * turnover rate + 0.335 * size in circulation + 0.357 * book to market equity

Falling_Prin 2 = -0.009* price + 0.508 * trading volume + 0.678 * turnover rate + 0.001 * size in circulation + -0.080 * book to market equity

Table 5: Correlation matrixes for price, trading volume, turnover rate, size in circulation and book to market equity.

	Price	TV	TR	SIC	BME
Price	1				
TV	0.203(**)	1			
TR	-0.035(**)	0.421(**)	1		
SIC	0.601(**)	0.491(**)	-0.171	1	
BME	0.797(**)	0.298(**)	-0.183(**)	0.853(**)	1

Note: TV (Trading Volume), TR (Turnover rate), SIC (Size In Circulation), BME (Book to Market Equity)

Table 6: Loadings of price, trading volume, turnover rate, size in circulation and book to market equity items.

Loadings		
Item	Rising_Prin 1	Rising_Prin 2
Price	0.819	-0.167
Trading volume	0.422	0.802
Turnover rate	-0.127	0.846
Size in circulation	0.917	0.094
Book to market equity	0.949	-0.190

Note: Sliding window within 5 days and the accumulated returns more than 10%

Loadings		
Item	Falling_Prin 1	Falling_Prin 2
Price	0.834	-0.052
Trading volume	0.453	0.711
Turnover rate	-0.125	0.931
Size in circulation	0.918	0.021
Book to market equity	0.975	-0.089

Note: Sliding window within 5 days and the accumulated returns more than -10%

For measuring the accuracy with pattern reappearing in test dataset, we proposed to use the accuracy rate and error rate as the accuracy with pattern reappearing. In the rising pattern, the accuracy rate is 76% and error rate is 24%. In the falling pattern, the accuracy rate is 81% and error rate is 19%. The experiment results about accuracy rate and error rate are shown in Table 7.

Table 7: Pattern reappearing results calculated with accuracy rate and error rate.

		Predicted		
		Rising pattern	Falling_ pattern	Total
Observed	Rising_ pattern	6021(76%)	1899(24%)	7920
	Falling_ pattern	1393(19%)	5830(81%)	7223
Total		7414	7729	

Logistic regression model can achieve high accuracy when linear variation of the data is small, however, if the variation range exceeds a certain limit, the linear models has a lower performance in estimated accuracy. In traditional method, this study discusses the use of logistic regression in financial database. This study uses ESA to find patterns in the time series database and verifies the pattern repeatable. Hegadeesh and Titman [Jegadeesh & Titman, 1993] showed that over a 3-12 month period, past winners (positive price or earnings momentum) outperform past losers. In this study, the sinner portfolios are the outstanding return from 1/1/2000 to 12/31/2001. Then it buys the stock between 2002 and 2003. The architecture uses the ESA algorithm to train the clustering rule. These clustering rules are used to classify the test data and calculate the returns. Finally, it compares the returns with the winner and loser portfolios. Table 8 provides the sample performance of the portfolio that meets most of these clustering rules. The experimental results show that the ESA can find time series patterns. The performance is better than logistic regression model and winner-loser portfolio.

Table 8: The compare table with clustering investment returns of Winner portfolio, Loser portfolio, Logistic Regression and ESA pattern.

Year	Item	Winner portfolio	Loser portfolio	Logistic Regression	ESA pattern
2002-2003	CR	-289%	1126%	3353%	9305%
	AV	-11%	45%	134%	372%

Note: *CR (Calculated Returns)*
AR (Average Returns)

In this study, there are evidences to support the relevance of the CHPD architecture model to informative pattern discovery. It also develops a novel, visual CHPD architecture prototype to prove the feasibility of the proposed model and to show better support to what-if analysis for the investor.

5. Conclusions and Future work

A decision support system is a computer program application that analyzes business data and presents it. So users can make business decisions more easily. CHPD mining architecture is an ideal decision support system to simulate human intelligence in finding patterns. Intelligent CHPD mining architecture can be employed to extract and generalize knowledge for major values which was hidden in the financial database for investors. The results show the feasibility of framework and the superiority of intelligent techniques.

In strategy contribution, accompany by the advent of the knowledge-based economy, it has became a rather critical issue of how to manage the knowledge well. The meaning of knowledge management is base in knowledge obtain, integrate, accumulate, share, transfer and innovation. It will get better effect in enhance knowledge management, if we can put these work into effect specifically through the time series database. In this study, we design a CHPD system based concept hierarchy of using the data rule extraction and data integrate characteristic in time series datasets. It can approach the knowledge sharing through rule extraction. We also give this platform the capability of self-learning, this capability let it can approach the innovation and creation of knowledge to improve the time series resource integrate and enhance the knowledge management.

In academic contribution, this paper is the study and research in financial and management literatures and for the theory and methodology of neural, clustering and classification literatures. This study builds an intelligent mining architecture that can extract and generalize knowledge for major values which hidden in the data available in the time series database for the investor.

In business contribution, CHPD system can extract the rules from historical data and describe the major values in attributes. The role of CHPD in the business is like a strategic information system and a knowledge extraction system in financial and management field. It also shows better support in what-if analysis for the investor.

There are several directions in which the present architecture can be utilized in future studies.

1. Apply CHPD to deal with an efficient association algorithm in the financial database for finding the frequent item sets.
2. Develop a knowledge innovation model, which consists of CHPD and many associated innovation issues or factors, like business fundamental index, news index or risk index to improve the performance in stock forecast. It can also incorporate other knowledge discovery technologies to build up a financial investment decision support system.

References

1. Agrawal, R., Imielinski, T. and Swami, A., "Mining association rules between sets of items in large databases," Proceedings of the ACM SIGMOD Conference on Management of Data 1993, pp: 207-216.
2. Becanovi, V., "Image object classification using saccadic search, spatio-temporal pattern encoding and self-organization," *Pattern Recognition Letters* (21:3) 2000, pp: 253-263.
3. Bollerslev, T., "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics* (31) 1986, pp: 307-327.
4. Brennan, M. J., Chordia, T. and Subrahmanyam, A., "Alternative factor specifications, security characteristics, and the cross-section of expected stock returns," *Journal of Financial Economics* (49) 1998, pp: 345-373.
5. Cai, Y., Cercone, N., and Han, J., *Attribute-oriented induction in relational databases*, 1991.
6. Carpenter, G. A. and Grossberg, S., "ART 3: Hierarchical search using chemical transmitters in self-organizing pattern recognition architectures," *Neural Networks* (3) 1990, pp: 129-152.
7. Chen, D. R., Chang, R. F. and Huang, Y. L., "Breast cancer diagnosis using self-organizing map for sonography," *Ultrasound in Medicine and Biology* (1:26) 2000, pp: 405-411.
8. Chen, M. S., Han, J. and Yu P. S., "Data mining: An overview from a database perspective," *IEEE Transactions on Knowledge and Data Engineering* (8) 1996, pp: 866-883.
9. Chen, S. H. and He, H., "Searching financial patterns with self-organizing maps," *Computational Intelligence in Economics and Finance* 2003, Springer.
10. Chen, S. H. and Tsao, C. Y., "Self-organizing maps as a foundation for charting or geometric pattern recognition in financial time series," Proceedings of 2003 International Conference on Computational Intelligence for Financial Engineering 2003, pp: 20-23.
11. Deboeck, G. J. and Alfred, U., "Picking Stocks with Emergent Self-organizing Value maps," *Neural Networks World* (10) 2000, pp: 203-216.
12. Deboeck, G. J. and Kohonen, T., *Visual Explorations in Finance with self-organizing maps*, Springer-Verlag, 1998.
13. Ester, M., Kriegel, H. P., and Sander, J., "Spatial data mining: A database approach," Proceedings of the fifth international symposium on spatial databases 1997, pp: 47-66.

14. Fama, E. F. and French, K., "The cross section of expected stock returns," *Journal of Finance* (47) 1992, pp: 427-465.
15. Fama, E. F. and French, K., "Size and book-to-market factors in earning and return," *Journal of Finance* (50) 1995, pp: 131-155.
16. Fayyad, U. M., Piatetsky, S. G. and Matheus, C. J., *Knowledge Discovery in Databases: An Overview*, 1991.
17. Fayyad, U. M., Piatetsky, S. G., Smyth, P. and Uthurusamy, R., *Advances in Knowledge Discovery and Data Mining*, 1996.
18. Fu, Y., Sandhu, K., and Shih, M. Y., "Clustering of web users based on access patterns," Proceedings of the 1999 KDD Workshop on Web Mining, 1999.
19. Guha, S., Rastogi, R. and Shim, K., "ROCK: A robust clustering algorithm for categorical attributes," Proceedings of the IEEE Conference on Data Engineering 1999, pp: 512-521.
20. Gujarati, D., *Essentials of Econometrics*, 2nd edition, McGraw-Hill, 1999.
21. Gunter, S. and Bunke, H., "Self-organizing map for clustering in the graph domain," *Pattern Recognition Letters* (23:4) 2002, pp: 405-417.
22. Hair, J. F., Anderson, R. E., Tatham, R. L., Black, W. C., *Multivariate Data Analysis*, Fifth Edition, Prentice-Hall International, Inc., New Jersey, 1998.
23. Han, J. and Kamber, M., *Data mining concepts and techniques*, San Francisco: Morgan Kaufmann, 2001.
24. Han, J., Cai, Y., and Cercone, N., "Knowledge discovery in databases: an attribute-oriented approach," Proceedings of the 18th VLDB conference, British Columbia, Vancouver, 1992, pp: 547-559.
25. Han, J., Cai, Y., and Cercone, N., "Data-driven discovery of quantitative rules in relational databases," *IEEE Transactions on Knowledge and Data Engineering* (5) 1993, pp. 29-40.
26. Hertz, J., Krogh, A., and Palmer, R. G., *Introduction to the Theory of Neural Computation*, Santa Fe Institute Studies in the Sciences of Complexity lecture notes. Addison- Wesley Longman Pub. Co., Inc., Reading, MA, 1991.
27. Hsu, C. C. and Wang, S. H., "An Integrated Framework for Visualized and Exploratory Pattern Discovery in Mixed Data," *IEEE Transactions on Knowledge and Data Engineering* (18:2) 2005, pp: 161-173.
28. Hsu, C. C., "Extending Attribute-Oriented Induction Algorithm for Major Values and Numeric Values," *Expert Systems with Applications* (27:2) 2004, pp: 187-202.
29. Huang, Y. P., Hsu, C. C and. Wang, S. H., "Pattern Recognition in Time Series Database: A Case Study on Financial Database," *Expert Systems with Applications*, (33) 2007, pp: 199-255.

30. Imielinski, T. and Mannila, H., "A database perspective on knowledge discovery," *Communications of ACM* (39) 1996, pp: 58-64.
31. Jain, A. K. and Mao, J., "Neural networks and pattern recognition," *Computational Intelligence: Imitating Life* 1994, pp: 194-212.
32. Jain, A. K., Murty, M. N., and Flynn P. J., "Data Clustering: A Review," *ACM Computing Surveys* (31:3) 1999, pp: 264-323.
33. Jegadeesh, N. and Titman, S., "Return to buying winners and selling losers," *Journal of Finance* (48) 1993, pp: 65-91.
34. Kiang, M. Y., Kulkarni, U. R. and Tam, K. Y., "Self-organizing map network as an interactive clustering tool — An application to group technology," *Decision Support Systems* (15:4) 1995, pp: 351-374.
35. Kiang, M., Michael, Y., Hu, Y. and Fisher, D. M., "An extended self-organizing map network for market segmentation—a telecommunication example," *Decision Support Systems*, (42:1) 2006, pp: 36-47.
36. Kohonen, T., *Self-organization and associative memory*, Springer Verlag, 1984.
37. Kohonen, T., "Engineering applications of the self-organizing map," Proceedings of the IEEE (84:10) 1996, pp: 1358-1384.
38. Koperski, K., Han, J., and Adhikary, J., "Mining knowledge in geographical data," *Communications of the Association For Computing Machinery* (26:1) 1998, pp: 65-74.
39. Kramer, A. A., Lee, D. and Axelrod, R. C., "Use of a Kohonen Neural Network to Characterize Respiratory Patients for Medical Intervention," *Artificial Neural Networks in Medicine and Biology*, 2000, pp: 192-196.
40. Kulkarni, U. R. and Kiang, M. Y., "Dynamic grouping of parts in flexible manufacturing systems-A self-organizing neural networks approach," *European Journal of Operational Research* (84) 1995, pp: 192-212.
41. Kuo, S. C., Li, S. T., Cheng, Y. C., and Ho, M. H., "Knowledge Discovery with SOM Networks in Financial Investment Strategy," The 4th International Conference on Hybrid Intelligent Systems (IEEE Press), 2004.
42. Laaksonen, J., Koskela, M., Laakso, S. and Oja, E., "PicSOM – content-based image retrieval with self-organizing maps," *Pattern Recognition Letters* (21:13), 2000, pp: 1199-1207.
43. Li, S. T. and Kuo, S. C., "Discovering Financial Investment Strategy through Wavelet-based SOM Networks," The 10th Annual Meeting of Asia-Pacific Decision Sciences Institute (APDSI 2005), 2005.
44. Li, S. T., "Leveraging a Web-aware SOM Tool for Clustering and Visualization," The First Asia-Pacific Conference on Web Intelligence (WI-2001), 2001.

45. Smith, K. A. and Ng, A., "Web page clustering using a self-organizing map of user navigation patterns," *Decision Support Systems* (35:2) 2003, pp: 245-256.
46. Toivanen, P. J., Ansamaki, J., Parkkinen, J. P. and Mielikainen, S. J., "Edge detection in multispectral images using the self-organizing map", *Pattern Recognition Letters* (24:16) 2003, pp: 2987-2994.
47. Tsai, F. C., Kuo, S. C. and Li, S. T., "Crime Trend Discovery using Fuzzy SOM Networks," The 11th Asia Pacific Management Conference APMC-2005, 2005.
48. Vesanto, J. E., Alhoniemi, J., Himberg, K. K. and Parviainen, J., "Self-organizing map for data mining in Matlab: the SOM Toolbox," *Simulation News Europe* 1999, pp: 25-54.
49. Wilson, D.R. and Martinez, T. R., "Improved heterogeneous distance functions," *Journal of Artificial Intelligence Research* (6) 1997, pp: 1-34.
50. Wu, S. and Chow, T. W. S., "Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density," *Pattern Recognition* (37:2) 2004, pp: 175-188.
51. Yin, H., "Data visualization and manifold mapping using the ViSOM," *Neural Networks* (15) 2002, pp: 1005-1016.
52. Yin, H., "ViSOM - a novel method for multivariate data projection and structure visualization," *IEEE Transactions on Neural Networks* (13:1) 2002, pp: 237-243.
53. Zaiane, O. R., Han, J., Li, Z. N., Chee, S. H., and Chiang, J. Y., "Multimedia miner: A system prototype for multimedia data mining," Proceedings of the ACM SIGMOD international conference on management of data, 1998, pp: 581-583.