Outlier Detection Based on Wavelet-HMM methods

Fang Liu[1, 2, 3], Ruzhen Dou[4], Chaoying Xia[1]
[1]*School of Electrical and Automation Engineering, Tianjin University, 300072 Tianjin, China*
[2]*Tianjin Qingyuan electric vehicle limited liability company Tianjin, China*
[3]*School of Computer Science & Software Engineering, TIANJIN Polytechnic University, 300387 Tianjin, China*
[4]*China Automotive Technology & Research Center, Tianjin, China*

Keywords: Outliers detection; Improved recursive wavelet transform; HMM; Process data

Abstract: According to the limitation of the principle of outlier detection based on wavelet, this paper proposes a new outlier detection method called Wavelet-Hidden Markov Model (W-HMM) algorithm. In this algorithm, the signal is decomposed in some scale, and when the wavelet decompositions of the signal are different from the most other wavelet decompositions, the signal will be seen as potential outlier. Aiming to make further accurate judgement, the similarity measure between the wavelet coefficient of this signal and that of normal signal will be done, and the final confirming is made by Viterbi algorithm which is used to HMM. The validity and practicality are proved by experimentation and application in this paper.

1. INTRODUCTION

The process data in the control industry can reflect the real-time situation of the overall control system. Therefore, the model parameters identification, process monitoring, fault diagnosis and end point prediction methods based on process data are widely used by people. However, due to the complicated on-site environment and sensor failures, it is difficult to avoid some erroneous data in the process data. The presence of these erroneous data directly affects the accuracy of the above data-based analysis methods. Therefore, it is very important to control the anomaly detection of process data.

In mathematics, the definition of anomaly is: The mutation of function amplitude or derivative is described by Lipschitz exponent, that is, the smaller the Lipschitz exponent is, the greater the possibility of anomaly [1]. Mallat [2] established the relationship between Lipschitz exponent and wavelet coefficients in 1992, and proposed the principle of wavelet transform modulus maximum analysis of abnormal data and was widely used in various fields by later generations [3, 4]. However, for the control process data, because it fluctuates greatly during the initial adjustment process, if the abnormality of the signal is judged only through the abrupt change of the amplitude or the derivative, it is less suitable; Secondly, the principle of modular maxima of wavelet transform considers white noise as an abnormal signal [5], that is, both the white noise and the Lipschitz exponent of abnormal data are less than zero, and show the same characteristics (as the scale increases, the modulus maxima will decrease). Therefore, using Mallat's wavelet transform modulus maximum principle is not easy to distinguish between noise signals and abnormal signals.

Considering the shortcomings of the existing wavelet methods, this paper proposes a W-HMM (Wavelet-Hidden Markov Model)-based abnormal data detection method for control process. This method is different from Mallat's idea of the modular maxima principle. Instead, it uses the basic features of wavelet transform to detect abnormal data, thus avoiding the difficulty of distinguishing between noise and abnormal data due to the modular maxima principle method. In order to avoid setting the detection threshold in advance, this method uses the Viterbi algorithm for solving HMM optimal state hidden chain to obtain the detection result of abnormal data [6]. Taking into account the large amount of data in the process data, the need for real-time detection and other characteristics, this article adopts an improved recursive wavelet transform (IRWT) that can run online to ensure the real-time detection process. Through verification and application, we can see that the abnormal data detection method based on Wavelet-HMM can accurately detect the outliers in the process data, and prove the effectiveness and practicality of the method.

2. WAVELET HMM ABNORMAL DATA DETECTION METHOD

Based on the Wavelet-HMM detection algorithm, the wavelet transform can simultaneously describe the characteristics of the signal in time and frequency, and the detected data is subjected to wavelet decomposition at a certain frequency, and the data anomaly is judged according to the change of the decomposed wavelet coefficient. That is, the algorithm considers that the normal signal should be composed of several signals with a fixed frequency [7]. The wavelet coefficient should fluctuate within a certain range. When the signal is abnormal, the wavelet coefficient will also change. Therefore, we can judge the abnormal situation of data by monitoring the change of wavelet coefficients. According to this idea, the detection algorithm based on Wavelet-HMM first performs wavelet decomposition on a certain scale in the data to be detected. Second, Gaussian function is used to detect the similarity between the wavelet coefficients and the normal data wavelet coefficients. Finally, the Viterbi algorithm is used to obtain the Markov state hidden chain which

representative data anomalies. The entire detection process does not need to set the detection threshold, can be detected online, and the amount of calculation is small.

2.1 Improve recursive wavelet algorithm

In 1996 Chaari proposed a damping wavelet in the form of [8]:

$$\Psi_{Chaari}(t)=(1+\sigma|t|+\frac{\sigma^2}{2}t^2)e^{-\sigma|t|}e^{i\omega_0 t} \quad (1)$$

In which $\sigma=2\pi/\sqrt{3}$; $\omega_0=2\pi$, at that time, $\Psi_{Chaari}(0)=0$, to ensure that the basic wavelet satisfies the admissibility condition.

However, since the wavelet-based recursive wavelet transform (Recursive Wavelet Transform) is composed of a combination of a positive transform and an inverse transform, the inverse transform needs to use the wavelet coefficients at a future time, and the amount of calculation is large. Therefore, the literature [9] proposed an improved recursive wavelet transform algorithm.

Define the function:

$$\Psi_1(t)=(\frac{\sigma^3t^3}{3}-\frac{\sigma^4t^4}{6}+\frac{\sigma^5t^5}{15})e^{(-\sigma+i\omega_0)t}u(t) \quad (2)$$

Let $\Psi(t)=\Psi_1^*(-t)$ be the fundamental wave, then the fundamental wave is:

$$\Psi(t)=(-\frac{\sigma^3t^3}{3}-\frac{\sigma^4t^4}{6}-\frac{\sigma^5t^5}{15})e^{(\sigma+i\omega_0)t}u(-t) \quad (3)$$

Where * denotes conjugate; choose $\sigma=2\pi/\sqrt{3}$; $\omega_0=2\pi$, where $\Psi(0)=0$, to ensure that the basic wavelet satisfies the admissibility condition.

Decomposition and discretization of the signal $x(t)$ according to the wavelet of equation (3) is:

$$\begin{aligned}W_{x,\Psi}(f,kT)&=\sqrt{f}T\sum_{n=1}^{\infty}x(nT)\Psi^*(f(nT-kT))\\&=\sqrt{f}T\sum_{n=1}^{\infty}x(nT)\Psi_1(f(kT-nT))\end{aligned} \quad (4)$$

The formula (4) is expressed as a convolution:

$$W_{s,\Psi}(f,kT)=T\sqrt{f}\,(x(nT)*\Psi_1(fnT)) \quad (5)$$

Among them, T is the sampling period; k,n is the integer mark sampling point; f is the reciprocal of the scale;

The Z-transformation of (5) can turn the convolutional form into a product form:

$$W_{x,\Psi}(Z)=T\sqrt{f}\,(x(Z)\cdot\Psi_1(Z)) \quad (6)$$

Where $\Psi_1(Z)$ can be expressed as:

$$\Psi_1(Z)=\frac{\delta_1Z^{-1}+\delta_2Z^{-2}+\delta_3Z^{-3}+\delta_4Z^{-4}+\delta_5Z^{-5}}{\lambda_1Z^{-1}+\lambda_2Z^{-2}+\lambda_3Z^{-3}+\lambda_4Z^{-4}+\lambda_5Z^{-5}+\lambda_6Z^{-6}} \quad (7)$$

In which

$$a=e^{-fT(\sigma-i\omega_0)}$$

$$\delta_1=[\frac{(\sigma fT)^3}{3}-\frac{(\sigma fT)^4}{6}+\frac{(\sigma fT)^5}{15}]\cdot a$$

$$\delta_2=[\frac{(\sigma fT)^3}{3}\cdot 2-\frac{(\sigma fT)^4}{3}\cdot 5+\frac{(\sigma fT)^5}{15}\cdot 26]\cdot a^2$$

$$\delta_3=[\frac{(\sigma fT)^3}{3}\cdot(-6)+\frac{(\sigma fT)^5}{5}\cdot 22]\cdot a^3$$

$$\delta_4=[\frac{(\sigma fT)^3}{3}\cdot 2+\frac{(\sigma fT)^4}{3}\cdot 5+\frac{(\sigma fT)^5}{15}\cdot 26]\cdot a^4$$

$$\delta_5=[\frac{(\sigma fT)^3}{3}+\frac{(\sigma fT)^4}{6}+\frac{(\sigma fT)^5}{15}]\cdot a^5$$

$$\lambda_1=-6a\,,\ \lambda_2=15a^2\,,\ \lambda_3=-20a^3,$$

$$\lambda_4=15a^4\,,\ \lambda_5=-6a^5\,,\ \lambda_6=a^6,$$

Substituting (7) into (6) yields:

$$\begin{aligned}&W_{x,\Psi}(Z)(1+\lambda_1Z^{-1}+\lambda_2Z^{-2}+\lambda_3Z^{-3}+\lambda_4Z^{-4}+\lambda_5Z^{-5}+\lambda_6Z^{-6})\\&=\sqrt{f}Tx(Z)\cdot(\delta_1Z^{-1}+\delta_2Z^{-2}+\delta_3Z^{-3}+\delta_4Z^{-4}+\delta_5Z^{-5})\end{aligned} \quad (8)$$

Converting equation (8) into discrete form is the formula for improving recursive wavelet decomposition:

$$\begin{aligned}&W_{x,\Psi}(kT,f)=\sqrt{f}T\{\delta_1x[(k-1)T,f]+\\&\delta_2x[(k-2)T,f]+\delta_3x[(k-3)T,f]+\\&\delta_4x[(k-4)T,f]+\delta_5x[(k-5)T,f]\}\\&-\lambda_1W_{x,\Psi}[(k-1)T,f]-\lambda_2W_{x,\Psi}[(k-2)T,f]\\&-\lambda_3W_{x,\Psi}[(k-3)T,f]-\lambda_4W_{x,\Psi}[(k-4)T,f]\\&-\lambda_5W_{x,\Psi}[(k-5)T,f]-\lambda_6W_{x,\Psi}[(k-6)T,f]\end{aligned} \quad (9)$$

From equation (9), we can know that the initial six wavelet coefficients can be calculated according to equation (4), and the wavelet coefficients at the current time can be calculated by using the signals of the first five moments and the wavelet coefficients of the first six moments. Based on this, online wavelet decomposition is implemented to meet the requirements of online abnormal data detection. Since the wavelet in (3) is a tightly supported wavelet, initializing the wavelet coefficients requires only supporting data within the range, without all data.

2.2 HMM and its Viterbi algorithm

2.2.1 HMM Structure

HMM is a first-order double stochastic process consisting mainly of two parts [10,11]: One of them is the

Markov chain, which describes the transition of states, described by the initial state probability π and state transition matrix $A=(a_{ij})_{N\times N}$, where $a_{ij}=P(S_t=j|S_{t-1}=i)$, $i,j\in S_s$, in where S_s represents a set of all states, S_t represents a state at time *t*, and *N* represents the total number of states of the model;

Wavelet-HMM detection algorithm uses Markov chain to represent data anomalies, that is, $N=2$. If 1 is used to indicate that the data is normal and 0 is abnormal, $S_s=\{0,1\}$; accordingly, the state transition matrix A indicates the transition probability between abnormal and normal data.

Another random process of the HMM describes the statistical correspondence between states and observations, described by the observed probability matrix $B=(b_{tk})_{N\times N}$, indicating the probability that the observation takes a certain value in a certain state S_t. In this paper, $b_{tk}=P(W_{ave},W_{x,\Psi}(t,f)|S_t=k)$ indicates the similarity probability of the wavelet coefficient $W_{x,\Psi}(t,f)$ and the normal data wavelet coefficient W_{ave} in the state k;

$$\begin{aligned}&P(W_{ave},W_{x,\Psi}(t,f)|S_t=s_k)=N(W_{x,\Psi}(t,f)|W_{ave},W_{\text{var}})\\&=\exp(-\frac{1}{2}(W_{x,\Psi}(t,f)-W_{ave})^T W_{\text{var}}^{-1}(W_{x,\Psi}(t,f)-W_{ave}))\end{aligned}\quad(10)$$

Among them, $N(\cdot|\cdot)$ represents Gaussian probability, W_{var} represents normal wavelet coefficient variance;

2.2.2 Viterbi algorithm

The Viterbi algorithm is widely used by scholars to solve the optimal sequence problem in three problems of HMM [12,13]. In order to make the Viterbi algorithm determine the value of the state chain online, [14] gives the real-time calculation method of Viterbi algorithm:

$$\begin{aligned}&\varphi_t(1)=a_{i1}\cdot P(W_{ave},W_{x,\Psi}(t,f)|S_t=1);\\&\varphi_t(0)=a_{i0}\cdot(1-P(W_{ave},W_{x,\Psi}(t,f)|S_t=1))\end{aligned}\quad 11)$$

Where $\varphi_t(1)$, $\varphi_t(0)$ are the indicators of $S_t=1$ (data normal) and $S_t=0$ (abnormal data), respectively. Taking a_{i1} as an example, it represents the state transition probability that the state at the previous moment is i and the current state is 1.

Therefore, simply comparing the size of $\varphi_t(1)$ and $\varphi_t(0)$ at each moment can determine the value of the HMM state chain at *t* (ie, the data anomaly):

$$\begin{aligned}&\varphi_t(1)\geq\varphi_t(0),\ S_t=1\\&\varphi_t(1)<\varphi_t(0),\ S_t=0\end{aligned}\quad(12)$$

2.3 Dynamic Update Parameters

Wavelet-HMM detection algorithm needs to be updated online to adapt to time-varying system parameters: normal wavelet coefficient mean W_{ave}, variance W_{var}, and HMM state transition matrix $A=(a_{ij})_{2\times 2}$.

The mean value and variance of normal wavelet coefficients are updated with forgetting factors:

$$W_{ave}(t)=rW_{ave}(t-1)+(1-r)W_{x,\Psi}(t,f)\quad(13)$$

$$\begin{aligned}&W_{\text{var}}(t)=rW_{\text{var}}(t-1)+(1-r)(W_{x,\Psi}(t,f)\\&-W_{ave}(t-1))^T(W_{x,\Psi}(t,f)-W_{ave}(t-1))\end{aligned}\quad(14)$$

Among them, r is forgetting factor, if the data is abnormal data, in order to avoid its impact on W_{ave}, W_{var}, it will not be updated.

The state transition matrix $A=(a_{ij})_{2\times 2}$ is calculated as:

$$\begin{aligned}&a_{01}=\frac{N(a_{01})}{N(a_{01}+a_{00})}\quad,\quad a_{11}=\frac{N(a_{11})}{N(a_{11}+a_{10})}\\&a_{00}=1-a_{01}\quad,\quad a_{10}=1-a_{11}\end{aligned}\quad(15)$$

Where $N(a_{ij})$ denotes the number of occurrences of the situation where the data state is i at the previous moment and *j* is the data state at the latter moment [15].

Since the proportion of abnormal data in the data to be detected is often small, when the initial value of the state transition matrix is set, this should be taken into consideration, ie, a_{11} and a_{01} should be correspondingly large. With the on-line detection, the state transition matrix is continuously updated according to (15). As the time increases, the subsequent state transition matrix will be more in line with the actual situation of the data to be detected. Therefore, the influence of the initial value of the state transition matrix on the overall detection result should be less than the detection threshold.

3. VERIFICATION AND APPLICATION

Verification

In order to verify the validity of Wavelet-HMM detection algorithm, this paper uses three models to generate three sets of data to verify the algorithm:

(1) Utilize the mechanism model of the electrode adjustment system of the three-phase AC arc furnace to generate the output signal of a set of electrode regulation systems, and add 10% white noise and 14 abnormal points to simulate the collapse and arc breaking of the actual electrode regulation system. Anomalies such as the signal to be detected and the test results are shown in Figure 1;

(2) Add 10% white noise and 18 abnormal points to the sine signal of amplitude 5 to form the second group of signals to be detected; the signal and the detection result are shown in Figure 2;

(3) Using the model in the Alex Alexandridis paper [16]: Generate the third set of data to be tested and add 10% white noise and eight outliers to it. The data and test results are shown in Figure 3.

In the three diagrams, (a) is the signal to be detected, the abscissa is the sampling point, and the ordinate is the signal amplitude; (b) is the wavelet decomposition in a certain scale, the abscissa is the sampling point, and the ordinate is the wavelet coefficient amplitude; From the three figures (b), it can be seen that the abnormity of the wavelet coefficients at the location of the anomaly is significantly different from that of the normal signal. In the three figures (c) are the detection results, the abscissa is still the sampling point, the ordinate is the detection result, "1" indicates that the data is normal, and "0" is abnormal. Among them, all abnormal points are detected in Figure 1-(c), and there are no missed detections and false detections. The accuracy rate is 100%. Figure 2-(c) has a misdetection at the initial stage. The accuracy rate was 99.93%; Figure 3-(c) also had a false detection in the initial stage with an accuracy of 99.9%. Therefore, we can see from the above three groups of detection results that the detection algorithm based on Wavelet-HMM can detect the outliers in the signal more accurately and have good anti-noise ability.

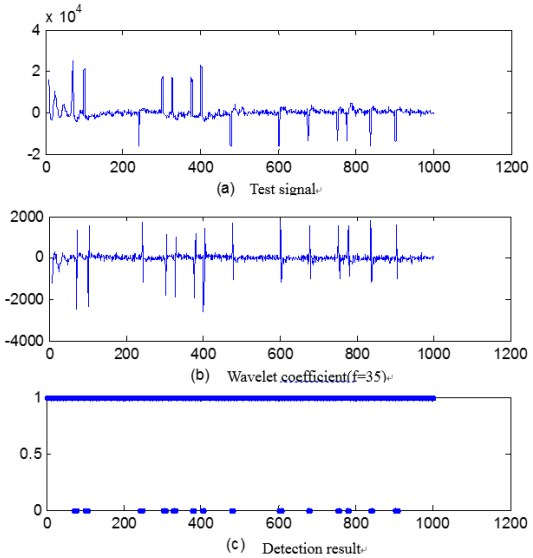

Fig. 1 Electrode regulation system signal and test result

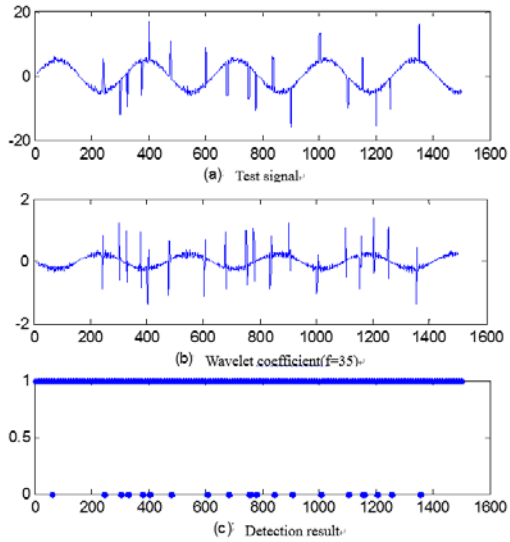

Figure 2 Sine signal and test results

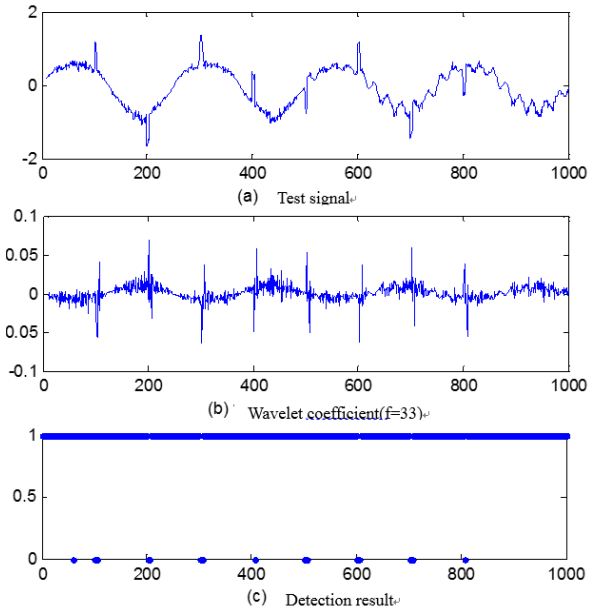

Figure 3 Alex model signal and test results

There are two reasons for analyzing the erroneous detection of the initial stage in Fig. 2 and Fig. 3: The first is the calculation of the probability b_{tk} of the observation of the HMM. From the formula (10), the accuracy of b_{tk} is determined by the mean value of the normal wavelet coefficient W_{ave} and variance W_{var}. Since there are less data in the initial stage, the estimates for W_{ave} and W_{var} are less accurate, resulting in the inaccuracy of b_{tk}; the second problem is the initial value of HMM state transition matrix A; Due to the above two reasons, there was a detection error at the beginning of the test. If more data can be used to make more accurate estimation of the above parameters before detection, the error can be effectively avoided.

4. CONCLUSIONS

This paper proposes anomaly data detection algorithm based on wavelet Hidden Markov Model, aiming at the limitations of the existing wavelet for defining abnormal data (the point where the function or its derivative discontinuity is abnormal data) and the deficiency of the detection process. The algorithm is based on the most basic principle of wavelet decomposition, using the improved recursive wavelet decomposition algorithm that can be decomposed online to decompose the test data at a certain scale. If the decomposed wavelet coefficient is obviously different from the other wavelet coefficients, the abnormal point is considered to exist. The combination of wavelet and HMM effectively avoids the need to set the detection threshold in advance. Through verification, it can prove the validity of the wavelet-HMM-based anomaly data detection algorithm, and its anti-noise and practicality.

ACKNOWLEDGEMENTS

This research is partially supported by National Natural Science Foundation of China under Grant 51607122, 51378350.

This research is partially supported by State Key Labora-tory of Process Automation in Mining & Meallurgy/ Bei-jing Key Laboratory of Process Automation in Mining & Metallurgy Research Fund Project BGRIMM-KZSKL-2017-01.

This research is partially supported by Tianjin Municipal Education Commission research project 2017KJ094.

This research is partially supported by Tianjin Science and Technology Project 17ZLZXZF00280.

REFERENCES

[1] Pittner S, Kamarthi S V. Feature extraction from wavelet coefficients for pattern recognition tasks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999, 21(1): 83-88.

[2] Stephane M, Wen Liang Hwang. Singularity Detection and Processing with Wavelets[J]. IEEE Transactions on information theory, 1992, 38(2): 617-642.

[3] Radu R, Valerie L D. Christian Heinrich, Didier Wolf, Iterative Wavelet-based Denoising Methods and Robust Outlier Detection[J]. IEEE Signal processing Letters, 2005, 12(8): 557-560.

[4] Zbigniew R, Struzik, Arno P J M. Wavelet transform based multifractal formalism in outlier detection and localization for financial time series[J]. Physica A, 2002, 309(3-4): 388-402.

[5] Tao L, Qi L, Shenghuo Zhu, Mitsunori Ogihara. A Survey on Wavelet Applications in Data Mining[J]. Sigkdd Explorations, 2002, 4(2): 49-68.

[6] Blake H, Paul Lee. Hidden Markov Model analysis of force/torque information in telemanipulation [J]. International Journal of Robotics Research, 1991, 10(5): 528-539.

[7] Robi Polikar. The engineer's ultimate guide to wavelet analysis-The Wavelet Tutorial[EB/OL]. (1999-03-07) [2001-01-12].http://users.rowan.edu/~polikar/WAVELETES/WTtutorial.html.

[8] Oinis C, Michel M,Francoise B. Wavelets: a new tool for the resonant grounded power distribution systems relaying[J]. IEEE Transations on Pover Delivery, 1996, 11(3): 1301-1308.

[9] Zhang C L, Huang Y Z,Ma X X,Lu W Z,Wang G X. A new approach to detect transformer inrush current by applying wavelet transform[C]. 1998 International Conference on Power System Technology. China: POWERCON '98, 1998: 1040-1044.

[10] Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257-286.

[11] Jeff A B. What HMMs Can Do[J]. IEICE-- Transactions on Information and Systems, 2006, E89-D (3): 869-891.

[12] Biing-Hwang Juang, Lawrence R. Mixture Autoregressive Hidden Markov Models for Speech Signals[J]. IEEE Transactions on acoustics and signal processing, 1985, 33(6):41-44.

[13] Kavcic A, Moura J M F. The Viterbi Algorithm and Markov Noise Memory[J]. IEEE Transactions on information theory, 2000, 46(1): 291-301.

[14] Hui L L, Moura J M F. Implementing the Viterbi algorithm, fundamentals and real-time issues for processor designers[J]. IEEE Signal Processing Magazine, 1995, 12(5): 42-52.

[15] Vikram K, George G Y. Recursive Algorithms for Estimation of Hidden Markov Models and Autoregressive Models with Markov Regime[J]. IEEE Transactions on Information Theory, 2002, 48(2): 458-476.

[16] Alex A, Haralambos S, George B. A new algorithm for online structure and parameter adaptation of RBF networks[J].Neural Networks, 2003, 16(7): 1033-1017