I.J. Engineering and Manufacturing, **2012,2, 23-28**
Published Online April 2012 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijem.2012.02.04

Available online at http://www.mecs-press.net/ijem

Application of AC Algorithm Based on RS in Stock Index Prediction

Xiaoguang Wang [a,*], Fuxian Liu [a], Hui Liu [a], Fei Ma[a]

[a] *Missile Institute, Air Force Engineering University, Sanyuan ,China*

Abstract

The AC Algorithm may easily get a lower pattern similarity when performing the AC under the situation of encountering multi-dimensional data, so this will affect the selection of similar patterns. Combining the Rough Set theory, the author makes the data dimension reduction processing. The experiment shows that the AC algorithm based on RS is practical and its performance efficiency and prediction accuracy are much higher than the AC algorithm.

Index Terms: Analog Complexing Algorithm, Pattern Similarity, Rough Set

1. Introduction

The stock data are collected according to the order of time, so it is a time sequence with the features of complicated fluctuations, strong uncertainty and nonlinea[1]. Research showed that stock data have the same chaotic characteristics as the meteorological data[2]. We can utilize these data's rich time sequence patterns to describe the system's dynamics, but it is very complicated to describe the system like the stock market because of the higher data dimension. Therefore, the document [7] adopts to RS theory to preprocess the data and reduce them as the input of BP neural network，and this method greatly improves the BP network's prediction efficiency and accuracy. But BP network's structure decides that it is a gradient descent algorithm in essence, so it has some disadvantages, such as poor convergence in the learning process, easy to trap in local minimum. So, it is not suitable to classify and predict the complicated and big system. The Analog Complexing algorithm developed by Lorence[3] was firstly applied in the weather prediction. AC algorithm can be considered as a sequence pattern recognition method to predict, cluster and classify the complicated objects. This method is based on the following hypothesis：the classic situations of the time process can be

* Corresponding author.
E-mail address: wxg0108018@163.com

repeated with some certain forms. For a designated multi-dimensional time process, the current phase of its development status in the history may have one or more similar phases. Thus, we can use the known development status in the history to alternate and composite to get a future development status of the current status and then to get the prediction.

When using the AC algorithm to perform the data prediction, the research object shall satisfy one of the most important conditions---the multi-dimensional process is fully representative, viz the data sets consist of the system's basic variables[4].How to select the most representative basic variables in system holds the key to the success of AC algorithm prediction. This article makes use of the RS theory to make the attribute reduction processing of the multi-dimensional data in stock market and the attribute reduction results are the most representative system's basic variables. Then, the author relies on the AC algorithm to predict the stock market.The experiment result shows that the AC algorithm based on RS is practical and the performance efficiency and prediction accuracy are all higher than the sole AC Algorithm.

2. Prediction Model of AC Algorithm Based on RS

The AC Algorithm based on AC includes the following five steps:

①Make the attribute reduction processingof the multi-dimensional data sets D_1 to get the new data sets D_2;

② Production of Selectable Patterns;

③ Conversion of Selectable Patterns;

④ Selection of the Similar Patterns;

It may need the step 5 if the algorithm is applied to the prediction:

⑤ Combine the similar patterns extension to get the prediction result.

2.1. Attribute Reduction of Rough Set

The knowledge representation system can be denoted as $S=(U,A,V,f)$ Here, U represents the domain, A is the attribute sets, $V=\bigcup_{a\in A}V_a$ is the sets of attribute value, and the $f:U\times A\to V$ is an information function. Such a "Attribute-Value" forms a two dimensional table, and we call it an information table. And, if $A=C\bigcup D$ exists and C and D here are the conditional attribute and result attribute respectively, a special information-decision table is formed. The decision table can be regarded as a family of equivalence relation, and that is called the knowledge base. Not all of the conditional attributes in the decision table are necessary and some of them are useless, therefore, removal of these useless attributes will not affect the original expression effect. Then:

Definition 1: Let R be a family of equivalence relation, $r\in R$, if $ind(R)=ind(R-\{r\})$ exists, then r shall be omitted in R, otherwise, r shall not be omitted in R. For any $r\in R$ which shall not omitted in R, then R is independent.

Definition 2: If the $Q=P-r$ exists, Q is independent and satisfies $ind(Q)=ind(P)$, then Q is called an attribute of P with the expression of $red(P)$.

Definition 3: A family of Equivalence Relation P may have several reductions, we define all the attributes interests as the core of P, which is marked with $core(P)$ and the $core(P)=\bigcap red(P)$ [5] exists.

The decision table reduction is to reduce the conditional attributes in the table, viz, getting ride of the useless conditional attributes and not affecting the original express effects[5] after deleting these attributes.

2.2. Production of Selectable Patterns

For a designated real value with N observations m-dimensional sequence $x_t = \{x_{1t}, \cdots, x_{mt}\}$, $(t = 1, 2, \cdots N)$ a pattern is defined that the table containing the k lines since the number i line, and the k here is called the pattern length $(i = 1, 2, \cdots, N-k+1)$:

$$\mathbf{P}_k(i) = \begin{bmatrix} x_{1i} & \cdots & x_{li} & \cdots & x_{mi} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1,i+j} & \cdots & x_{l,i+j} & \cdots & x_{m,i+j} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1,i+k-1} & \cdots & x_{l,i+k-1} & \cdots & x_{m,i+k-1} \end{bmatrix}_{k \times m} \tag{1}$$

Comparing all possible selectable patterns $P_k(i) = (i = 1, ..., l, ..., N-k+1)$ with the reference pattern P^R , it hopes to research the system behavior by depending on the patterns which are similar to the reference pattern. Since the AC algorithm combines the similar patterns extension to form the development status of the reference pattern, this method can realize the prediction field exactly equals to the reference pattern's extension. Therefore, we select the nearest known pattern to the prediction starting point as the reference pattern, that is $P^R = P_k(N-k+1)$.

2.3. Conversion of Selectable Patterns

According to work principle, for the reference pattern with the pattern length of k , there may have one or some similar patterns with the length of k in the sample data. But considering that the system is dynamic, the similar patterns may have different average values and standard variances in different phases.

To measure the similarity among patterns, it must to seek the conversiton between the selectable patterns and the reference pattern to describe these differences,viz,alternating the patterns to the same base point to let them be comparable. The conversion always be performed by using the linear conversion, and the converted patters are:

$$\mathbf{T}_i[P_k(i)] = \begin{bmatrix} x^*_{1i} & \cdots & x^*_{li} & \cdots & x^*_{mi} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x^*_{1,i+j} & \cdots & x^*_{l,i+j} & \cdots & x^*_{m,i+j} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x^*_{1,i+k-1} & \cdots & x^*_{l,i+k-1} & \cdots & x^*_{m,i+k-1} \end{bmatrix} \tag{2}$$

$x^*_{l,i+j} = a^i_{0l} + a^i_{1l} x_{l,i+j}, j = 0, 1, \cdots, k-1; i = 1, 2, \cdots, N-k+1; l = 1, 2, \cdots, m$ and parameter a^i_{0l} can be explained as the status difference between the reference patter and the similar pattern $P_k(i)$, but the parameter a^i_{1l} just is defined as some uncertain factors.

Select the relative date of the reference pattern $x_{ij}(i = N-k+1, N-k+2, \cdots N; j = 1, 2, \cdots m)$ as the base value, for each selectable pattern $P_k(i)$, to use the least squares to calculate the unknown weight a^i_{0l} and a^i_{1l}, and then to get the error sum of square for the patter similarity calculation.

2.4. Selection of Similar PatternsPattern Similarity

The main purpose of calculating the pattern similarity is to identify the similarity among the different pattern shapes. In order to measure the similarity between a selectable pattern $P_k(i)$ which is converted according to the step 3 and the reference pattern P^R, it needs to measure the distances between the k observations with m system variables of the two patterns.

Commonly, the distance from the number i selectable pattern to the reference pattern can be defined as:

$$d_i = \frac{1}{k+1}\sum_{j=0}^{k-1}\sqrt{\sum_{r=1}^{m}\left(x_{r,i+j} - x_{r,N-k+j+1}\right)^2} \tag{3}$$

The pattern similarity shall be measured by the distance, and the number i pattern's similarity to the reference pattern- s_i is defined as: $s_i = 1/d_i$, It is clearly seen that the bigger the distance value, the smaller the pattern similarity.

2.5. E Selection of the Similar Pattern for Prediction

When calculating the pattern similarity, we can accord to the pattern similarity to select the similar patterns. It all depends on the application categories to decide which patterns shall be picked. There are four problems to be solved when it is applied to the prediction:

① Variable Sets;

② Pattern Length;

③ Select the number of similar pattern;

④ Value the Weight Coefficient.

The first problem is the inevitable one when digging all the data, but we can use the known algorithm of rough sets attribute reduction to get the most effective variables. The selection of pattern length is discussed in document [6], and the author put forwards three methods to get the suitable pattern length. In practical application, we can set a extent of the pattern length based on the concrete problems, that is to say, get the $k_{\max}$ and $k_{\min}$ with considering all patterns with the pattern length $k \in [k_{\min}, k_{\max}]$.

Here, Set that there are m variables, the pattern length is k and the reference patter is x^R, the selected similar patterns are $x_1, x_2, \cdots, x_F$, so each pattern has $k \times m$ data. We can depend on the GMDH Algorithm to select the number of similar patterns and determine the weight coefficient when processing the patterns combination at the same time. Using the linear input and output GMDH model, we can get: $x^R = \sum_{j \in J} g_j x_j$. J represents the subsets of the sets $\{1, 2, ..., F\}$, $\{g_j\}$ represents the weight coefficient of the relative patterns, and patterns with those marks shall be combined. When the $k \times m$ is very small, we commonly take as the weight coefficient and the equation $\sum_{j \in J} g_j = 1$ exists.

3. Research on Shanghai Composite Index

3.1. Data Selection and Pre-processing

Since the Shanghai Stock market started earlier and had a great variety of data, the composite index are affected by various influences, so the index are more representative. Therefore, this article picks the Shanghai Composite Index Daily Close Index as the research objective with the research goal of predicting the Shanghai Composite Index's Incoming Day's Variation Trend. This article defines the following 5 variation trends:

① Slump(T=0) $R \in (-\infty, -2.0\%]$;

② Down(T=1) $R \in (-2.0\%, -0.5\%]$;

③ Adjust(T=2) $R \in (-0.5\%, 0.5\%]$;

④ Up(T=3) $R \in [0.5\%, 2.0\%)$;

⑤ Soar(T=4) $R \in [2.0\%, +\infty)$.

R represents the rising range.

When analyzing the stock market system, we may have aids of six attribute index of 15 characteristic values and original data which are commonly used to reflect the stock market[7]. Based on the narration as above, using NaiveSacler Discretization and Johnson's Attribute Reduction to pre-process the Shanghai Stock Market Data in 2005, we can get one of the attribute reduction results of the Shanghai Composite index Daily Variation Trend--$\{F_1, F_5, F_6\}$ (Here, F_1 represents the open price, F_5 represents the Trading Volume and F_6 represents the Quantity Relative Ratio).The Shanghai Composite Index from Feb.17, 2005 to Aug.19, 2005 from the above-mentioned data are selected as the learning sample data, so there are $N = 125$ learning samples and $m = 4$ variations in the AC algorithm with the data is omitted here.

3.2. Establishment of AC Model

Setting the Pattern Length to be $k = 5$,the AC algorithm automatically selects the data ranging from Aug. 15, 2005 to Aug. 19, 2005 as the reference pattern. The AC automatically selects 5 similar patterns to combine and gets the smallest pattern similarity of 0.8750. See Fig 1 for the results produced from the computer performing.

Table 1. Similar Pattern and its Similarity

Reference Pattern/Day	Similar Pattern/Day	Similarity
15-Aug~ 19-Aug	25-Apr~29-Apr	0.9526
	20-Jul~26-Jul	0.9317
	22-Apr~28-Apr	0.9317
	15-Jun~21-Jun	0.9147
	21-Apr~27-Apr	0.8750

Table 2. Predicted Results of the Variation Trend

Date	22-Aug	23-Aug	24-Aug	25-Aug	26-Aug
Actual Value	3.0000000	1.0000000	3.0000000	2.0000000	2.0000000
Predicted Value	2.7777777	1.2046686	3.0000000	2.1191918	1.9461279
Relative Error (%)	-7.4	20.5	0.0	6.0	-2.6

See Fig1 for the Shanghai Composite Index Daily Variation Trend.In Fig1, the blue line represents the Actual Shanghai Composite Index Daily Variation Trend; the red line represents the Predicted Shanghai

Composite Index Daily Variation Trend. Seeing in the Fig, it is obviously seen that the AC algorithm based on RS can get a more accurate data for the index prediction.

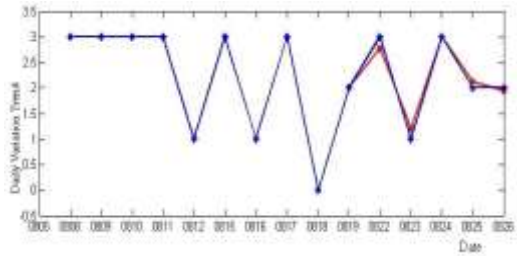

Fig.1. Composite Index Daily Variation Trend

4. Conclusion

According to the features of the stock market, it is a practical endeavor to apply the AC algorithm to predict the stock index daily variation trend. Relying on the experiments, we can say that the AC algorithm based on RS is a realistic method in predicting the stock index daily variation trend. This method not only improves the searching speed of the similar pattern, but also has a higher accuracy. Meanwhile, by using the GMDH technology to automatically combine the similar pattern, it enables this method to be a noteworthy new method in the fields of pattern recognition and artificial intelligence research.

References

[1] Guangqiang Wang, Peiling Zhou,"Application of Neuron Network in Stock Index Predicton[J],"Computer Engineering,vol.32,no.1,pp. 211-212, January 2006.
[2] Dengfeng Wu, "Application of GMDH network with chaos characteristic in rainfall predicton[J]," Journal of Chinese Computer Systems,vol. 21,no. 2,pp. 135-137. February 2000.
[3] Lorence,E.N. "Athmospheric predictability is revealed by naturally occurring analogues[J]," J.Atmos.Sci,no. 4,pp. 636-646. April 1969.
[4] Hema R.Madala,Alexy G.Ivakhnenko. "Inductive Learning Algorithms for Complex Systems Modeling [M],"Boca Raton,Ann Arbor: CRC Press Inc, 1994.
[5] Wenxiu Zhang, "Rough Set Theory and Method [M],"Scientific Publishing Company,2005.
[6] Motnikar，B.S.u.a, "Time-series forecasting by pattern imitation[J],"OR Spektrum, , vol. 18,no. 1,pp. 43-49. January 1996.
[7] Lin Zhu, "Rough Set Integrated Neural Network Model For Forecasting Stock Price[J],"Chinese Journal of Management Science,vol. 10,no. 4,pp. 7-12. April 2002.