Pattern recognition in time series database: A case study on financial database

Yan-Ping Huang a,b,*, Chung-Chian Hsu a, Sheng-Hsuan Wang a

a Department of Information Management, National Yunlin University of Science and Technology, 123, Sec. 3, University Road, Douliu, Yunlin 640, Taiwan, R.O.C.
b Department of Management Information System, Chin Min Institute of Technology, 110, Hsueh-Fu Road, Tou-Fen, Miao-Li 351, Taiwan, R.O.C.

Abstract

Today, there are more and more time series data that coexist with other data. These data exist in useful and understandable patterns. Data management of time series data must take into account an integrated approach. However, many researches face numeric data attributes. Therefore, the need for time series data mining tool has become extremely important. The purpose of this paper is to provide a novel pattern in mining architecture with mixed attributes that uses a systematic approach in the financial database information mining. Time series pattern mining (TSPM) architecture combines the extended visualization-induced self-organizing map algorithm and the extended Naïve Bayesian algorithm. This mining architecture can simulate human intelligence and discover patterns automatically. The TSPM approach also demonstrates good returns in pattern research.

Keywords: Time series analysis; Portfolio strategy; Data mining; Cluster analysis; Self-organizing map; Classification analysis

1. Introduction

Knowledge discovery in database is a nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in the data. Data mining is a step of knowledge discovery process consisting of particular algorithms to produce patterns. 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, Imielinski, & Swami, 1993; Han & Kamber, 2001).

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. How to find a pattern from the time series datasets and how to prove the pattern be useful become more and more important. These new methods for knowledge discovery and data mining in time series datasets are based on unsupervised neural networks and specifically on self-organizing maps (SOM) (Kohonen et al., 1981). These applications of SOM are focused on how to find a pattern in a particular database. However, there are still problems. When a database has numeric attributes and categorical attributes, the traditional approach cannot preserve and present these mixed attributes. 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 paper addresses this issue by proposing a framework for time series datasets.

This paper describes a novel pattern mining architecture in the financial database. The time series pattern mining architecture combines the extended visualization-induced self-organizing map (EViSOM) algorithm (Wang & Hsu, 2005) and the extended Naïve Bayesian (ENB) algorithm.

* Corresponding author. Address: Department of Information Management, National Yunlin University of Science and Technology, 123, Sec. 3, University Road, Douliu, Yunlin 640, Taiwan, R.O.C. Tel.: +88 637627153; fax: +88 697605684.
E-mail addresses: sunny@chinmin.edu.tw (Y.-P. Huang), hsucc@mis.yuntech.edu.tw (C.-C. Hsu), g9223722@yuntech.edu.tw (S.-H. Wang).

(Chang & Hsu, 2005) to automatically discover the patterns in the financial database. The TSPM mining architecture is an ideal tool for simulating the human intelligence in finding or creating patterns that summarizes and stores useful aspects of our perceptions.

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

In the time-domain models, Engle (1982) described ARIMA models (autoregressive integrated moving average process) (Fama & French, 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. The feature-based model (Chen & He, 2003; Chen & Tsao, 2003) selected relevant prior observations based on the symbolic or geometric characteristics of the time series in question, rather than their location in time.

Most algorithms for pattern finding nearly were variations of the SOM algorithm (Chen, Chang, & Huang, 2000; Kohonen, 1981; Kramer, Lee, & Axelrod, 2000) the SOM had several applications, including financial forecasting and management (Kohonen, 1996; Vesanto, Alhoniemi, Himberg, & Parviainen, 1999) as well as medical diagnosis (Chen et al., 2000; Deboeck & Kohonen, 1998). The SOM, proposed by Kohonen, was an unsupervised neural network, which projected high-dimensional data onto a low-dimensional grid. In recent years, many researchers tried to use SOM algorithm or other related techniques to discover the patterns from a huge financial database to support investors to make decisions (Chen & He, 2003; Chen & Tsao, 2003; Deboeck & Alfred, 2000; Deboeck & Kohonen, 1998).

A non-linear multi-dimensional projection method had been proposed from SOM and the visualization-induced SOM (named ViSOM) (Yin, 2002a, 2002b). The objective of ViSOM was to preserve the data structure and the topology as faithfully as possible. The ViSOM considered the distance between two neurons (winner and neighborhood) in the data space and on the map, respectively, and used a resolution parameter that controlled the inter-neuron distance on the map (Hsu, 2004). The EViSOM algorithm integrates the concept hierarchies such that the extended system properly handles the mixed data (Hsu, 2004). The EViSOM algorithm is described in Section 3.2.

Naive Bayesian classifier has been widely used in data mining as a simple and effective classification algorithm (Fayyad & Irani, 1993). Naive Bayesian classifiers have been proven successful in many domains, despite the simplicity of the model and the restrictiveness of the independent assumptions it made. Naive Bayesian algorithm handles only categorical data, but could not reasonably express the probability between two numeric values and preserve the structure of numeric values. Extended Naive Bayesian algorithm is used in data mining as a simple and effective classification algorithm. The ENB algorithm has integrated the concept hierarchies such that the extended system could properly handle the categorical data.

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. The EViSOM algorithm is used to calculate the distance between the categorical and numeric data. The extended Naive Bayesian algorithm integrates the concept hierarchies such that the extended system properly handles the mixed data. This paper uses EViSOM algorithm to discover the pattern and ENB algorithm to prove the pattern be profitable or informative in the TSPM architecture. Therefore, this article provides a brief architecture to find the pattern, which is defined by the user’s request returns and prove that the pattern is profitable.

3. Time series pattern mining architecture

3.1. A pattern mining model for time series dataset

Time series pattern mining architecture provided four steps of stock information mining. These steps included: preprocess, pattern recognition analysis, pattern evaluation analysis and comparison evaluation. Fig. 1 shows the TSPM architecture.

1. Preprocess: There are two general steps. First, it removes noise and inconsistent data. Second, it retrieves the relevant data from the database and transforms the data format for pattern recognition.

2. Pattern recognition: It applies the intelligent methods to extract data patterns, and tries to identify the patterns representing knowledge based on EViSOM algorithm. The EViSOM algorithm is described in Section 3.2.

3. Pattern evaluation: It evaluates the target pattern features or classes from source classes. The source class is specified by the EViSOM algorithm, whereas the target pattern class is specified by the ENB algorithm. The ENB algorithm is described in Section 3.3.

4. Comparison evaluation: Patterns are referred to as actionable. Patterns can be used in strategy. Measuring the interest in patterns is essential for the efficient discovery of the value of patterns.

3.2. Example

The time series datasets based on the financial dataset are implemented in Table 1. 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 1 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 TSPM architecture. Table 2 shows an example of original financial dataset.

About attributes selection, this paper chooses the attributes like return, trading volume, momentum, book to market, and the ratio of book to market. About the attributes of book to market and the ratio of book to market attributes, Fama and French (1992; Fama and French (1995) described the portfolio strategies and found out that size and book-to-market factors could affect stock returns (Fama & French, 1995; Jegadeesh & Titman, 1993). Jegadeesh and Titman (1993) described that momentum strategies could affect the stock returns (Brennan, Chordia, & Subrahmanyam, 1998). Brennan et al. (1998) described the trading volume factor and it is the cross-section of the expected stock returns (Hsu, 2006). Therefore, this article uses these core stock indices like income stocks, return, trading volume, momentum, book to market, and the ratio of book to market as dataset attributes.

Table 1
The example of time series dataset format

<table>
<thead>
<tr>
<th>No.</th>
<th>$att_1$</th>
<th>$att_2$</th>
<th>$att_3$</th>
<th>$att_4$</th>
<th>$att_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_{11}$</td>
<td>174.4</td>
<td>17243</td>
<td>4.53</td>
<td>1.5</td>
<td>1146</td>
</tr>
<tr>
<td>$x_{12}$</td>
<td>170.4</td>
<td>7926</td>
<td>-2.31</td>
<td>0.69</td>
<td>1146</td>
</tr>
<tr>
<td>$x_{13}$</td>
<td>169.9</td>
<td>8497</td>
<td>-0.3</td>
<td>0.74</td>
<td>1146</td>
</tr>
<tr>
<td>$x_{14}$</td>
<td>166.3</td>
<td>6184</td>
<td>-2.08</td>
<td>0.54</td>
<td>1146</td>
</tr>
<tr>
<td>$x_{15}$</td>
<td>174.9</td>
<td>15577</td>
<td>5.15</td>
<td>1.36</td>
<td>1146</td>
</tr>
<tr>
<td>$x_{21}$</td>
<td>170.37</td>
<td>7926</td>
<td>-2.3</td>
<td>0.69</td>
<td>1146</td>
</tr>
<tr>
<td>$x_{22}$</td>
<td>169.87</td>
<td>8497</td>
<td>-0.3</td>
<td>0.74</td>
<td>1146</td>
</tr>
<tr>
<td>$x_{23}$</td>
<td>166.34</td>
<td>6184</td>
<td>-2.1</td>
<td>0.54</td>
<td>1146</td>
</tr>
<tr>
<td>$x_{24}$</td>
<td>174.91</td>
<td>15577</td>
<td>5.15</td>
<td>1.36</td>
<td>1146</td>
</tr>
<tr>
<td>$x_{25}$</td>
<td>172.39</td>
<td>18476</td>
<td>-1.4</td>
<td>1.61</td>
<td>1146</td>
</tr>
</tbody>
</table>

Table 2
The example of financial dataset

<table>
<thead>
<tr>
<th>Id</th>
<th>Stocks ID</th>
<th>Stocks name</th>
<th>Date</th>
<th>Price</th>
<th>Trading volume</th>
<th>Return</th>
<th>Momentum</th>
<th>Book to market</th>
<th>Book to market ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>5118</td>
<td>2315</td>
<td>神達 (MiTAC)</td>
<td>2000/1/4</td>
<td>36.86</td>
<td>41313</td>
<td>2.29</td>
<td>5.38</td>
<td>37759</td>
<td>0.30</td>
</tr>
<tr>
<td>5119</td>
<td>2315</td>
<td>神達 (MiTAC)</td>
<td>2000/1/5</td>
<td>35.81</td>
<td>29787</td>
<td>-2.85</td>
<td>3.88</td>
<td>36685</td>
<td>0.29</td>
</tr>
<tr>
<td>5120</td>
<td>2315</td>
<td>神達 (MiTAC)</td>
<td>2000/1/6</td>
<td>35.81</td>
<td>39324</td>
<td>0.00</td>
<td>5.12</td>
<td>36685</td>
<td>0.29</td>
</tr>
<tr>
<td>5121</td>
<td>2315</td>
<td>神達 (MiTAC)</td>
<td>2000/1/7</td>
<td>36.11</td>
<td>51504</td>
<td>0.84</td>
<td>6.71</td>
<td>36992</td>
<td>0.29</td>
</tr>
</tbody>
</table>
Time series pattern mining architecture provided four steps of stock information mining. These steps included: preprocess, pattern recognition analysis, pattern evaluation analysis and comparison evaluation.

Step 1. **Preprocess:** In this stage, the trading signals by the sliding-window calculate the return for days more than the given threshold. The threshold can be defined by the user or the investor. The investor can define which performance pattern liked by the user. Such that \( f(x) = \sum_{i=1}^{n} (k), i \) time interval and \( j \) is the dataset in each interval data. If \( f(x) > \text{threshold} \), then it uses these datasets to predict based on pattern recognition analysis.

Step 2. **Pattern recognition:** This step applies the EViSOM algorithm to extract data patterns. During training, the EViSOM forms an elastic net that folds onto the “cloud” formed by the input data. Data points lay near each other in the input space that are mapped onto the nearby map units. Thus, the EViSOM can be interpreted as a topology preserving mapping from input space onto the 2-D grid. After the EViSOM algorithm, it can identify the patterns representing knowledge and generates the rules of clustering. If \( \text{Sim}(x, BNU) < \text{threshold} \), then the cluster lever is 1 else 0.

Step 3. **Pattern evaluation:** It evaluates the target pattern features or classes from source classes. This step uses EViSOM algorithm to train and generate the clustering rules. It also uses the test data set to improve the pattern profitable or informative. The target pattern class is specified by the ENB algorithm.

Step 4. **Comparison evaluation:** This step uses test data set with ENB algorithm to find which stock can be bought. There are two classification levels – to buy or not to buy from the ENB algorithm. If the class level is buy single, then it supports buying in the stock opening quotation and selling it in closing quotation, otherwise, do nothing. Since this paper does not have intraday data, it cannot know the effects of the bid-ask spreads on return calculations. Obviously, a judicious use of limit orders would be warranted in attempting to implement such a trading strategy. And it calculates the return to check the pattern profitable or informative. It compares the portfolio strategy form TSPM mining architecture with winner and loser portfolio strategy. The winner and loser portfolio strategy is described by Jegadeesh and Titman (1993), Brennan et al. (1998).

3.3. Extended visualization-induced self-organizing map algorithm

The EViSOM algorithm consists of a regular, usually two-dimensional (2-D), grid of map units. Each unit is represented by a prototype vector, which is input vector dimension. The units are connected to the adjacent by a neighborhood relation. The EViSOM clustering algorithms usually have the following steps:

Step 1. It pre-processes task-relevant records in the dataset. Step 2. It initializes the map or weights either to the principal components or to small random values. Each of the neurons 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 calculates the Euclidean distances between \( x(t) \) and all neurons. A winning neuron \( w_v \) can be found according to the minimum distance to \( x(t) \). Find the winner neuron such that

\[
\nu = \arg \min_i \| x(t) - w_i(t) \|, \quad i \in \{1, \ldots, M\}. 
\]

Step 3. 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 such that

\[
w_v(t + 1) = w_v(t) + \alpha(t) \times h_v(t) \times [x(t) - w_v(t)],
\]

where \( \alpha(t) \) is the learning rate and \( h_v(t) \) is the neighborhood kernel at time \( t \), respectively. Both \( \alpha(t) \) and \( h_v(t) \) decrease monotonically with time within 0 and 1. The neighborhood kernel \( h_v(t) \) is a function defined over the lattice points.

Step 4. Update the weights of neighborhood neurons such that

\[
w_k(t + 1) = w_k(t) + \alpha(t) \times h_k(t)
\]

\[
\times \begin{cases} \left[ |x(t) - w_v(t)| + \left| w_i(t) - w_k(t) \right| \left( \frac{d_{vk}}{d_{vk}^*} - 1 \right) \right], & \text{if } w_v(t) \text{ between } x(t) \text{ and } w_k(t) \\ \left[ |x(t) - w_v(t)| - \left| w_i(t) - w_k(t) \right| \left( \frac{d_{vk}}{d_{vk}^*} - 1 \right) \right], & \text{if } w_k(t) \text{ between } x(t) \text{ and } w_v(t) \\ \left[ |x(t) - p| + |p - w_k(t)| \left( \frac{d_{vk}}{d_{vk}^*} - 1 \right) \right], & \text{otherwise} \end{cases}
\]

where \( d_{vk} \) and \( A_{vk} \) are the distances between neurons \( v \) and \( k \) in the data space on the map, respectively, and \( \lambda \) is a positive pre-specified resolution parameter. It represents the desired inter-neuron distance that reflects in the input space and
depends on the size of the map, data variance, and requires resolution of the map.

Step 5. Refresh the map randomly and choose the neuron weight, which is the input at a small percentage of updating time.

Step 6. Repeat steps 2–5 until the map converges.

3.4. Extended Naïve Bayesian classification algorithm

The ENB algorithm has been widely used in data mining as a simple and effective classification algorithm. It integrates concept hierarchies such that the extended system properly handles the mixed data. For a categorical attribute, the conditional probability that an instance belongs to a certain class \( c \) given that the instance has an attribute value \( A = a \), \( P(C = c | A = a) \) is given by

\[
P(C = c | A = a) = \frac{P(C = c \cap A = a)}{P(A = a)} = \frac{n_{ac}}{n_a}
\]

where \( n_{ac} \) is the number of instances in the training set which has the class value \( c \) and an attribute value of \( a \), while \( n_a \) is the number of instances which simply has an attribute value of \( a \). Due to horizontal partitioning of data, each party has partial information about every attribute. Each party can locally compute the local count of instances. The global count is given by the sum of the local counts. For a numeric attribute, the necessary parameters are the mean \( \mu \) and variance \( \sigma^2 \) for all different classes. Again, the necessary information is split between the parties. In order to compute the mean, each party needs to sum the attribute values of the appropriate instances having the same class value. These local sums are added together and divided by the total number of instances having the same class to get the mean for that class value.

The ENB clustering algorithm usually has the following steps:

Step 1. A training dataset requires \( x_i \) parties, \( C_k \) class values and \( w \) attribute values. If there occur numeric data attributes, calculate the mean value and the variance value. However, the time is counted if categorical data attributes occur.

Step 2. It calculates the probability of an instance having the class and the attribute value. If there occur numeric data attributes, the probability of attribute value \( x_i \) in class \( C_k \) determines the probability such that

\[
p(x_i | C_k) = p(C_k) \prod_{j=1}^{attj} p(w_{ij} | C_k) \prod_{j=atti+1}^{attj} 2
\]

\[
\times p \left( z \geq \frac{\left( \bar{x}_j - \bar{x}_j' \right)}{\sqrt{\frac{\sigma_j^2}{n_j} + \frac{\sigma_j'^2}{n_j'}}} \times m \right).
\]

If there occur categorical data attributes, the probability of attribute value \( x_i \) domain value \( w_{ij} \) in class \( C_k \) determines the probability such that

\[
P(C_i | x) > P(C_j | x), \ x \ is \ in \ class \ C_i;
\]

else \( x \) is in class \( C_j \).

The Bayesian approach to classify the new instance is to assign the most probable target value, \( P(C_i | x) \), given the attribute values \( \{w_1, w_2, \ldots, w_m\} \) that describe the instance.

\[
P(C_i | x) = \frac{P(C_i \cap x)}{P(x)} = \frac{P(x | C_i)P(C_i)}{P(x)}.
\]

The ENB classifier makes the simplification assumption that the attribute values are conditionally independent given the target value. Therefore,

\[
P(x | C_i) = \prod_{j=1}^{n} P(x_j | C_i).
\]

The categorical values have been normalized before calculating the probability of each class. It determines the normalized equation such that

\[
P(w_{ij} | C_k) = \frac{1 + \sum_{c \in C_k} N(w_{ij}, x_i)p(C_k | x_i)}{|V| + \sum_{c \in C_k} \sum_{i=1}^{n} N(w_{ij}, x_i)p(C_k | x_i)},
\]

where \( |V| \) is the total domain value in the attribute value \( x_i \).

Step 3. All parties calculate the probability of each class, such that

\[
p_i = \prod_{j=1}^{m} p(\hat{j})
\]

Step 4. It selects the maximal of the probability of each class such that:

\[
\max_{i=1}^{\text{classCount}} (p(\hat{i})).
\]

4. Experiments and results

The architecture is developed using Borland C++ Builder 6, access database. In the experiments, it presents the results of the TSPM 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. There are 125,000 observations and test datasets from 1/1/2001 to 12/31/2003. Index-based investment alternatives have surfaced recently.

Among the index tracking stocks, various types of 3P stocks (iPOD, PHS and GPRS) are the most popular. For stock indices, each index can be limited to the types of 3P stocks. The 3P companies in Taiwan include iPOD, PHS and GPRS companies. The GPRS companies include Atech.
This paper considers the problem of finding out the pattern rules from a large time series database. It proposes an efficient TSPM mining architecture, for exploring the patterns. This architecture combines EViSOM algorithm to find patterns and ENB algorithm to improve the pattern reusable. In order to provide the functionality of visualization to a decision maker, this project utilizes the self-organization map that transforms high-dimensional, complicated, and nonlinear data into low-dimensional maps with topology preservation. For clustering, the TSPM architecture is applied to validate the clusters. It not only endeavors to improve the accuracy in the classification of trading signals, but also attempts to maximize the profits of trading. This architecture can also apply in other time series databases, like medical databases.

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

1. TSPM mining architecture will combine some other associated innovation issues or factors, like business fundamental index, news index, technology index, business or risk index to improve the performance in stock forecast.
2. It will propose an efficient association algorithm in the financial database to find the frequent item sets.
3. It will combine knowledge discovery technology to build up a financial investment decision support system.

Table 4

<table>
<thead>
<tr>
<th>Year</th>
<th>The winner portfolio</th>
<th>The loser portfolio</th>
<th>The EViSOM pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculated</td>
<td>2000–2001</td>
<td>1501%</td>
<td>−2164%</td>
</tr>
<tr>
<td>the returns</td>
<td>2002–2003</td>
<td>−289%</td>
<td>1126%</td>
</tr>
<tr>
<td>Summarized</td>
<td>1212%</td>
<td>−1038%</td>
<td>14664%</td>
</tr>
</tbody>
</table>

and verifies the pattern repeatable. Jegadeesh and Titman (1993) showed that over a 3–12 months period, past winners (positive price or earnings momentum) outperform past losers. In this study, the winner 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 ENB 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 4 provides the sample performance of the portfolio that meets most of these clustering rules. The experimental results show that the EViSOM can find time series patterns. These performances are better than winner–loser portfolio.

In this paper, there are evidences to support the relevance of the TSPM architecture model to informative pattern discovery.

5. Conclusions and future work

This paper considers the problem of finding out the pattern rules from a large time series database. It proposes an efficient TSPM mining architecture, for exploring the patterns. This architecture combines EViSOM algorithm to find patterns and ENB algorithm to improve the pattern reusable. In order to provide the functionality of visualization to a decision maker, this project utilizes the self-organization map that transforms high-dimensional, complicated, and nonlinear data into low-dimensional maps with topology preservation. For clustering, the TSPM architecture is applied to validate the clusters. It not only endeavors to improve the accuracy in the classification of trading signals, but also attempts to maximize the profits of trading. This architecture can also apply in other time series databases, like medical databases.

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

1. TSPM mining architecture will combine some other associated innovation issues or factors, like business fundamental index, news index, technology index, business or risk index to improve the performance in stock forecast.
2. It will propose an efficient association algorithm in the financial database to find the frequent item sets.
3. It will combine knowledge discovery technology to build up a financial investment decision support system.

Fig. 2. The stocks clustering map in iPOD, PHS and GPRS industries.
References


