computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

A NEW ALGORITHM FOR TIME SERIES DATA MINING BY USING ROUGH SET

Fei Hao [1], Ling Hei Yeung [2]

[1] Korea Advanced Institute of Science and Technology
373-1, guseong-Dong, Yuseong-Gu, Daejeon 305-701, Korea
fhao@islab.kaist.ac.kr
[2] Hong Kong Baptist University, Hong Kong
lightisgood2005@yahoo.com.hk

Abstract: *This paper is to apply Rough Set to data mining of time series. Firstly, we process the time series data by attribute selection and similarity sequence search. Secondly, the time series is partitioned into some sets of pattern by Mobile Window Method (MWM) and each pattern is a trend of time series. Thirdly, an information table is made by predicting attributes and targeting attribute in trending variation ratio structure sequence (TVRSS). Then, the original information table is made suitably for rough set to discover knowledge. Finally, the extracting rules can predict the time series behavior in the future. The total process is four steps. In the end, we show some examples to demonstrate our method on the time series data of stock market.*

Keywords: *TVRSS, Time series, Rough Set, Prediction.*

1. INTRODUCTION

Time-series data is becoming more and more vital in much field, especially in economic and finance area. With broad application of the information technology and diversification approach of obtaining data, time series information is increasing rapidly. People often try to mine the potential knowledge and useful information in some effective method or technologies when they confront a great deal of time series data. For example, consumers often make use of historical stock closing price data to predict future closing price. The procedure of converting historical data to useful knowledge or information for human is paid more attention in modern society.

Time-series data mining is an important way which mining some useful and potential knowledge from a great deal of time-series [1]. More and more people began to pay more attention to time-series data mining [2, 3, 4]. Nowadays, the study about time-series mining is mainly about similarity mining and time sequences query. Due to nonlinear, aperiodic and chaos of time series, it is difficult to predict future precision for human. But in a certain situation, it is possible to predict some potential information in time-series. For example, if you have necessary information about trend of stock closing price. We can predict the approximately trend in the future in spite of we can't predict the precise closing price.

In this paper, we mainly discussed the time-series data prediction based on TVRSS and rough set. Section 2 is devoted to introducing time-series and trending variation ratio structure sequence. Section 3 deals with some basic notions related to rough set theory (RST). In section 4, a novel time-series prediction method based on TVRSS and a time-series data mining model based on rough set are proposed respectively. Section 5 is experimental analysis. Conclusion is section 6.

2. TIME SERIES AND TVRSS

Time series is a series of observation data according to a certain time sequence [5]. It is a aggregate which has time and event. Time series is a data set in the planar or multidimensional space. Time-series data mining is a important way which mining some useful and potential knowledge from a great deal of time-series [6].

Time series possess characteristic as follows:

1. In a planar time series X, a certain point x_t is composed of abscissa time T and y-axis X (t, x) in planar space $T \times X$.
2. Time series is not reversible.
3. Information transfer is unilateral and not reversible; Information transfer direction is same as

the time. For example, a time series X={x_t|t=1, 2,…, n}. Information will transfer with $x_{t1}$to x_{t2} ($t_1<t_2$).

Definition 1 [5] *For a planar time series X={x_t|t=1, 2, …,n}, a certain point x_t is composed of abscissa time T and y-axis X (t, x) in planar space T × X. We can write time series in vector form*

$$X=\{x_1, x_2, \ldots, x_n\}$$

Definition 2 Let X={x_t|t=1, 2,…,n} be time series, λ={λ_1, λ_2,…, λ_{n-1}}.λ is a trending variation ratio structure sequence of X. In which, $\lambda i=\lambda i=\frac{\triangle^{i}_{x}}{\triangle^{i}_{T}}$, $\triangle^{i}_{x}=x_{i+1}-x_i$, $\triangle^{i}_{t}=t_{i+1}-t_i$ Obviously, λ is also a time series.

$$\lambda=\left\{\frac{\triangle^{1}_{x}}{\triangle^{1}_{T}},\frac{\triangle^{2}_{x}}{\triangle^{2}_{T}},\ldots,\frac{\triangle^{n-1}_{x}}{\triangle^{n-1}_{T}}\right\} \quad (1)$$

in which, I={1, 2,…, n-1}.

Definition 3 Let X = {x_t|t=1, 2, …, n} be time series, λ={λ_1, λ_2, …, λ_{n-1}} be trending variation ratio structure sequence of X. (x_i, x_{i+1},…, x_{i+k-1}), (λ_j, λ_{j+1},…, λ_{j+k+1}) (I=1, 2, …, n-k+1; j=1, 2, …, n-k) are called k=time sub series for X and λ respectively. {λ_h, λ_{h+1},…, λ_{h+k-1}} is a trending variation ratio structure sequence of time sub series {x_h, x_{h+1}, …, x_{h+k-1}} in λ. In which, $\lambda_m=\frac{\triangle^{m}_{x}}{\triangle^{m}_{T}}$, m=h, h+I,…, h+k-1 (h+k-1<n), xi∈X; {x_h, x_{h+1},…, x_{h+k-1}} is a latest time sub series which length is k of X.

k-time sub series is defined as follows,

$$S(X, k)=\{(x_i, x_{i+1},\ldots, x_{i+k-1})|(i=1,\ldots, n-k+1)\} \quad (2)$$

$$S(\lambda, k)=\{(\lambda_j, \lambda_{j+1},\ldots, \lambda_{j+k-1})|(j=1,2,\ldots, n-k)\} \quad (3)$$

Here, |S(X, *k*)|=n-k+1, S(λ, *k*)=n-k.

Definition 4 Suppose *s_1, s_2∈S(X, k)-{(x_{n-k+1}, x_{n-k+2},…, x_n)}, ι_1, ι_2∈S(λ,k), l_1,l_2 are trending variation ratio structure sequence of s_1,s_2 respectively. If d(ι_1, ι_2)=0, then the trending variation ratio structure of s_1 is same as s_2. In which, d(ι_1, ι_2) denotes the Euclid distance, d(ι_1, ι_2)=$\sum_i^k|\lambda_{1i}-\lambda_{2i}|$, ι_1={λ_{11}, λ_{12}, …, λ_{1k}}, ι_2={λ_{21}, λ_{22}, …, λ_{2k}}.*

Definition 5 *Suppose X={x_t|t=1, 2,…,n} be time series, $\triangle^{i}_{x}=x_{i+1}-x_i$, $\triangle^{i}_{T}=t_{i+1}-t_i$, i=1, 2, …, n-1, $\lambda=\left\{\frac{\triangle^{1}_{x}}{\triangle^{1}_{T}},\frac{\triangle^{2}_{x}}{\triangle^{2}_{T}},\ldots,\frac{\triangle^{n-1}_{x}}{\triangle^{n-1}_{T}}\right\}$ is the trending variation ratio structure sequence of X, then the trending variation ratio structure sequence of X are redefined as follows according to various $\frac{\triangle^{i}_{x}}{\triangle^{i}_{T}}$,*

$$\lambda^k=\begin{cases}\left(\left\{\frac{\triangle^{i}_{x}}{\triangle^{i}_{T}}|\frac{\triangle^{i}_{x}}{\triangle^{i}_{T}}>0\right\}\right)\\ \left(\left\{\frac{\triangle^{i}_{x}}{\triangle^{i}_{T}}|\frac{\triangle^{i}_{x}}{\triangle^{i}_{T}}=0\right\}\right)\\ \left(\left\{\frac{\triangle^{i}_{x}}{\triangle^{i}_{T}}|\frac{\triangle^{i}_{x}}{\triangle^{i}_{T}}<0\right\}\right)\end{cases} \quad i=1, 2, \ldots, n-1, k=1, 2, 3 \quad (4)$$

in which, $\triangle^{i}_{T}>0$ corresponding to ascending trending structure sequence λ^1. $\triangle^{i}_{T}=0$ corresponding to ascending trending structure sequence λ^2. $\triangle^{i}_{T}<0$ corresponding to ascending trending structure sequence λ^3.

3. DECISION-MAKING INFORMATION SYSTEM IN ROUGH SET THEORY

Information is a cognitive result about external world and a decision-making principle about human behavior. People can make some useful decision-making by historical knowledge, especially in prediction field and reasoning technology. Information system is a abstract description of database. Data in information system is usually denoted by relation table which row denotes objects and column denotes attributes. Information of objects is expressed by attribute value. It is an important tache between objects and attribute. It is also an information foundation for knowledge discovery.

Definition 6 [8] *Let U be a non-empty, finite set of objects called the universe, R be an equivalence relation on U. The pair (U, R) as approximation space. And a set of equivalent classes with respect to R is as follows*

$$U/R=\{[x_i]_R|x_i\in U\},$$

and where $[x_i]=\{x_j|(x_i, x_j)\in R\}$.

Definition 7 [8] *Information System (IS) is a quaternary presentation*

$$S=(U, A, V, f)$$

in which, U denotes a non-empty, finite set of objects called the universe, it is expressed as U={x_1, x_2, …, x_n}. A denotes a noe-empty, finite set of attribute it is expressed as A={a_1, a_2,…, a_m}. V is range of attribute. V={v_1, $_2$,…, v_m} where v_i is a value of a_i. f is an information function, f: U × A→V, $f(x_i, a_i)\in V_i$

Definition 8 [8] Decision Information System (DIS) is a quaternary presentation

$$S=(U, C\cup D, V, f)$$

where $A=C \cup D$ and $C\cap D\neq 0$ in which, C is the set of conditional attribute and D is the set of decision attribute.

Definition 9 [8] Each non-empty subset $p\subseteq A$ determines an indiscernibility relation as follows:

$$IND(p)=\{(x, y)\in U\times U|f_i(x)=f_i(y), a_i\in p\},$$

then IND(p) is an equivalence relation on U and let:

$$[x_i]_{IND(p)}=\{x_j|(x_i,x_j)\in IND(p)\}.$$

Property 1 IND(p) is an equivalence relation on U and $IND(p)=I_{a\in p}IND(a)$

Definition 10 Let S={U, A, V, f} be a decision-making information system, where $A=C\cup D$ and $C\cap D\neq 0$ $X_i=U/C$, $Y_j=U/D$ Decision-making rules are defined as follows:

$$r_{ij}: des(X_i)\rightarrow des(Y_j) \quad X_i\cap Y_j\neq 0 \qquad (5)$$

in which $des(X_i)$ denotes the description of each object on conditional attributes, and $des(Y_j)$ denotes the description of each object on decision-making attribute.

Corollary 1 Reliability gene of each rule is defined as follows,

$$\mu(X_i, Y_j)=|X_i\cap Y_j|/|X_i|$$

where, $0\leq\mu(X_i, Y_j)$

if $\mu(Xi,Yj)=1$, then r_{ij} is certain;

if $0<\mu(Xi,Yj)<1$, then r_{ij} is uncertain

Example 1 Give a decision-making table of insurance investigation in TaiWan (see Table 1), in which U={person1, person2,…, person8}={*x*1, *x*2,…, *x*8}, A={area, occupation, age, salary, insurance}={a1, a2, a3, a4, a5},

C={area, occupation, age, salary}={a1, a_2, a_3, a_4},

D={insurance}={a_5}

In above insurance investigation, conditional attributes and decision attribute are described as follows,

area={north, south}={1, 2};
Professional={worker, professor}={1, 2};
Age={young, middleage, old}={1, 2, 3};
Salary={low, high}={1, 2};
Insurance={living, poverty, house}={1, 2, 3};
Xi=U/C={{x1, x3}, {x2}, {x4, x5, x7}, {x6, x8}}
Yi=U/D={{x1, x2, x3}, {x4, x5, x7}, {x6, x8}}

For example, according to Eq. (4), we can obtain some useful decision-making rules easily:

Table 1. Insurance investigation

person	area	occupation	age	salary	insurance
x_1	1	1	3	2	1
x_2	2	1	2	1	1
x_3	1	1	3	2	1
x_4	1	2	2	1	2
x_5	1	2	2	1	2
x_6	1	2	1	1	3
x_7	1	2	2	1	2
x_8	1	2	1	1	3

Rule 1 If a person who lived in north of Taiwan, is a old worker, and have high salary, then he or she will buy living insurance;

Rule 2 If a person who lived in south of Taiwan and is middle age worker, **and** have low salary, then he or she will buy living insurance;

Rule 3 If a person who lived in north of Taiwan and is old worker, and have low salary, then he or she will buy living insurance;

Rule 4 If a person who lived in north of Taiwan and is middle age professor, and have high salary, then he or she will buy poverty insurance;

According to corollary 1, above rules are certain, because of their $\mu(X_i, Y_j)=1$.

4. TIME-SERIES DATA PREDICTION BASED ON TVRSS AND ROUGH SET

After introduced the basic concept of time series and decision-making information system in rough set theory. In this section, a model of time-series data mining based on rough set is established firstly, then a novel time-series data prediction method based on trending variation ratio structure sequence and rough set is proposed.

4.1. TIME-SERIES DATA MINING MODEL BASED ON ROUGH SET

Generally, data mining procedure is divided into steps as follows,

1. Establish data mining goal
2. Analysis and prepare data
3. Establish mining model according correlative algorithm
4. Evaluate mining model
5. Implement data mining

Without exception, time-series data mining model based on rough set contains data collection, data pretreatment, data reduction, extracting rules, classification and prediction [11].

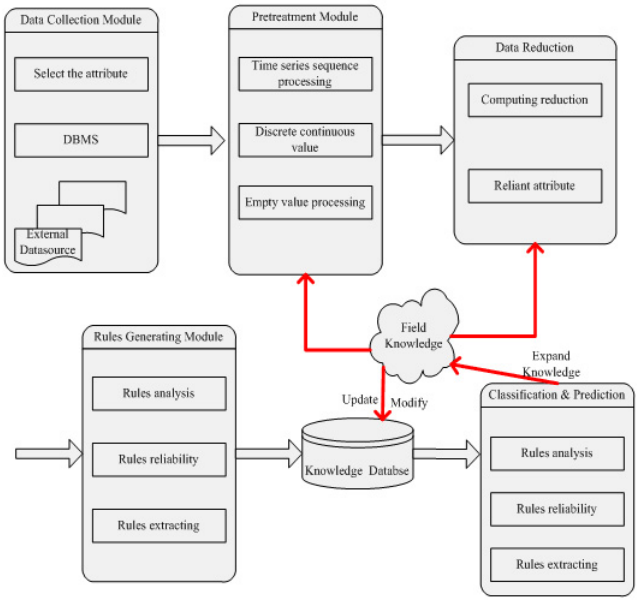


Fig. 1 – Time-series data mining model based on rough set

4.2. CONVERSION FROM TVRSS TO DECISION IN-FORMATION SYSTEM

Suppose time series $X=\{x_1,x_2,\ldots,x_n\}$, trending variation ratio structure sequence $\lambda=\{\lambda_1,\lambda_2,\ldots,\lambda_{n-1}\}$ Conversion step as follows:

step 1 Data standard pretreatment for λ by mobile window method [9].

step 2 Make trending variation ratio structure sequence convert into standard time-series data stylebook set.

step 3 Make standard time-series data stylebook set convert into decision-making information system.

Suppose mobile window size=k, we can construct trending variation ratio structure information system according to above steps, (see Table 2). Here, trending variation ratio structure information system reflects feature sub sequence property. So, we can analyze the time-series data with feature sub sequence.

Table 2. TVRSS information system

U\A	a_1	a_2	…	a_{k-2}	a_{k-1}	a_k
1	λ_1	λ_2	…	λ_{k-2}	λ_{k-1}	λ_k
2	λ_2	λ_3	…	λ_{k-1}	λ_k	λ_{k+1}
3	λ_3	λ_4	…	λ_k	λ_{k+1}	λ_{k+2}
4	λ_4	λ_5	…	λ_{k+1}	λ_{k+2}	λ_{k+3}
…	…	…	…	…	…	…
n-k	λ_{n-k}	λ_{n-k+1}	…	λ_{n-3}	λ_{n-2}	λ_{n-1}

Table 3. TVRSS predictable information system

U\A	a_1	a_2	…	a_{k-2}	a_{k-1}	a_k
1	λ_1	λ_2	…	λ_{k-2}	λ_{k-1}	λ_k
2	λ_2	λ_3	…	λ_{k-1}	λ_k	λ_{k+1}
3	λ_3	λ_4	…	λ_k	λ_{k+1}	λ_{k+2}
4	λ_4	λ_5	…	λ_{k+1}	λ_{k+2}	λ_{k+3}
…	…	…	…	…	…	…
n-k	λ_{n-k}	λ_{n-k+1}	…	λ_{n-3}	λ_{n-2}	λ_{n-1}
s_q	λ_{n-k+1}	λ_{n-k+2}	…	λ_{n-2}	λ_{n-1}	q

4.3. ONE-STEP PREDICTION BASED ON TVRSS PREDICTABLE INFORMATION SYSTEM

In time series $X=\{x_1,x_2,\ldots,x_n\}$, One-step prediction is that predict X_{n+1} value. But, it can't predict precise value, only predict the trend on time n+1. Time-series trending variation ratio structure in latest mobile window is $\{\lambda_{n-k+1}, \lambda_{n-k+2},\ldots,\lambda_{n-1}\}$, decision attribute value q is unknown, decision attribute value q is the prediction aim(q=-1,0,1) (see Table 3).

In Table3, different trending variation ratio depicts different pattern. These values are continuous, it is necessary to discrete these values according to its variation ratio. Here, fuzzy discrete method is adopted.

4.3.1 DATA PRETREATMENT

Definition 11 Suppose $X=\{x_t|t=1,2,\ldots,n\}$ be time series, the trending variation ratio structure $\lambda=(\frac{\Delta^1 X}{\Delta^1 T},\frac{\Delta^2 X}{\Delta^2 T},\ldots,\frac{\Delta^{n-1} X}{\Delta^{n-1} T})$, the membership functions of trending variation ratio on, $\lambda_1,\lambda_2,\lambda_3$ are defined respectively as follows,

$$\mu_{ascend_{rapid}}=\begin{cases}0, & |\lambda_i|\le b_1\\ \dfrac{|\lambda_i|-|b_1|}{a_1-b_1}, & b_1\le|\lambda_i|\le a_1\\ 1, & |\lambda_i|\ge a_1\end{cases}$$

$$\mu_{ascend_{moderate}}=\begin{cases}0, & |\lambda_i|\le b_2\\ \dfrac{|\lambda_i|-|b_2|}{a_2-b_2}, & b_2\le|\lambda_i|\le a_2\\ 1, & |\lambda_i|\ge a_1\end{cases}$$

$$\mu_{ascend_{slow}}=\begin{cases}0, & |\lambda_i|\le b_3\\ \dfrac{|\lambda_i|-|b_3|}{a_3-b_3}, & b_3\le|\lambda_i|\le a_3\\ 1, & |\lambda_i|\ge a_1\end{cases}\quad(6)$$

In Eq. (6), $\lambda_i\in\lambda^1$

$$\mu_{equal} = 1 \quad c \le |\lambda_i| \le d \qquad (7)$$

Here, $\lambda_i \in \lambda^2$

$$\mu_{descend_{rapid}} = \begin{cases} 0, & |\lambda_i| \le f_1 \\ \dfrac{|\lambda_i| - |f_1|}{e_1 - f_1}, & f_1 \le |\lambda_i| \le e_1 \\ 1, & |\lambda_i| \ge e_1 \end{cases}$$

$$\mu_{descend_{moderate}} = \begin{cases} 0, & |\lambda_i| \le f_2 \\ \dfrac{|\lambda_i| - |f_2|}{e_2 - f_2}, & f_2 \le |\lambda_i| \le e_2 \\ 1, & |\lambda_i| \ge e_2 \end{cases}$$

$$\mu_{descend_{slow}} = \begin{cases} 0, & |\lambda_i| \le f_3 \\ \dfrac{|\lambda_i| - |f_3|}{e_3 - f_3}, & f_3 \le |\lambda_i| \le e_3 \\ 1, & |\lambda_i| \ge e_3 \end{cases} \qquad (8)$$

In Eq. (8), $\lambda_i \in \lambda^3$

In which, $a_1,a_2,a_3,b_1,b_2,b_3,c,d,e_1,e_2,e_3,f_1,f_2,f_3$ is defined by users

The trending variation ratio structure sequence $л = \left\{ \frac{\Delta^1 X}{\Delta^1 T}, \frac{\Delta^2 X}{\Delta^2 T}, \ldots, \frac{\Delta^{n-1} X}{\Delta^{n-1} T} \right\}$ is redefined according to above membership functions of trending variation radio as follows,

$$\tilde{\lambda} = (\mu_1^1, \mu_2^1, ..., \mu_7^1, ..., \mu_1^{n-1}, \mu_2^{n-1}, ..., \mu_7^{n-1}) \qquad (9)$$

Moreover, Eq. (9) is formalized according to the maximum membership grade principle like this,

$$\text{e.g. } \tilde{\lambda} = (0,1,0,0,0,0,0|0,0,0,0,1,0,0| \\ 0,0,0,1,0,0,0|1,0,0,0,0,0,0) \qquad (10)$$

Eq. (10) depicts the feature of trending on variation ratio. So, its feature is explained as follows,

$$ascend_{moderate} \rightarrow descend_{rapid} \rightarrow equal \rightarrow ascend_{rapid}$$

Obviously, redefined $\tilde{\lambda}$ is constituted by seven-dimensional vector space. However, we can get the trend of future time according to the $\tilde{\lambda}$ corresponding to its state space vector.

4.3.2 PATTERN MATCHING

In this paper, we consider each trending structure sequence is a pattern.Due to this idea, Hamming distance is adopted in this paper.

Definition 12 *Suppose a certain pattern* $l_{n-k} = (\lambda_{n-k}, \lambda_{n-k+1}, ..., \lambda_{n-1}) \in S(\lambda, k)$, *the trending structure state space vector of* l_{n-k} *is denoted as follows,*

$$\tilde{\lambda}^{n-k} = (p_{n-k1}, p_{n-k2}, ..., p_{n-kn-k+1}) \\ k=1,2...,n-k+1 \qquad (11)$$

Definition 13 Let $s_1, s_2 \in S(X,k) - \{(x_{n-k+1}, ..., x_n)\}, l_1, l_2, \in S(\lambda, k), l_1, l_2$ are the time sub series of $s_1 m, s_2$ respectively. $\tilde{\lambda}^{(1)}, \tilde{\lambda}^{(2)}$ are the trending structure state space vector of l_1, l_2. Hamming distance between $\tilde{\lambda}^{(2)}$ and $\tilde{\lambda}^{(2)}$ is defined as follows,

$$H(\tilde{\lambda}^{(1)}\tilde{\lambda}^{(2)} = \sum_{i=1}^{n-k+1} h_i$$

$$h_i = \begin{cases} 1, & p_i \ne p_2, \lambda_{1i}\lambda_{2i} > 0 \\ 0, & \tilde{\lambda}^{(1)} = \tilde{\lambda}^{(2)} \\ 2, & p_i \ne p_2, \lambda_{1i}\lambda_{2i} < 0. \end{cases} \qquad (12)$$

in which, $l_1 = (\lambda_{11}, ..., \lambda_{1k}), l_2 = (\lambda_{21}, ..., \lambda_{2k})$

Property 2 $0 \le H(\tilde{\lambda}^{(1)}, \tilde{\lambda}^{(2)}) \le 2k$

If $H(\tilde{\lambda}^{(1)}, \tilde{\lambda}^{(2)}) = 0$, then K-time sub sequence s_1 has the same trending variation ratio with s_2.

The smaller $H(\tilde{\lambda}^{(1)}, \tilde{\lambda}^{(2)})$ is, the better pattern matching is.

Here, we give a pattern difference measurement method, we can easily get the same or similar pattern with goal pattern (predictable sub time sequence). In a intuitionistic sense, $H(\tilde{\lambda}^{(1)}, \tilde{\lambda}^{(2)})$ reflects the synthetical similarity between l_1 and l_2 $i = 1, 2..., n-k$. So, obtaining the relationship and extracting useful rules among these patterns is a problem.

The pattern matching model is devised as follows,

$$Min\ H(s_q, \tilde{\lambda}^{(i)})$$

$$s.t.\ \lambda_{si}\lambda_{ji} > 0$$

$$i = 1,2,...n-k$$

4.3.3 EXTRACTING PREDICTABLE RULES

In fact, each object in Table3 denotes a known predictable rule. These predictable rules are wrote like this according to Definition 10.

$$eg.if\ \text{descend}_{\text{rapid}} \rightarrow ascend_{moderate} \\ \rightarrow ascend_{rapid} \rightarrow ascend_{slow} \rightarrow ascend_{moderate} \\ \text{the } decend_{slow}$$

In the trending variation ratio structure predictable information system (See table 3), q is our prediction aim, s_q is predictable pattern, we can

predict the q by conditional expression C={$\lambda_{n-k+1}, \lambda_{n-k+2}, \ldots, \lambda_{n-1}$} in s_q.

Definition 14 *Let Sim(x, y) be the function of the similarity about x and y. For prediction sequence s_q and l_i, its similarity is defined as follows*

$$Sim(s_q, l_i) = \frac{1}{H(s_q, l_i)} \tag{13}$$

in which q=-1,0,1;i=1,2,…,n-k.

Note 1 The smaller $H(s_q, l_i)$ is, the better $Sim(s_q, l_i)$ is, i.e, the trusty q is.

Remark: The feature of the algorithm of trending variation ratio structure sequence based on rough set is concluded as follows,

Consider the conditional attribute sets belong to in-discernibility equivalence relation on objects, but not other objects. Hence, calculated capacity is lesser.

Description of trend structure classification is subtle.

Fuzzy discrete method accords with human intuitivism.

4.4 ALGORITHMS

A: Algorithm of Data Conversion and Resolution on $S(\lambda, k)$

Step1: Input $X = \{x_1, x_2, ...x_n\}$

Step2: Computing TVRSS $\lambda = (\frac{\Delta_X^1}{\Delta_T^1}, \frac{\Delta_X^2}{\Delta_T^2}, ..., \frac{\Delta_X^{n-1}}{\Delta_T^{n-1}}) = (\lambda_1, \lambda_2, ..., \lambda_{n-1})$ according to $\lambda_i = \frac{\Delta_X^i}{\Delta_T^i}$.

Step3: $S(\lambda, k) \leftarrow \phi$

For (t=1; t<n-k; t++) {

$U(t) = (\lambda_1, \lambda_2, ...\lambda_{t+k-1})$;

$S(T_r, k) \leftarrow S(T_r, k) \cup \{U(t)\}$

}

Step4: output $S(\lambda, k)$

B: Adjacence Algorithm of Computing $S(\lambda, k-1)$ and l_{n-1}

Step1: Input $S(\lambda, k-1)$

$M \leftarrow 0$

Step2. For ($i = 1; i < |S(\lambda, k-1)|; i++$) {

$\tilde{\lambda} \leftarrow F(\lambda)$

For $\lambda^{n-k} \in \tilde{\lambda}$, calculating $H(s_q, \lambda^{n-k})$

$Min(H(s_q, \lambda^{n-k}))$

}

5. EXPERIMENT

As we all known, stock data is a typical time series database. Stock code, opening price and closing price for every day are kept in the stock database. Here, stock code is static attribute, but opening price and closing price are dynamic attributes which change with time. Investor pay more attention to the alteration of opening price and closing price. In this paper, closing price is considered and analyzed.

We make short-term closing price forecast of Shen zhen Developing Bank (China) stock data according to the 32 transaction data (from March 1, 2005 to April 13, 2005, see Table4) [10].

5.1 EXPERIMENT STEP

Step 1: After computing trending variation ratio structure sequence, we can compute the trending variation ratio structure sequence of stock time series according to definition 2. (see table 4)

The stock series is expressed as follows according to definition 1:

X={6.48, 6.40, 6.38, 6.44, 6.33, 6.38, 6.36, 6.28, 6.10, 6.12, 6.10, 5.99, 5.94, 6.0, 5.93, 5.99, 5.54, 5.37, 5.40, 5.38, 5.30, 5.30, 5.09, 5.20, 5.52, 5.42, 5.5, 5.7, 6.29, 6.8, 6.88, 6.7}

Table 4. Transaction data (t, p denote time and closing price respectively)

t	1	2	3	4	5	6	7	8
p	6.48	6.40	6.38	6.44	6.33	6.38	6.36	6.28
t	9	10	11	12	13	14	15	16
p	6.10	6.12	6.10	5.99	5.94	6.0	5.93	5.99
t	17	18	19	20	21	22	23	24
p	5.54	5.37	5.40	5.38	5.30	5.30	5.09	5.20
t	25	26	27	28	29	30	31	32
P	5.52	5.42	5.5	5.7	6.29	6.8	6.88	6.7

Trending variation ratio structure sequence of stock time series is,

λ ={-0.02, 0.06, -0.08, -0.04, -0.11, -0.06, -0.13, -0.04, -0.0133, -0.09, -0.06, 0.04, -0.04, 0.0033, -0.4, -0.18, 0.04, 0, -0.0333, -0.2, 0.11, 0.39, -0.06, 0.07, 0.12, 0.56, 0.62, 0.0967, -0.44, 0.24}

Step 2: Let mobile windows size k=5, k-time sub series are obtained according to definition 3. (see Table 5)

Step 3: Convert TVRSS into predictable decision-making information system according to table 3 (see Table 5). Here, q is our prediction aim.

Step 4: Fuzzy discrete trending variation ratio according to Data pretreatment method in section.4.3.1. (see table 6)

Step 5: Computing the Hamming distance between s_q and l, i.e, $H(s_q, l)$. Due to definition 13, $H(s_q, l)$ are solved as follows (see Table 7).

Step 6: Computing the maximal similarity degree of Hamming distance according to Eq. (13).

$$H(s_q, s_{25}) = \frac{1}{3}$$

Finally, we learnt similarity degree of Hamming distance according is maximal when $s = s_{25}$. So, the similarity is the best among all feature sub sequences. Hence, the rules of sequence s_{25} is inserted in knowledge database, helping make future prediction.

Because of above analysis, we can get the experimental result that the closing price at time 33 is higher than its former closing price. In fact, observing closing price at time 33 is higher than its former closing price. From the decision-making rule point of view, we can gain such rule as follow,

Decision-making rule If $A_1 = ascend_{\mathrm{mod}erate}, A_2 = ascend_{apid}, A_3 = descend_{\mathrm{mod}erate}, A_4 = ascend_{\mathrm{mod}erate}$, then $D = ascend_{\mathrm{mod}erate}$. i.e, If closing price of the stock market keeps rising for five days and declines next day, then the closing price will be higher than former closing price in the future.

Table 5. Trending prediction table of stock time series

	A_1	A_2	A_3	A_4	D
s_1	-0.02	0.06	-0.08	-0.04	0.11
s_2	0.06	-0.08	-0.04	0.11	-0.06
s_3	-0.08	-0.04	0.11	-0.06	-0.13
s_4	-0.04	0.11	-0.06	-0.13	-0.04
s_5	0.11	-0.06	-0.13	-0.04	-0.0133
s_6	-0.06	-0.13	-0.04	-0.0133	-0.09
s_7	-0.13	-0.04	-0.0133	-0.09	-0.06
s_8	-0.04	-0.0133	-0.09	-0.06	0.04
s_9	-0.0133	-0.09	-0.06	0.04	-0.04
s_{10}	-0.09	-0.06	0.04	-0.04	0.0033
s_{11}	-0.06	0.04	-0.04	0.0033	-0.4
s_{12}	0.04	-0.04	0.0033	-0.4	-0.18
s_{13}	-0.04	0.0033	-0.4	-0.18	0.04
s_{14}	0.0033	-0.4	-0.18	0.04	0
s_{15}	-0.4	-0.18	0.04	0	-0.0333
s_{16}	-0.18	0.04	0	-0.0333	0
s_{17}	0.04	0	-0.0033	0	-0.2
s_{18}	0	-0.0033	0	-0.2	0.11
s_{19}	-0.0033	0	-0.2	0.11	0.39
s_{20}	0	-0.2	0.11	0.39	-0.06
s_{21}	-0.2	0.11	0.39	-0.06	0.07
s_{22}	0.11	0.39	-0.06	0.07	0.12
s_{23}	0.39	-0.06	0.07	0.12	0.56
s_{24}	-0.06	0.07	0.12	0.56	0.62
s_{25}	0.07	0.12	0.56	0.62	0.0967
s_{26}	0.12	0.56	0.62	0.0967	-0.44
s_{27}	0.56	0.62	0.0967	-0.44	0.24
s_q	0.62	0.0967	-0.44	0.24	goal

Table 6. Trending prediction table of stock time series

	A_1	A_2	A_3	A_4	D
s_1	0000001	0010000	0000010	0000001	0100000
s_2	0010000	0000010	0000001	0100000	0000010
s_3	0000010	0000001	0100000	0000010	0000010
s_4	0000001	0100000	0000010	0000010	0000001
s_5	0100000	0000010	0000010	0000001	0000001
s_6	0000010	0000010	0000001	0000001	0000010
s_7	0000010	0000001	0000001	0000010	0000010
s_8	0000001	0000001	0000010	0000010	0010000
s_9	0000001	0000010	0000010	0010000	0000001
s_{10}	0000010	0000010	0010000	0000001	0010000
s_{11}	0000010	0010000	0000001	0010000	0000100
s_{12}	0010000	0000001	0010000	0000100	0000100
s_{13}	0000001	0010000	0000100	0000100	0010000
s_{14}	0010000	0000100	0000100	0010000	0000100
s_{15}	0000100	0000100	0010000	0000100	0000001
s_{16}	0000100	0010000	0000100	0000001	0001000
s_{17}	0010000	0000100	0000001	0001000	0000100
s_{18}	0000100	0000001	0001000	0000100	0100000
s_{19}	0000001	0001000	0000100	0100000	1000000
s_{20}	0001000	0000100	0100000	1000000	0000010
s_{21}	0000100	0100000	1000000	0000010	0100000
s_{22}	0100000	1000000	0000010	0100000	0100000
s_{23}	1000000	0000010	0100000	0100000	1000000
s_{24}	0000010	0100000	0100000	1000000	1000000
s_{25}	0100000	0100000	1000000	1000000	0100000
s_{26}	0100000	1000000	1000000	0100000	0000100
s_{27}	1000000	1000000	0100000	0000100	1000000
s_q	1000000	0100000	0000100	1000000	goal

Table 7. H(sq, si) table

s	$H(s_q, s_i)$	s	$H(s_q, s_i)$	s	$H(s_q, s_i)$
s1	2+1+1+2=6	s_2	1+2+1+1=5	s_3	2+2+2+2=8
s4	2+0+1+2=5	s_5	1+2+1+2=6	s_6	2+2+1+2=7
s7	2+2+1+2=7	s_8	2+2+1+2=7	s_9	2+2+1+1=6
s10	2+2+2+2=8	s_{11}	2+1+1+1=5	s_{12}	1+2+2+2=7
s13	2+1+0+2=5	s_{14}	1+2+0+1=4	s_{15}	2+2+2+2=8
s16	2+1+2+2=7	s_{17}	1+2+1+2=6	s_{18}	2+2+2+2=8
s19	2+2+0+1=5	s_{20}	2+2+2+0=6	s_{21}	2+0+2+2=6
s22	1+1+1+1=4	s_{23}	0+2+2+1=5	s_{24}	2+0+2+0=4
s25	1+0+2+0=3	s_{26}	1+1+2+1=5	s_{27}	0+1+2+2=5

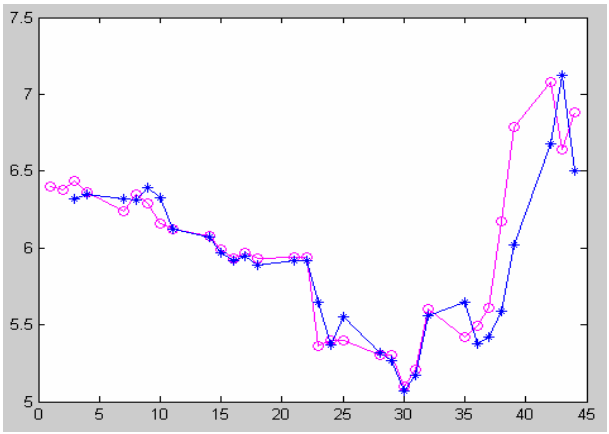
5.2 EXPERIMENT RESULT

In this paper, we utilized different Mobile window size to calculating the prediction result and accuracy.

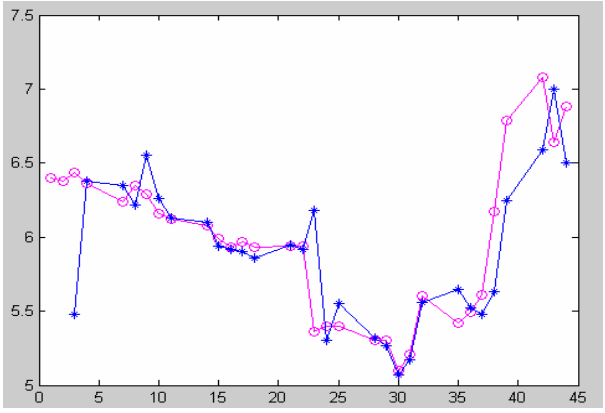
Table 8. Pattern results for different mobile window sizes

Factors / Size	Hamming Distance	Matching No	Prediction Results	Confidence
$k=3$	0+1=1	2	Descend moderate	50%
	0+1=1		Descend slow	50%
$k=4$	1+1+1=3	2	Ascend moderate	50%
	1+1+1=3		Ascend rapid	50%
$k=5$	1+1+1+1 =4	1	Ascend moderate	100%
$k=6$	#	0	#	
$k=7$	#	0	#	

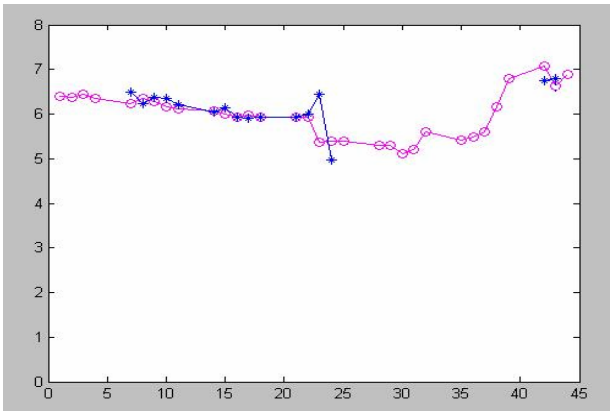
So, we can easily get the optimal mobile window size=3, and the prediction results (see Fig2.(a)).



(a) Size=3



(b) Size=4



(c) Size=5

Fig. 2 – Actual curve and prediction curve for different mobile window sizes

6. CONCLUSION

This paper mainly discussed the time-series data prediction based on trending variation ratio structure sequence and rough set. To obtain various k-time series, mobile window method is adopted. As a matter of fact, each sub time series is pattern. Trending variation ratio structure sequence is applied to sub time series in order to get important attribute which depict each sub time series. To predict unknown data or information in the future, trending variation ratio structure predictable information system is constructed. Hamming distance is adopted in pattern identification. Our method is applied to stock market data prediction.

7. REFERENCES

[1] Shao Fengjing, Yu Zhongqing. *Principle and Algorithm of Data Mining,* Water conservancy & water electric press of China, Beijing, 2003.

[2] P.Lee J, S.Kim. "Trend Similarity and Prediction in Time-series Databases [A]". *In: Proc. of SPIE on Data Mining and Knowledge Discovery: Theory, Tools, and Technology II. Washington: SPIE,* 2000,pp.201-212

[3] R.Agrawal, C.Faloutsos, A.Swami. "Efficient Similarity Search in Sequence Database". *Springer Verlag,* 1993, pp.69-84.

[4] G.Das, D.Gunopulous, H.Mannila. "Finding Similar Time Series", *In: Proc. of 1st European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD97),* Komorowski J, Zytkow J (Eds.), 1997.

[5] Z.Stefan. Data Mining for Prediction. Financial Series Case. *Doctoral thesis, Department of Computer and System Sciences.* The Royal Institute of Technology, 2003.

[6] R.J.Bayardo Jr., R.Agrawal. "Mining the Most Interesting Rules", *In: Proc. of the 5th ACM*

SIGKDD Inter. Conf. on Knowledge Discovery and Data Mining, August 1999.

[7] Yong Wang, Time-series data mining study and its application in prediction of water quality. *Doctor degree dissertation, Department of automatization.* Guangdong university of technology, 2005.

[8] Z. Pawlak, "Rough sets", *International Journal of Computer and Information Sciences,* vol.11, 1982, pp.341-356.

[9] J.K.Baltzersen. An attempt to predict stock data: a rough sets approach. *Diploma thesis, Konwledge Systems Group, Department of Computer Systems and Telematics,* The Norwegian Institute of Technology, University of Trondheim, 1996.

[10] http://quote.stock.163.com/h/data/dayhtml/000001.htm

[11] Keyun Hu, *Research and design of knowledge discovery system based on rough set,* HeFei university of technology, 1998.

Fei Hao *received the bachelor's degree & Master's degree in school of Mathematics & Computer Engineering from Xihua university in 2005 and 2008 respectively. Currently he is a PHD candidate in department of Computer Science, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea.*

His research interests include intelligent information processing, rough sets theory and time series data mining, data stream.

Ling Hei Yeung *graduated 2004 ~ 2007 Bsc (Hons) in Computing Mathematics, Department of Mathematics, City University of Hong Kong. 2007~ 2009 Double Msc in Operational Research and Business Statistics, Jointly by Hong Kong Baptist University and Kent University of UK.*

Distinction: Dean's List, 2005, Fall Semester, City University of Hong Kong.

Research Interests: Operational Research and Business Statistics, Data Mining.