# Research on Data Flow Partitioning Based on Dynamic Feature Extraction

**Min Zhang , Wei Wang**

Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission,

Tianjin Normal University, Tianjin, China,300387
E-mail: weiwang@tjnu.edu.cn

**Abstract.** With the rapid development of the Internet of Things, social networks and e-commerce, the era of big data has arrived. Although big data has great potential for many areas such as industry, education and healthcare, getting valuable knowledge from big data can be a daunting task. Big data has the characteristics of high-speed change, and its content and distribution characteristics are in dynamic changes. Most current models are static learning models that do not support on-line updating, making it difficult to learn dynamically changing big data features in real time. In order to solve this problem, this paper proposed a method to support incremental recursive least squares (IRLS) regression parameter estimation and variable sliding window algorithm to analyze and judge the trends of dynamic characteristics of data streams, which can provide early warning, status assessment and decision support for monitoring objects and improve the accuracy and adaptability of data flow classification. The computational real-time and analysis accuracy are obviously improved than the traditional algorithm, the simulation results verify the effectiveness of the proposed algorithm.

**Keywords:** Trend analysis, dynamic data mining, incremental recursive least square method, variable sliding window.

## 1. Introduction

Dynamic data flow refers to an ordered sequence of data consisting of a large number of continuously arriving, potentially infinitely long, fast-changing data with transient, real-time, and infinite characteristics. In recent years, data flow has been widely used in many fields such as stock exchange, telephone communication record, network traffic monitoring, sensor network and so on. Data stream contains a lot of information, which can be used as the basis for intelligent decision-making. The purpose of feature extraction and description of dynamic data stream is to extract the change information and analyze the data stream signal to achieve decision support, such as stock market curve analysis and audio waveform analysis, ECG waveform recognition analysis and industrial production curve identification. At present, however, most of the data stream are generated by variety of dynamic systems in real time, which is a high-speed incoming data, so the data stream data transmission, calculation and storage become very difficult. Due to the different of its application fields, the characteristics of the dynamic data flow curve vary widely, and most of them exist in a process system of nonlinear, multivariable coupling, events and large time delays. Therefore, the data flow trend analysis method requires strong adaptability and high analysis accuracy. The established data model can be applied to dynamic data flow analysis and prediction in various fields [1].

The traditional data flow curve trend analysis algorithm has sliding window algorithm(SW) [2-4], extrapolated online data segmentation algorithm(OSD) [5], Bottom-Up algorithm [4], Top-Down algorithm [4], but they have obvious deficiencies. In the following, we give some shortcomings of SW and OSD algorithm in data segmentation:

A. There is no limitation to the maximum length of sliding window in SW algorithm. When the threshold of detection point is relatively large, the length of window may be too long and the fitting error becomes large. And the arriving data needs a complete regression modeling and data segment segmentation point judgment. As the data increases, the computational efficiency is extremely;
B. The OSD algorithm sets the minimum sliding window on the basis of the SW algorithm, and only substitutes the newly arrived data into the established model to analyze the extrapolated cumulative error. Although the efficiency is improved, the transition point in the sliding window cannot be detected;
C. The SW and OSD algorithm have no limitation on the maximum length of the sliding window. When the detection threshold is relatively large, the length of the window may be so long that the trend analysis error becomes larger;
D. Both SW and OSD use the conventional least-squares method for curve fitting. Compared with the conventional least squares method, incremental recursive least squares method is more efficient and has higher fitting accuracy.

In order to improve the precision of trend analysis and make the data stream feature more accurate, this paper improves the traditional data flow trend analysis algorithm, and proposes incremental recursive least-squares (IRLS) [6] linear regression modeling to get the model fitting parameters. Using variable sliding window algorithm [1] according to the trend of eigenvalues to split the data stream. The algorithm limits the lower limit of the length of a data stream. And the algorithm only recursively calculates the parameters of the regression model before the new data stream segmentation starts until the minimum segmentation length is reached, and does not perform segmentation point detection. The experimental results show that the proposed algorithm can get better fitting accuracy, update fitting parameters in real time, segment data segments online, and has high computational efficiency and small fitting error. It is suitable for the dynamic model of data. Moreover, based on the real-time trend analysis of data flow, the segment data segment can get better classification accuracy and recognition efficiency for the later target recognition processing. It is widely used in agriculture, commerce, healthcare and other fields. In this paper, a real-time algorithm that can adjust the trend parameters online according to the elements of the new arrival data stream is proposed. In addition, the variable sliding window algorithm is used to quickly detect the data stream segmentation points and determine the segmentation. The algorithm has high precision and low computational time complexity. The remainder of the paper is organized as follows. In the section 2 the data flow problem will be described. Then data stream linear regression modeling and variable sliding window algorithm will be introduced in section 3. Experimental results for blocking the data flow will be showed in the section 4. Conclusion and discussion is in section 5.

## 2. Data flow problem description

The real-time trend analysis of data flow is based on real-time segmentation according to some statistical property index (such as mean square error, cumulative error and generalized likelihood ratio statistics), so that the data in the divided data segments are subject to the same statistical model, Segment obeys different statistical models.

In order to facilitate the study, we define: one-dimensional continuous of data flow is:

$$Y = \{\upsilon_{t1}, \cdots, \upsilon_{ti}, \cdots, \upsilon_{tc}, \cdots\} \tag{1}$$

Where $\mathrm{t}_c$ is the current moment.

Data stream segmentation (This paper uses the mean square error comparison) to divide Y into a series of continuous non-empty data segments (i.e. sliding windows): $\{Y_1, \cdots, Y_j, \cdots, Y_s, \cdots\}$

The jth data segment is:

$$Y_j = \{\upsilon_{t_{j,1}}, \cdots, \upsilon_{t_{j,\lambda}}, \cdots, \upsilon_{t_{j,n_j}}\} \tag{2}$$

Corresponding data arrival time:

$$t_{j,\lambda} \in \{t_1, \cdots, t_i, \cdots, t_c, \cdots \mid j \in N, 1 \le j \le s; \lambda \in N, 1 \le \lambda \le n_j\} \tag{3}$$

In equations (2)~(3),the length of the data segment $Y_j$ is denoted by $n_j$ ,that is, the length of the sliding window $Y_j$ ,and $t_{1,1} = t_1$ .Note that the data segment $Y_s$ includes the data segment of the current data $\upsilon_{t_c}$ .

Suppose the data in $Y_j$ can be fitted by linear regression model, that is:

$$\upsilon(t) = f(t, \theta_j) + \varepsilon_j(t), \quad t \in \{t_{j,1}, \cdots, t_{j,n_j}\} \tag{4}$$

Where $f(t, \theta_j) = a_j t + b_j$ is the linear regression model of data segment $Y_j$ ,

$\theta_j = [a_j, b_j]^T$ is the model parameter vector, parameter $a_j$ is the trend characteristic value of data segment $Y_j$, and $\varepsilon_j(t)$ is the independent and identically distributed zero-mean white noise. For the purpose of algorithm description, let the first data element $\upsilon_{t_{j+1,1}}$ of data segment $Y_{j+1}$ be the dividing point of $Y_j$.

The basic task of the real-time trend analysis of data stream is to perform the following calculation on the newly arrived data stream element $\upsilon_{t_c}$ based on the currently accepted data sequence $Y_{s,n} = \{\upsilon_{t_{s,1}}, \cdots, \upsilon_{t_{s,n}}\}$ :1) Split point detection (such as detecting whether $\upsilon_{t_c}$ is used as the division point of data segment $Y_{s,n}$; 2) Establish a regression model for the current data segment to calculate the current fitted model parameter values $a_j$ and $b_j$.

## 3. Data flow real-time trend analysis method

In order to overcome the shortcomings of existing trend analysis algorithms, this paper presents a new real-time trend analysis algorithm for data streams to overcome the drawbacks of high computational complexity of SW algorithm and low accuracy of OSD algorithm in traditional methods.

## 3.1 Incremental recursive least squares algorithm(IRLS)

In order to identify the time-varying system in real time, Zhou et al proposed an incremental recursive least squares (IRLS) algorithm [7]. The algorithm can use the newly arrived data to correct the original model parameters to obtain new model parameters.

Let the current linear regression model constructed for the data sequence $Y_{s,n} = \{\upsilon(t_{s,1}), \cdots, \upsilon(t_{s,n})\}$ be $f(t, \theta_{s,n})$, and the data sequence be expressed as $Y_{s,n} = [\upsilon(t_{s,1}), \cdots, \upsilon(t_{s,n})]^T$ in vector form, then $Y_{s,n} = U_{s,n} + \varepsilon_{s,n}$. among them:

$$U_{s,n} = \begin{bmatrix} t_{s,1} & \cdots & t_{s,n} \\ 1 & \cdots & 1 \end{bmatrix}^T, \quad rank(U_{s,n}) = 2 < n \tag{5}$$

$\varepsilon_{s,n}$ is the expected random error vector of zero. The parameter $\theta_{s,n}$ is estimated by the least square method, even if $\theta_{s,n}$ is satisfied

$$\min \mu_{s,n} \left\|\varepsilon_{s,n}\right\|^2 = \left\|Y_{s,n} - U_{s,n}\theta_{s,n}\right\|^2 \qquad (6)$$

Let $p_{s,n} = U_{s,n}^T U_{s,n}$, $q_{s,n} = U_{s,n}^T Y_{s,n}$. Derivative of $\mu_{s,n}$ and make its derivative of 0, can be obtained $\theta_{s,n} = p_{s,n}^{-1} q_{s,n}$. When the data stream element $\upsilon(t_c)$ is reached, it is marked as $\upsilon(t_{s,n+1})$, and the current data sequence is expanded to $Y_{s,n+1}$, then there is:

$$p_{s,n+1} = U_{s,n+1}^T U_{s,n+1} = U_{s,n}^T U_{s,n} + U_{n+1}^T U_{n+1} = p_{s,n} + U_{n+1}^T U_{n+1} \qquad (7)$$

Where $U_{n+1} = \left[\, t_{s,n+1} \; 1 \,\right]$ .Similar to equation (7), for $q_{s,n+1}$ there is:

$$q_{s,n+1} = U_{s,n+1}^T Y_{s,n+1} = q_{s,n} + U_{n+1}^T \upsilon(t_{s,n+1}) \qquad (8)$$

In this case, the parameter vector of the regression model satisfies the recurrence equation

$$\theta_{s,n+1} = p_{s,n+1}^{-1} q_{s,n+1} = \theta_{s,n} + p_{s,n+1}^{-1} U_{n+1}^T (\upsilon(t_{s,n+1}) - U_{n+1}\theta_{s,n}) \qquad (9)$$

In order to avoid the inversion in Eq. (9), a matrix inversion lemma is introduced

$$[A + BCD]^{-1} = A^{-1} - A^{-1}B[DA^{-1}B + C^{-1}]^{-1} DA^{-1} \qquad (10)$$

.Let $A = p_{s,n}, B^T = D = U_{n+1}, C = 1, \beta_{s,n} = p_{s,n}^{-1}$ ,available

$$\beta_{s,n+1} = p_{s,n+1}^{-1} = \beta_{s,n} - (1 + U_{n+1}\beta_{s,n}U_{n+1}^T)^{-1} \beta_{s,n} U_{n+1}^T U_{n+1} \beta_{s,n} \qquad (11)$$

The initial value $\beta_{s,0} = \varsigma I$ , $\varsigma$ is a positive number, $I$ is $2 \times 2$ unit matrix. When the data sequence noise is larger, $\varsigma$ takes a smaller value; on the contrary, $\varsigma$ takes a larger value. Substituting (11) into (9) yields a recursive formula for regression model parameters: $\theta_{s,n+1} = \theta_{s,n} + \Delta\theta_{s,n}$

Where $$\Delta\theta_{s,n} = \beta_{s,n} U_{n+1}^T (1 + U_{n+1}\beta_{s,n}U_{n+1}^T)^{-1} (\upsilon(t_{s,n+1}) - U_{n+1}\theta_{s,n}) \qquad (12)$$

Where $(1 + U_{n+1}\beta_{s,n}U_{n+1}^T)$ is a scalar, thus avoiding matrix inversion. The initial value $\theta_{s,0}$ of Equation (12) is generally taken as zero vector.

## 3.2 Variable sliding window algorithm

The maximum length of the sliding window in the SW algorithm is not limited. When the detection threshold is relatively large, the length of the window may be so long that the trend analysis error becomes larger. However, the OSD algorithm defines the minimum sliding window length so that the mutation in the minimum sliding window point cannot be detected.

Aiming at the shortcomings of sliding window in SW and OSD algorithms, this paper presents a variable sliding window algorithm which dynamically changes the setting window length to segment the data sequence reasonably. The algorithm first sets the length of reference window and the length of the longest data window. Starting from the starting point of the current data segment, regression modeling is re-established for each newly arrived data stream element to improve accuracy. When the current data segment is smaller than the length of the reference window, the fitted standard deviation of the model is compared with the return value of the noise function G to detect whether there is an abnormal point in the reference window. If the length of the current data segment is greater than or equal to the reference window length, When the mean square error is greater than the standard segmentation threshold set in advance, the newly arrived data is considered as the segmentation point of the current data segment. If the length of the current data segment is greater than the length of the longest data window, the data point that has the closest fitting square-variances to the standard segmentation point threshold is searched as the dividing point of the current data segment from the beginning of the current data segment. The variable sliding window algorithm solves the fixed window problem of the SW algorithm and the OSD algorithm and realizes the reasonable segmentation of the data stream, so the accuracy of the trend analysis is improved.

## 4 Experiments results

In order to verify the validity of the proposed algorithm, the monitoring data collected in 60 days from 54 mica2 sensor nodes in Intel Berkeley Research Lab's sensor network lab are used. (mica2dot nodes: processor is Atmegal28L, RF chip is cc1000, flash is 128kb, RAM is 4kb and the transmission rate is 76.8kbps) The data set is sampled every 31s by temperature, humidity, light intensity and node voltage. The data is collected in the TinyDB network query processing system and the system is built on the TinyOS platform. Sensor layout shown in Figure 1, And gives the original data plot of individual sensors(figure2) and overlay sensors(figure3).

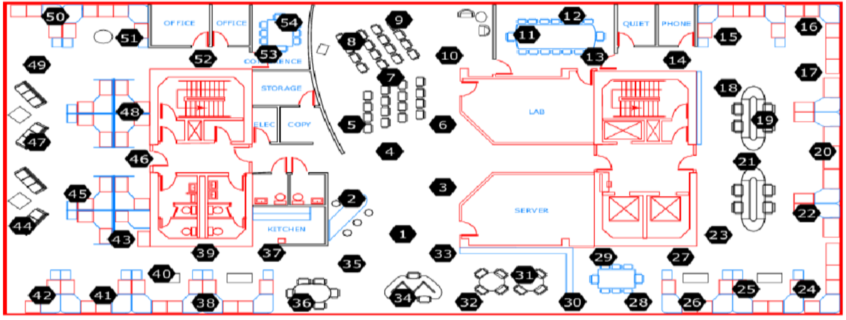

Figure 1 Sensor layout

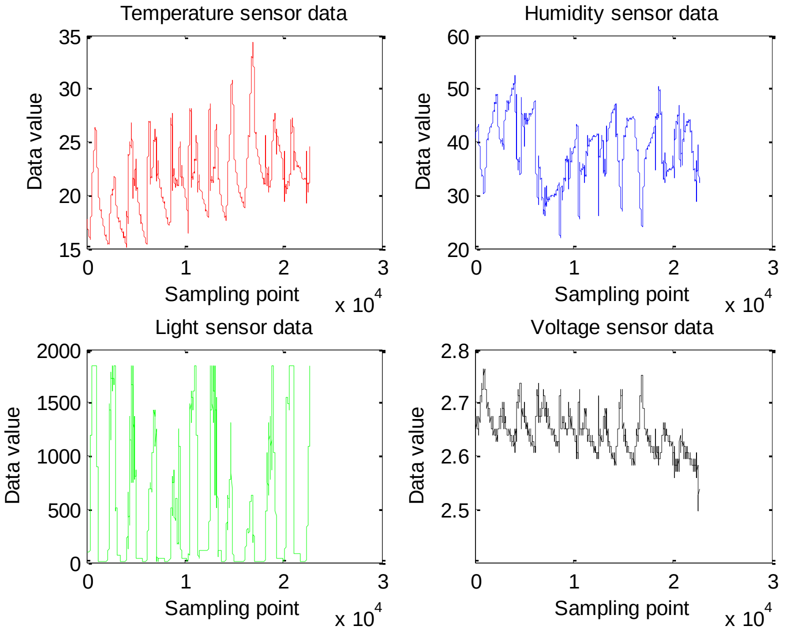

Figure2 The original data plot of individual sensors

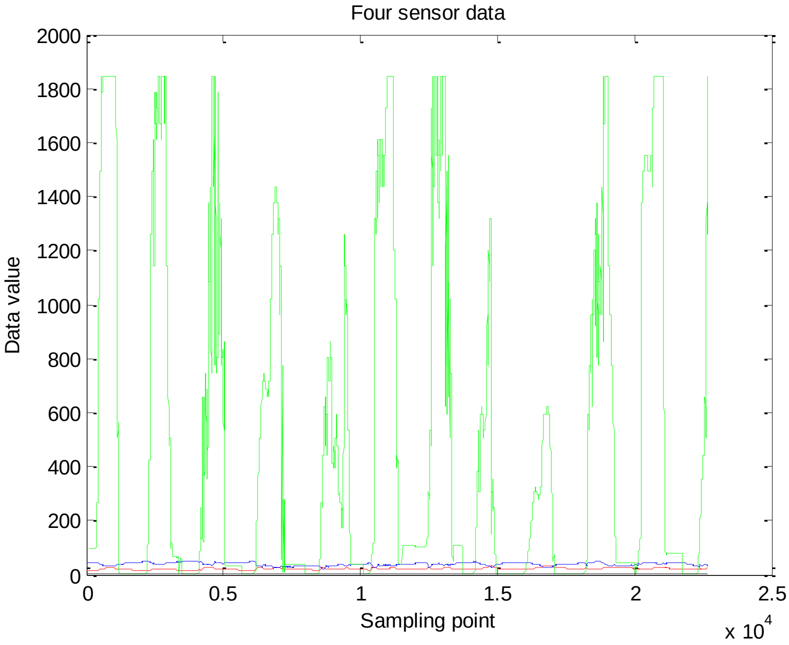

Figure3 The original data plot of overlay sensors

Based on the computer of P42.5GHZ,512M RAM and MATLAB R2014a simulation experiment,

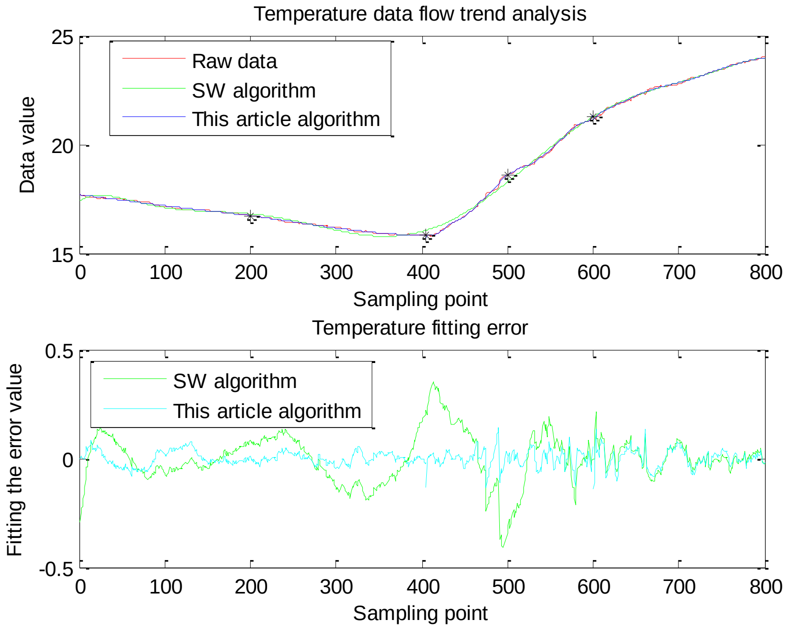

Figure 4 Temperature data flow trend analysis

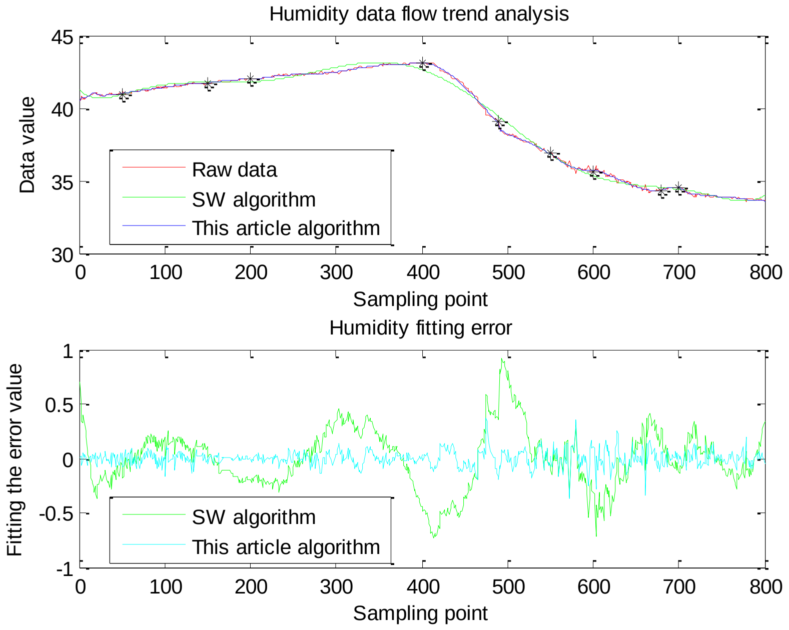

Figure 5 Humidity data flow trend analysis

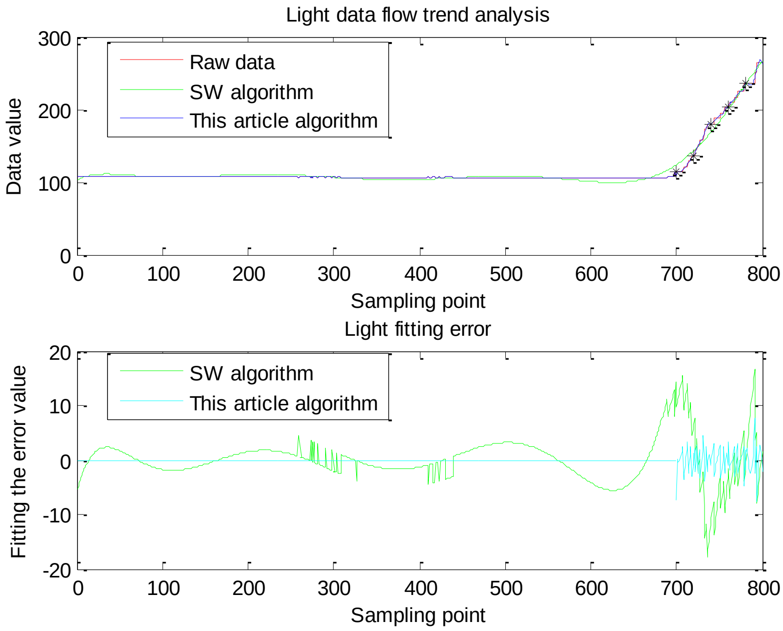

Figure 6 Light data flow trend analysis

Table 1 Three sensor linear modeling runtime based on two algorithms

| Data | This article algorithm | Sliding window(SW) |
|---|---|---|
| Temperature sensor data | 0.1719s | 1.6296s |
| Humidity sensor data | 0.1652s | 1.6699s |
| Light sensor data | 0.0788s | 1.9244s |

## 5. Conclusions

In this paper, incremental recursive least square method and variable sliding window algorithm are combined to propose a data stream real-time trend extraction algorithm. In order to solve the data flow elements that arrive constantly, this algorithm uses incremental mechanism to determine the data sequence regression model parameters and dynamically segment data segments, and extract the trend characteristics of data flow in real time. The algorithm not only calculated faster, but also higher accuracy. After this method is used to segment the data stream, the data can be extracted later to analyze and process a piece of data, which provides a good solution for some algorithms that cannot handle dynamic data nowadays, and lays a good data foundation for later target recognition.

## Acknowledge

This paper is supported by Natural Youth Science Foundation of China (61501326, 61401310), and Natural Science Foundation of China (61271411). It also supported by Tianjin Research Program of Application Foundation and Advanced Technology (15JCZDJC31500), and Tianjin Science Foundation(16JCYBJC16500).

## References

[1] Chenliang Wang, Xu Pang, Zhijian Lu, ea.al. Research on data flow classification based on dynamic feature extraction and neural network. Computer system and application,2010,30(6):1539-1542.
[2] Koski A, Juhola M, Meriste M. Syntactic recognition of ECG signals by attributed finite automata. Pattern Recognition,1995,28(12):1927-1940.
[3] Shatkay H, Zdonik S. Approximate queries and representations for large data sequences. Proceedings of 12th IEEE International conference on data engineering. Washington: IEEE computer society,1996:546-553.
[4] Keogh E, Chu S, Hart D, et al. Segmenting time series: A survey and novel approach. Proceedings of IEEE International conference on data mining. Los Jose: IEEE computer society, 2001:289-296.
[5] Sylvie C, Carlos G B, Cathering C, et al. Trends extraction and analysis for complex system monitoring and decision support. Engineering applications of artificial intelligence,2005,18(1):21-36.
[6] Qian Zhou, Tiejun Wu. Research and application of a data flow trend analysis method. Control and decision making,2008,23(10):1182-1185.
[7] Zhou Q, Cluett W. Recursive identification of time-varying systems via incremental estimation. Automatica,1996,32(10):1427-1431.
[8] Http://berkeley.intelresearch.net/lab data.