

Real-Time Segmenting Time Series Data

Aiguo Li, Shengping He, and Zheng Qin

Department of Computer Science, Xi'an Jiaotong University, 710049, Xi'an, Shaanxi, China
liag@xust.edu.cn, zhqin@xjtu.edu.cn

Abstract. There has been increased interest in time series data mining recently. In some cases, approaches of real-time segmenting time series are necessary in time series similarity search and data mining, and this is the focus of this paper. A real-time iterative algorithm that is based on time series prediction is proposed in this paper. Proposed algorithm consists of three modular steps. (1) Modeling: the step identifies an autoregressive moving average (ARMA) model of dynamic processes from a time series data; (2) prediction: this step makes k steps ahead prediction based on the ARMA model of the process at a crisp time point. (3) Change-points detection: the step is what fits a piecewise segmented polynomial regressive model to the time series data to determine whether it contains a new change-point. Finally, high performance of the proposed algorithm is demonstrated by comparing with Guralnik-Srivastava algorithm.

1 Introduction

There has been increased interest in time series data mining and similarity search recently [1-4,7-10]. The application background of segmenting time series methods includes using data-mining techniques to extract interesting patterns from time series data generated by sensors. Some batch or incremental algorithms have been proposed for segmenting time series [5,6]. However, in some real-time application situations, it is necessary for real-time detection of events from time series data. We will consider a real-valued time series denoted by $x_t, t = 1, 2, \dots$, where t is a time varying parameter. When a crisp observed value x_t is obtained at time point t , we need an algorithm to determine whether the time point t is a new change-point or not before next time point $t + 1$. The problem is the focus of this paper.

We propose two iterative real-time segmenting time series algorithms based on time series prediction. Proposed algorithms consist of three modular steps: Modeling, Prediction, and Change-point detection. (1) Modeling: the step identifies the ARMA (Autoregressive Moving Average) model of a dynamic process from time series data; (2) Prediction: this step predictions that future k time points states based on the ARMA model of the process. The k steps Kalman predictor of ARMA model is employed in this paper; (3) Change-points detection: the step is that fits a piecewise poly

nomial regressive model to a time segment, and maximum likelihood principles is applied to determine whether it contains a new change-point or not.

The remainder of the paper is organized as follows: section 2 describes the segmenting time series problem briefly. Section 3 presents the real-time segmenting time series algorithms. Section 4 describes experiments involved in comparing our algorithms with the batch algorithm proposed by Guralnik and Srivastava [6]. Finally, section 5 concludes the paper.

2 Real-Time Segmenting Time Series

It is supposed that a real valued time series $x_t, t = 1, 2, \dots, N$ can be modeled mathematically, where each model is characterized by a set of parameters, the segmenting time series problem becomes the change-point detection problem [6], so we don't discriminate segmenting time series from change-point detection in this paper.

2.1 Segmenting Time Series

Consider a real-valued time series denoted by

$$x_t, t = 1, 2, \dots, N \quad (1)$$

Where t is a time varies. We can find a piecewise segmented model M , given by

$$X = \begin{cases} f_1(t, w_1) + e_1(t), (0 < t \leq \alpha_1) \\ f_2(t, w_2) + e_2(t), (\alpha_1 < t \leq \alpha_2) \\ \dots \\ f_k(t, w_k) + e_k(t), (\alpha_{k-1} < t \leq \alpha_k = N) \end{cases} \quad (2)$$

Where $f_i(t, w_i), 1 \leq i \leq k$ is a basis class function (with its vector of parameters w_i) that is fit in segment i ; the vector $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_k)$ is the change points set of time series $x_t, t = 1, 2, \dots, N$; and $e_i(t), i = 1, 2, \dots, k$ is error term in i th segment.

Likelihood L is defined as below:

$$L = \sum_{i=1}^k l_i = \sum_{i=1}^k s_i \quad (3)$$

Where k is the number of change-points; l_i is the likelihood of i th segment; and s_i is the residual sum of squares for the model of i th segment. Here s_i is defined as below

$$s_i = \sum_{j=0}^{m_i} (x_{\alpha_{i-1}+j} - f_i(\alpha_{i-1} + j, w_i))^2 . \quad (4)$$

Where $m_i = \alpha_i - \alpha_{i-1}$ is the number of time points in i th segment.

The maximum likelihood estimate (MLE) of change points set $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_k)$ and parameters vector w_i of $f_i(t, w_i), i = 1, 2, \dots, k$ can be found by means of minimizing the likelihood L .

2.2 Real-Time Segmenting Time Series

Consider a time series defined in Eq. (1), we suppose that actual sample value x_1, x_2, \dots, x_{i-1} and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_k)$ have been known. When a crisp sample value x_i is obtained at sample time instant i , it is necessary to determine whether time instant i is a new change point or only a candidate by means of minimizing the likelihood L .

3 Proposed Algorithms

The real-time algorithms we have proposed consist of three modular steps: modeling, prediction, and change-point detection. The step of modeling identifies the ARMA model of a dynamic process from time series data. The ARMA model of the time series can be identified either in a priori or in iterative processes of the algorithms. The step of prediction is to predict future k time point states based on the ARMA model. Many approaches of modeling and prediction have been proposed [11]. In our experiments of the paper, the k steps Kalman predictor of ARMA model is employed (see section 4). The step of change-points detection is that fits a piecewise regressive model to a time segment, and maximum likelihood principles is applied to determine whether it contains a new change-point or not.

The basis idea of the real-time algorithms is that the k steps prediction are made which are denoted by $x_i(1), x_i(2), \dots, x_i(k)$ at every crisp sample i , then i is examined to see whether it is a new change point or a candidate, according to x_i and $x_i(1), x_i(2), \dots, x_i(k)$. The algorithms work under the assumption that a dynamic process can be described by an ARMA(n, m) model:

$$\varphi(q^{-1})x_t = \theta(q^{-1})a_t . \quad (5)$$

Where

$$\varphi(q^{-1}) = 1 + \varphi_1 q^{-1} + \varphi_2 q^{-2} + \dots + \varphi_n q^{-n} . \quad (6)$$

$$\theta(q^{-1}) = 1 + \theta_1 q^{-1} + \theta_2 q^{-2} + \dots + \theta_m q^{-m} . \quad (7)$$

The parameters vector of the ARMA model is

$$\Phi = (\varphi_1, \varphi_2, \dots, \varphi_n; \theta_1, \theta_2, \dots, \theta_n) . \quad (8)$$

3.1 The Algorithm A

Many approaches to identify the ARMA model of dynamic processes from time series data have been proposed [11]. If there are enough time series data, the ARMA model of the time series could be identified a priori. When the ARMA model of the time series is known a priori, the framework of the real-time segmenting algorithm is described as follows:

- 1) Given a time series Eq.(1); regressive model Eq(2), and model set Mset of $f_i(t, w_i), i = 1, 2, \dots, k$; ARMA model
- 2) Initialize $x_0, x_{-1}, \dots, x_{-n-m}; \alpha = \{ \alpha_1 = 0 \}, L=0$;
maximum prediction steps K ; new change point $ncp=0$.
- 3) for $t= 2:1:N$
 - a). k steps prediction: $x_t(1), x_t(2), \dots, x_t(k)$
Where $k \leq K$;
 - b). for $i= ncp:1:t$ % change point detection
Compute likelihood: $l1=l(ncp, t+k)$; and $l2=l(ncp, i)+ l(i, t+k)$;
if $((l1- l2)/ l1) > \mu$
 $ncp=i; l1= l2; l3= l(ncp, i)$;
end if
end for
 $\alpha = \alpha \cup ncp$; $L= L+ l3$;
- end for
- 4) Output α, L .
- 5) End.

3.2 The Algorithm B

If the ARMA model of a dynamic process is unknown, we have to on-line identify the ARMA model of the dynamic process. Many iterative algorithms of identifying the ARMA model have been proposed. An assumption is what the orders of the ARMA model are known a priori. In this case, the framework of the real-time segmenting algorithm is described as follows:

- 1) Given a time series Eq.(1); regressive model Eq(2), and model set Mset of $f_i(t, w_i), i = 1, 2, \dots, k$; the orders of ARMA model n and m
- 2) Initialize $x_0, x_{-1}, \dots, x_{-n-m}; \alpha = \{ \alpha_1 = 0 \}, L=0$;
maximum prediction steps K , new change point: $npc=0$.
- 3) for $t= n:1:N$
 - a) Identifying ARMA model parameters Φ_t
 - b) k steps prediction: $x_t(1), x_t(2), \dots, x_t(k)$
Where $k \leq K$
 - c) for $i= npc:1:t$ % change point detection
Compute likelihood: $l1= l(npc,t+k)$; and $l2= l(npc,i)+ l(i,t+k)$;
if $(l1- l2)/ l1 > \mu$
 $npc=i; l1= l2; l3= l(npc,i)$;
end if
end for
 $\alpha = \alpha \cup npc$; $L= L+ l3$;
- end for
- 4) Output α, L ;
- 5) End.

4 Experimental Results

The data used in our experiments is taken from a vibration experiment of a bus. In the raw data set D, the sample period is 5 milliseconds, and the data length is 512. At each sample instant, the amplitude of the bus vibration is recorded, and the unit of the amplitude is millimetre. The raw data set D is divided into two data subsets D1, and D2. D1 contains 400 data that come from the front elements of set D, and D2 contains 112 data that come from the rest data of set D. Set D, D1, and D2 are denoted respectively by $D= \{ d_1, d_2, \dots, d_{512} \}$; $D1= \{ d_1, d_2, \dots, d_{400} \}$; $D2= \{ d_{401}, d_{402}, \dots, d_{512} \}$. The raw data is shown in Figure 1.

Data set D1 is used to off-line identify the ARMA model of the time series for real-time segmenting algorithm A described in section 3.1. However, set D2 is used to examine the real-time algorithms described in above.

We are interested in how our real-time segmenting algorithms performed compared to the batch algorithm proposed by Guralnik and Srivastava [6].

For experimental purposes, the regressive model set Mset of $f_i(t, w_i)$, $i = 1, 2, \dots, k$ in Eq. (2) is a group of polynomials as below:

$$f_i(t, w_i) = 1 + w_{i,1}t + w_{i,2}t^2 + \dots + w_{i,p}t^p \tag{9}$$

$$i = 1, 2, \dots, k$$

In our experiments, $p_{max}=3$.

The k steps ahead prediction algorithm used in our experiments is a Kalman prediction algorithm. Supposed the ARMA(n, m) model is described by Eq. (5), (6), and $n=m$. The k steps ahead Kalman prediction algorithm is described as below:

$$\theta(q^{-1})x_t(k) = G_k(q^{-1})x_t \tag{10}$$

Where the coefficient vector of $G_k(q^{-1})$ is

$$\mathbf{g}_k = (g_{k1}, g_{k2}, \dots, g_{kn})^T = A^{k-1}K_P \tag{11}$$

$$A = \begin{bmatrix} -\varphi_1 & & & \\ -\varphi_2 & I_{n-1} & & \\ \dots & & & \\ -\varphi_n & 0 & \dots & 0 \end{bmatrix} \tag{12}$$

$$K_P = (\theta_1 - \varphi_1, \theta_2 - \varphi_2, \dots, \theta_n - \varphi_n)^T \tag{13}$$

Obviously, only would Eq. (13) be changed slightly, when $n \neq m$.

The experimental results are shown in Figure 2, 3, and 4, and Table 1, 2. Figure 2 shows the results of Guralnik-Srivastava (GS) batch algorithm; Figure 3 and Figure 4 show the results of our algorithms described respectively in section 3.

The orders of ARMA model the process used in our algorithms are $n= 6$ and $m= 5$. Box-Jenkins method is used to off-line identify the ARMA model of the process in the real-time segmenting algorithm A. However, iterative least square method is used to on-line identify the ARMA model of the process in the algorithm B.

According to the results in Table 1, the likelihood value L of GS batch algorithm is nearly 3 times than that of proposed algorithm A and B, the likelihood value L of the algorithm B is slightly larger than that of the algorithm A, so the segmenting result of algorithm A is best, and that of the GS batch algorithm is worst. The compute speeds of our algorithm A and B are great faster than that of GS batch algorithm. In fact, the memory demand of the algorithm A and B are less than that of batch algorithm too.

Fig. 5 shows daily closing price data of IBM stock from Jan. 1, 1980 to Oct. 8, 1992 (<http://www-personal.buseco.monash.edu.au/~hyndman/TSDL/korsan/dailyibm.dat>),

totally 3333 points. The 1000 points of the front of the raw dataset were used for estimating prediction model, and the remaining 2333 points for evaluating the proposed algorithms. ARMA(5, 4) models were used in algorithm A and B. The basis class functions were polynomial functions, and pmax=3. Because computing efficiency of batch GS algorithm is low in large dataset, incremental GS algorithm [6] is used for comprising with algorithm A and B. The experimental results are shown in Table 3. In Table 3, the likelihood value of algorithm A and B are very small, so the evaluating results of algorithm A and B are better than that of GS algorithm.

5 Conclusions

In this paper, we presented two real-time segmenting time series algorithms that based on time series prediction. We have analyzed how ARMA model of a process could be used in segmenting time series. The experiments show that proposed algorithms are superior to Guralnik-Srivastava batch algorithm.

Table 1. Comparison of likelihood estimation and run time of the three algorithms

Algorithm	μ	L	CPU Time (sec)
GS Algorithm	0.02	4747.4	41.08
Algorithm A	0.9	1052.9	2.03
Algorithm B	0.9	1190.4	3.51

Table 2. The results of segmenting the bus vibration data by the three algorithms respectively

Algorithm	Chang Point Set
GS Algorithm	0, 13, 21, 29, 37, 44, 57, 66, 74, 84, 97, 105, 112
Algorithm A	0, 7, 14, 21, 28, 35, 43, 50, 57, 73, 80, 87, 94, 101, 108
Algorithm B	0, 7, 15, 25, 32, 44, 51, 58, 65, 72, 79, 86, 93, 100, 108

Table 3. The results of segmenting daily closing price of IBM stock with GS algorithm and proposed algorithms respectively

Algorithm	μ	L	CPU Time (sec)	Number of change points
GS Algorithm	0.02	155080	14572	10
Algorithm A	0.2	1497.9	14.44	328
Algorithm B	0.9	1437.6	38.05	334

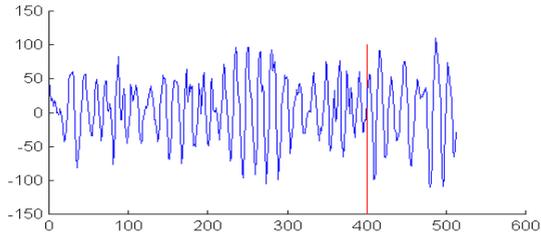


Fig. 1. Bus vibration time series data

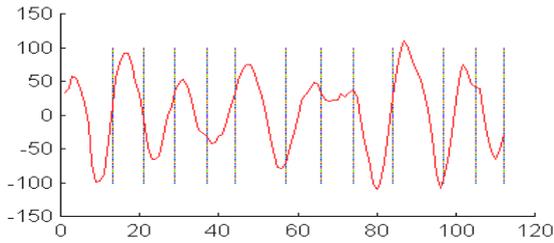


Fig. 2. Result of segmenting bus vibration time series using GS batch algorithm

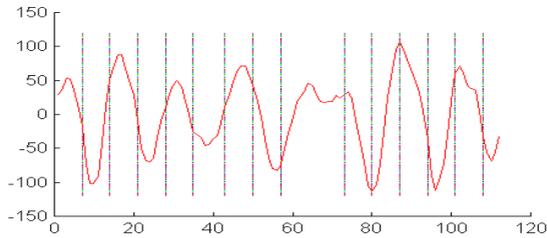


Fig. 3. Result of segmenting bus vibration time series using the algorithm A

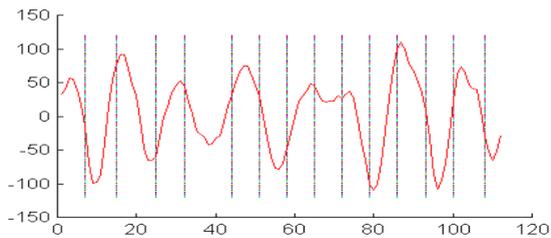


Fig. 4. Result of segmenting bus vibration time series using the algorithm B

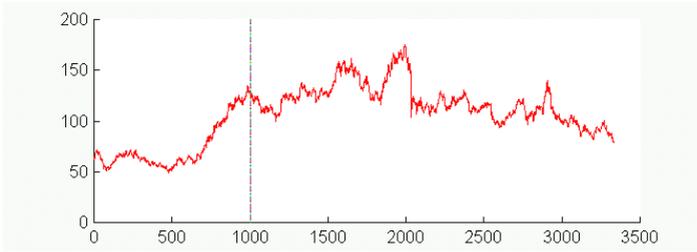


Fig. 5. Daily closing price time series data of IBM stock

References

1. Agrawal R., Faloutsos C., Swami A.: Efficient Similarity Search In Sequence Databases. In: Proc of the 4th Conf on FODO, 1993, 69–84
2. Shim K., Srikant R., Agrawal R.: High-Dimensional Similarity Joins. *IEEE Trans. on Knowledge and Data Engineering* (2002) 14(1): 156–171
3. Rafiai D., Mondelzon A. O.: Querying Time Series Data Based on Similarity. *IEEE Trans on Knowledge and Data Engineering* (2000) 12(5): 675–693
4. Keogh E., Chakrabarti K., Pazzani M., Mehrotra S.: Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. *Knowledge and Information Systems* (2001) 3(3): 263–286
5. Keogh E., Chu S., Hart D., et al.: An Online Algorithm for Segmenting Time Series. In: *IEEE Int'l Conf on Data Mining* (2001)
6. Guralnik V., Srivastava J.: Event Detection from Time Series Data. In: *Proc of SIGKDD* (1999) 33–42
7. Agrawal R., Lin K. I., Sawhney H. S., Shim K.: Fast Similarity Search in the Presence of Noise, Scaling, and translation in Time-Series Databases. In *Proc of the 21st VLDB* (1995) 490–50
8. Faloutsos C., Ranganathan M., Manolopoulos Y.: Fast Subsequence Matching in Time-Series Databases. In *Proc. of the ACM SIGMOD Conf. on Management of Data* (1994) 419–429
9. Chan K. P., Fu A. W.: Efficient Time Series Matching by Wavelets. In *Proc of the 15th ICDE* (1999)
10. Perng C. S., Wang H. X., Zhang S. R., et al: Landmarks: A New Model for Similarity-Based Pattern in Time Series Databases. In *Proc of the 16th IEEE Int'l Conf on Data Engineering* (2000) 475–693
11. Kantz H., Schreiber T.: *Nonlinear Time Series Analysis*, Cambridge University Press, Cambridge, England (1997)