在线分割时间序列数据*

李爱国[1,2+], 覃 征[2]

[1](西安科技大学 计算机科学技术系,陕西 西安 710054)
[2](西安交通大学 计算机科学技术系,陕西 西安 710049)

On-Line Segmentation of Time-Series Data

LI Ai-Guo[1,2+], QIN Zheng[2]

[1](Department of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an 710054, China)
[2](Department of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China)
+ Corresponding author: Phn: +86-29-82663979, E-mail: liag@xust.edu.cn
Received 2003-07-02; Accepted 2004-02-05

Li AG, Qin Z. On-Line segmentation time-series data. *Journal of Software*, 2004,15(11):1671~1679.
http://www.jos.org.cn/1000-9825/15/1671.htm

Abstract: Segmentation of time series is one of the important tasks in time series data mining. Segmentation has two major uses: It may be performed either to detect when the system that creates the time series has changed or to create a high level representation of the time series for indexing, clustering, and classification. Approaches to on-line segmentation of time series are necessary when identifying and predicting temporal patterns in real-time time series databases are needed, and this is the focus of this paper. A formal description of segmenting time series problem and a criterion for the evolution of segmentation algorithms are presented. An on-line iterative algorithm of segmenting time series, called OLS (on-line segmentation), is then proposed. OLS is independent of a priori knowledge about the segmented time series. Experimental results demonstrate that OLS can on-line detect the critical change points of time series with less 'over fit' than that of competitive algorithms.
Key words: data mining; knowledge acquisition; time series; segmentation

摘 要: 时间序列分割是时间序列数据挖掘研究的重要任务之一.它主要有两个应用:检测生成时间序列的系统何时发生变化;创建时间序列的高级数据表示,从而对时间序列进行索引、聚类和分类.在实时时间序列数据挖掘应用中,需要在线时间序列分割算法,以便实时发现和预测时态模式.在对时间序列分割问题进行形式化描述的基础上,提出了一种评估时间序列的分割结果以及分割算法性能的评价指标,并提出了一种在线分割时间序列数据的递推算法(on-line segmentation,简称 OLS).OLS 的一个显著特点是不依赖有关时间序列的先验知识.实验结果说明,OLS 算法能够有效地在线检测出数据挖掘应用中感兴趣的关键变化点,而且"过拟合"程度低.
关键词: 数据挖掘;知识获取;时间序列;分割

* Supported by the Key Science-Technology Project of the 'Tenth Five-Year-Plan' of Shaan'xi Province of China under Grant No.2000K08-G12 (陕西省科学技术发展计划"十五"攻关项目)

作者简介: **李爱国**(1966－),男,甘肃张掖人,博士,副教授,主要研究领域为机器学习,数据挖掘,信息融合;**覃征**(1956－),男,博士,教授,博士生导师,主要研究领域为复杂环境下自适应信息处理,信息融合,计算机系统集成与电子商务,分布式并行信息处理.

中图法分类号: TP311 文献标识码: A

时间序列数据在一些新的数据库应用,如数据仓库以及数据挖掘等领域中日益重要.与传统的统计分析方法不同,在这些应用中,人们试图基于某种相似性度量,从时间序列数据中抽取感兴趣的模式,以便进行查询、分析和发现规则等处理.虽然根据具体应用的不同,对"模式"的定义会有所区别,但是一个共同点是将模式定义为一个相似时间序列的集合.例如,在给定时间段内,那些价格大幅下降的股票价格序列数据的集合就是一个模式(大幅下降模式),其中每支股票的价格数据就是该模式的一个样本.又例如,同一支股票数据序列,连续 5 个交易日价格单调上升的子序列的集合构成一个模式(单调上升模式).

从时间序列数据抽取模式的一般方法是,先将原始时间序列分割,并将所得的子序列转换为某种高级的数据表示,如符号序列或者某个特征空间中的点,然后在此符号序列或者特征空间中进行聚类(或分类),生成模式或模式集合[1,2].其中关键问题之一是如何分割时间序列数据.

文献[1]指出,时间序列分割主要有两个应用:系统模型变化检测,即当产生时间序列的系统的模型(或参数)发生变化时,应用分割算法可以检测到这种变化是何时发生的;应用分割算法创建时间序列的高级数据表示,以便对时间序列进行索引、聚类和分类.因此,研究时间序列分割算法具有重要的理论意义和实际应用价值,并已成为时间序列数据挖掘研究的主要任务之一[1].

文献[3]研究用一组人工神经网络模型分割时间序列.而在时间序列数据挖掘研究中,常采用分段直线表示(piecewise linear representation,简称 PLR)方法分割时间序列[4].因为 PLR 法比较符合人们的直观经验,而且通常索引结构维数低、计算速度较快,所以被较多人采用.但是,正如文献[4]所指出的,对各种 PLR 法的深入研究还比较欠缺.文献[4]对 3 种 PLR 法做了实验比较研究,但是缺乏系统的理论分析.文献[5]提出了基于分段回归分析技术的时间序列分割算法,也没有给出理论分析.文献[2]系统地给出了一种基于分段多项式回归分析技术的最优时间序列分割及高级数据表示方法(piecewise polynomial representation,简称 PPR),并据此系统化地研究了时间序列相似性比较和模式抽取方法.文献[2]证明了这种基于分段多项式回归分析的高级数据表示以及相似模式发现技术具有与 DFT(discrete Fourier transform)方法和 DWT(discrete wavelet transform)方法一样好的数学性质,而且也证明某些 PLR 方法实际上是这种方法的特例.这一研究结果为基于分段多项式回归的时间序列高级数据表示和相似模式发现技术奠定了理论基础.这种技术的基本思路是用一个分段回归模型近似时间序列数据,从而"自然"地把时间序列数据分割为一个不重叠的有序子序列集合.文献[6]进一步改进了文献[5]的算法,大幅度提高了时间序列分割算法的计算效率.

在实时时间序列数据挖掘的应用场合,需要对实时得到的时间序列数据进行在线分割,以便实时发现和预测时态模式.但是,上述算法均不适合此种应用.文献[7]针对可预测的时间序列,探讨了在线分割时间序列的问题,并提出了两种基于时间序列多步预测的实时分割算法.尽管时间序列预测的方法有很多[8,9],但是许多时间序列是不可预测的,因此有必要研究更加一般的分割算法.

本文在对时间序列分割问题进行形式化描述的基础上,研究了评估时间序列分割结果以及分割算法的评价指标,并提出了一种在线分割时间序列数据的递推算法(on-line segmentation,简称 OLS).对比实验结果说明,OLS 算法能够有效地在线检测出数据挖掘应用中感兴趣的关键变化点,而且"过拟合"程度低.

1 时间序列分割

1.1 时间序列分割问题的形式化描述

为了叙述简便,我们使用如下记号:

X :长度为 N 的时间序列的集合; M :候选的模型集合,且 $M \neq \varnothing$; $\varnothing$ 是空集合.

定义 1. $\forall \boldsymbol{x} \in X$, $\exists P \in M$,使得 $\hat{\boldsymbol{x}} = P(\boldsymbol{x})$,称 $\hat{\boldsymbol{x}}$ 为原始序列 $\boldsymbol{x}$ 经由模型 P 产生的时间序列.

定义 2. 给定阈值 $\varepsilon > 0$ 以及距离度量 d,如果 $\forall \boldsymbol{x} \in X$, $\exists P \in M$,有

$$d(\hat{\boldsymbol{x}}, \boldsymbol{x}) \leq \varepsilon \tag{1}$$

成立,则称模型 P 与 $\boldsymbol{x}$ 近似,简记为 $P \cong \boldsymbol{x}$.

定义 3. $\forall \boldsymbol{x} \in X$,找到一个有序的模型集合 $Q=(Q_1,Q_2,...,Q_K)$,$Q_i \in M$,$i=1,2,...,K$,将 $\boldsymbol{x}$ 分割成不重叠的有序子序列集合 $\boldsymbol{s}=(\boldsymbol{s}_1,\boldsymbol{s}_2,...,\boldsymbol{s}_K)$,且满足

$$Q_i \cong \boldsymbol{s}_i, \quad i=1,2,...,K \tag{2}$$

称 Q 为 $\boldsymbol{x}$ 的一种分割.

根据时间序列分割的两个不同应用,对分割算法也有不同的要求.在监测系统变化的应用中,要求分割算法能识别出系统模型变化的时刻(在统计学中称为变化点检测问题);在对时间序列进行索引、聚类和分类等应用中,要求分割算法"合理"地将时间序列分割成一组不重叠的子序列.本文研究的分割算法针对后一种应用.那么,如何评价分割结果是否"合理",换言之,如何评估不同的分割算法对同一时间序列分割的结果的优劣呢?这就需要一个客观的评价指标.

定义 4. $\forall \boldsymbol{x} \in X$,如果存在使某个性能函数

$$J=\sum_{i=1}^{K} J_i\big|_{Q_i \cong \boldsymbol{s}_i} \to \min \tag{3}$$

的一个有序模型集合 $Q=(Q_1,Q_2,...,Q_K)$,$Q_i \in M$,$i=1,2,...,K$,将 $\boldsymbol{x}$ 非平凡地分割成不重叠的有序子序列集合 $\boldsymbol{s}=(\boldsymbol{s}_1,\boldsymbol{s}_2,...,\boldsymbol{s}_K)$,则称 $\boldsymbol{s}$ 是 $\boldsymbol{x}$ 的最优分割.

下面,我们给出一种基于最小均方误差的最优分割问题描述.

1.2 最优时间序列分割

考虑一个一维实值时间序列 $\boldsymbol{x}=x_1,x_2,...,x_N$,设 $\boldsymbol{x}$ 能用如下分段模型描述:

$$\boldsymbol{x}=\begin{cases} f_1(t,\boldsymbol{w}_1)+e_1(t), & 1 \le t < \alpha_1 \\ ... & \\ f_k(t,\boldsymbol{w}_k)+e_k(t), & \alpha_{k-1} \le t < \alpha_k = N+1 \end{cases} \tag{4}$$

其中,$f_i(t,\boldsymbol{w}_i) \in M, 1 \le i \le k$;$\boldsymbol{w}_i$ 是其系数向量;候选模型集合 M 可以有许多种形式,如线性多项式模型、小波变换模型以及傅立叶变换模型等;$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,...,\alpha_k)$ 是 $\boldsymbol{x}$ 的时间分割点集合,而 $e_i(t), i=1,2,...,k$ 是零均值白噪声.

定义均方误差为

$$L=\sum_{i=1}^{k} l_i=\sum_{i=1}^{k}\sum_{j=0}^{m_i}(x_{\alpha_{i-1}+j}-f_i(\alpha_{i-1}+j,\boldsymbol{w}_i))^2 \tag{5}$$

其中,$m_i=\alpha_i-\alpha_{i-1}$ 是第 i 段的采样点数目.

时间序列最优分割就是使式(5)定义的均方误差 L 达到最小的分割.

1.3 分割结果的评价指标

对一个时间序列的分割结果是否合理?如何比较不同分割算法对同一时间序列的分割结果?这是困扰人们的两个问题.

在统计学家们感兴趣的时间变化点检测研究中,通常根据人类专家的意见评价分段的合理性[10];而在数据挖掘领域,有些研究者采用式(5)作为评价指标[5~7],或者式(5)结合人类专家意见的评价方法[5].然而,有些研究者却忽视了这个问题[4].

式(5)刻画了分段模型近似被分割时间序列的程度.L 值越小,说明分段模型越近似被分割时间序列.可见,用式(5)评价分割结果是有一定的合理性的.可是,如果我们平凡地将时间序列 x 分割为 $N/2$ 个子序列,每个子序列包含 2 个采样值,并对子序列用直线模型近似,那么 $L=0$,达到了最小值.但是,这样的分割是没有意义的.因此,用式(5)评价分割结果也有其局限性.

一个好的分割结果应满足什么条件呢?我们提出如下假设:

假设. 一个合理的分割结果应当是以尽可能少的子序列数获得尽可能小的 L 值.

根据这个假设,提出如下的评价指标:

$$J=\begin{cases}\dfrac{L}{\sum_{i=1}^{N}(x_i-\bar{x})^2}+\dfrac{k}{N}, & \sum_{i=1}^{N}(x_i-\bar{x})^2>0\\ \dfrac{k}{N}, & \sum_{i=1}^{N}(x_i-\bar{x})^2=0\end{cases} \tag{6}$$

其中,L 由式(5)定义;k 是分段数,$1\le k\le N/2$;$\bar{x}$ 是 $\boldsymbol{x}$ 的均值.

下面分 3 种情况加以讨论:

(1) 当 $\sum_{i=1}^{N}(x_i-\bar{x})^2=0$ 时,表明时间序列 $\boldsymbol{x}$ 是一条水平线,此时

$$J=\frac{k}{N} \tag{7}$$

当 $k=1$ 时,$J_{\min}=k/N$,即 $\boldsymbol{x}$ 不宜分段.

(2) 当 $\sum_{i=1}^{N}(x_i-\bar{x})^2>0$ 时,令

$$\hat{\boldsymbol{s}}_j=f_j(t,\boldsymbol{w}_j),\ j=1,2,...,k;\ \alpha_{j-1}\le t<\alpha_j \tag{8}$$

式(5)改写成

$$L=\sum_{j=1}^{k}(\boldsymbol{s}_j-\hat{\boldsymbol{s}}_j)^2=\sum_{j=1}^{k}\langle\boldsymbol{e}_j,\boldsymbol{e}_j\rangle \tag{9}$$

其中,$\langle\bullet,\bullet\rangle$ 是向量的内积符号.

不失一般性,假设式(4)中的 M 是正交变换模型集合,则

$$\sum_{i=1}^{N}(x_i-\bar{x})^2=\sum_{j=1}^{k}\langle\hat{\boldsymbol{s}}_j+\boldsymbol{e}_j-\mathbf{1}\bar{x},\hat{\boldsymbol{s}}_j+\boldsymbol{e}_j-\mathbf{1}\bar{x}\rangle=\sum_{j=1}^{k}(\langle\boldsymbol{e}_j,\boldsymbol{e}_j\rangle+\langle\hat{\boldsymbol{s}}_j-\mathbf{1}\bar{x},\hat{\boldsymbol{s}}_j-\mathbf{1}\bar{x}\rangle+2\boldsymbol{e}_j^T(\hat{\boldsymbol{s}}_j-\mathbf{1}\bar{x})) \tag{10}$$

其中,$\mathbf{1}$ 表示与 $\hat{\boldsymbol{s}}_j$ 和 $\boldsymbol{e}_j$ 相同维数的元素全为 1 的向量.注意到 $\bar{x}$ 是常数,因 M 是正交变换模型集合,故

$$2\boldsymbol{e}_j^T(\hat{\boldsymbol{s}}_j-\mathbf{1}\bar{x})=0,\ j=1,2,...,k \tag{11}$$

于是

$$\sum_{i=1}^{N}(x_i-\bar{x})^2=\sum_{j=1}^{k}(\langle\boldsymbol{e}_j,\boldsymbol{e}_j\rangle+\langle\hat{\boldsymbol{s}}_j-\mathbf{1}\bar{x},\hat{\boldsymbol{s}}_j-\mathbf{1}\bar{x}\rangle) \tag{12}$$

结合式(9)和式(12),有

$$J=\frac{\sum_{j=1}^{k}\langle\boldsymbol{e}_j,\boldsymbol{e}_j\rangle}{\sum_{j=1}^{k}(\langle\boldsymbol{e}_j,\boldsymbol{e}_j\rangle+\langle\hat{\boldsymbol{s}}_j-\mathbf{1}\bar{x},\hat{\boldsymbol{s}}_j-\mathbf{1}\bar{x}\rangle)}+\frac{k}{N} \tag{13}$$

易知,当 $\sum_{j=1}^{k}\langle\boldsymbol{e}_j,\boldsymbol{e}_j\rangle=0$ 且 $k=1$ 时,

$$J_{\min}=\frac{1}{N} \tag{14}$$

而当 $\sum_{j=1}^{k}\langle\hat{\boldsymbol{s}}_j-\mathbf{1}\bar{x},\hat{\boldsymbol{s}}_j-\mathbf{1}\bar{x}\rangle=0$ 且 $k=N/2$ 时,

$$J_{\max}=\frac{3}{2} \tag{15}$$

综合式(14)和式(15),得到:

$$\frac{1}{N}\le J\le\frac{3}{2} \tag{16}$$

(3) 如果 M 不是正交变换模型集合,因为 $\sum_{i=1}^{N}(x_i-\bar{x})^2>0$,则式(14)仍然成立.

2 在线分割算法

在线分割时间序列要解决的一个关键问题是:

问题 1. 设在当前采样时刻 t,时间序列的值是 x_t,距离 t 最近的一个变化点是 r,判定 t 是否为一个新的变化点.

如果 t 是一个变化点,则 $\boldsymbol{s}_t=(x_r,x_{r+1},...,x_{t-1},x_t)$ 被 t 分割为两个子序列: $\boldsymbol{s}_{t1}=(x_r,x_{r+1},...,x_{t-1})$ 和 $\boldsymbol{s}_{t2}=(x_t)$.

引理 1. $\forall \boldsymbol{x}\in X$,存在 $\boldsymbol{x}$ 的近似模型 P,使得残差序列

$$\boldsymbol{e}=\boldsymbol{x}-\hat{\boldsymbol{x}} \tag{17}$$

服从零均值正态分布,即 $\boldsymbol{e}\sim N(0,\sigma^2)$.

证明:线性多项式回归模型就是一类这样的模型,因此引理成立. □

基于引理 1,我们可以找到子序列 $\boldsymbol{s}_{t1}$ 和 $\boldsymbol{s}_t$ 的近似模型 P_{t1}, P_t 使得残差序列

$$\boldsymbol{e}_{t1}=\boldsymbol{s}_{t1}-P_{t1}(\boldsymbol{s}_{t1}) \tag{18}$$

和

$$\boldsymbol{e}_t=\boldsymbol{s}_t-P_t(\boldsymbol{s}_t) \tag{19}$$

均服从零均值正态分布,即 $\boldsymbol{e}_{t1}\sim N(0,\sigma_{t1}^2)$, $\boldsymbol{e}_t\sim N(0,\sigma_t^2)$.

于是,问题 1 可以等价为问题 2.

问题 2. 残差序列 $\boldsymbol{e}_t$ 是否分割为 $\boldsymbol{e}_{t1}$ 和 $\boldsymbol{e}_{t2}$ 这两段更好?

$\boldsymbol{e}_{t2}$ 是由 $\boldsymbol{s}_{t2}$ 产生的残差序列,显然, $\boldsymbol{e}_{t2}=0$.

问题 2 就是统计学中变化点检测问题.回答了问题 2,就解决了问题 1.

借鉴文献[10]的思想,我们提出解决问题 2 的方法如下:

设有序列 $\boldsymbol{e}=(e_1,e_2,...,e_m)$,且 $\boldsymbol{e}\sim N(0,\sigma^2)$,令

$$F(1,m)=Q(0,m) \tag{20}$$

是 $\boldsymbol{e}$ 只分成 1 段的似然函数;而

$$F(2,m)=F(1,m-1) \tag{21}$$

是将 $\boldsymbol{e}$ 从 m 处分为 2 段的似然函数.其中,

$$Q(h,m)=(m-h)\lg(S^{*2}) \tag{22}$$

式中 S^{*2} 是序列 $e_{h+1},e_{h+2},...,e_m$ 的样本方差.如果

$$F=\frac{(m-2)[F(1,m)-F(2,m)]}{F(2,m)}>\delta \tag{23}$$

则将 $\boldsymbol{e}$ 从 m 处分为 2 段更合理. $\delta>0$ 是阈值.

在式(23)中, F 服从 $F_\alpha(m,m-1)$ 分布.当给定显著水平 $\alpha=0.10$ 时, $F_\alpha(m,m-1)$ 的值在(1,9.16)之间,因此 δ 可取 1~10 之间的值.我们根据大量实验结果发现,这样选取 δ 的值在多数情况下较好.

下面给出在线分割时间序列算法流程如下:

给定时间序列 $\boldsymbol{x}$;候选模型集合 M;给定允许的最小子序列长度 c;变化点集合 cp.

在每一个当前采样时刻 t; DO

if $t-r\geq c$

选择合适的模型近似 $\boldsymbol{s}_t$ 和 $\boldsymbol{s}_{t1}$;

计算 S_t^{*2} 和 S_{t1}^{*2};计算 $F(r,t)$ 和 $F(r,t-1)$;计算 F;

if $F>\delta$

$cp = cp \cup t$; $r = t$;

end if

end if

end do

3 实 验

3.1 模型选择

在实验中,我们选择的候选模型集合 M 是多项式模型,即 $\boldsymbol{s}_t$ 和 $\boldsymbol{s}_{t1}$ 在最小均方误差意义下用如下多项式函数来近似:

$$f(t,\boldsymbol{w}) = w_0 + w_1 t + w_2 t^2 + ... + w_p t^p \tag{24}$$

即将 $\boldsymbol{s}_t$ 和 $\boldsymbol{s}_{t1}$ 分别映射到多项式基$\{1,t^1,\ldots,t^p\}$张成的 $p+1$ 维特征空间中,有

$$\hat{\boldsymbol{s}}_t = P_1(\boldsymbol{s}_t) \tag{25}$$

$$\hat{\boldsymbol{s}}_{t1} = P_2(\boldsymbol{s}_{t1}) \tag{26}$$

而相应的残差序列分别为

$$\boldsymbol{e}_t = \boldsymbol{s}_t - \hat{\boldsymbol{s}}_t \tag{27}$$

$$\boldsymbol{e}_{t1} = \boldsymbol{s}_{t1} - \hat{\boldsymbol{s}}_{t1} \tag{28}$$

$\boldsymbol{e}_t$ 和 $\boldsymbol{e}_{t1}$ 服从零均值正态分布 $\boldsymbol{e}_t \sim N(0,\sigma_1^2)$, $\boldsymbol{e}_{t1} \sim N(0,\sigma_2^2)$.

3.2 实验结果

实验数据是 IBM 公司 1980 年 1 月 1 日~1992 年 10 月 8 日的每日收盘价时间序列(http://www-personal.buseco.monash.edu.au/~hyndman/TSDL/korsan/dailyibm.dat),共 3 333 个数据.该时间序列数据的曲线如图 1 所示.

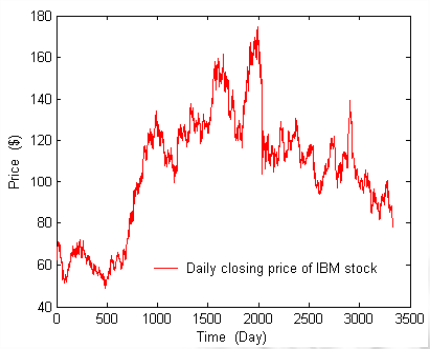


Fig.1 Sequence of daily closing price of IBM stock

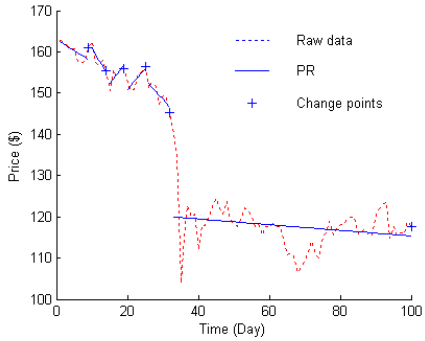
图 1 IBM 公司每日股票收盘价格序列

为了便于视觉比较,首先用较小的数据集进行实验.选取第 2001~第 2100 的 100 个数据作为一个序列(记做 T)进行在线分割实验.实验中分别选取多项式的阶 p 为 1 或者 2;阈值 δ 的值为 1 或 2.实验结果见表 1、图 2(a)、图 2(b)和图 2(c).作为对比的算法是滑动窗口(SW)算法[4]、自底向上(BU)算法[4]和 ROS 算法[6].它们的阈值是 α.实验结果见表 2、图 2(d)~图 2(g).需要说明的是,SW 算法、BU 算法和 ROS 算法并不是在线分割算法,但却是数据挖掘中较有代表性的分割算法.我们尽可能地使每种算法的阈值取得比较合适.在实验中,SW 算法和 BU 算法的阈值均取 0.7,而 ROS 算法的阈值取 0.002.

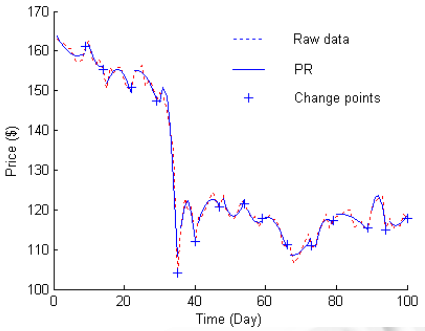
Table 1 Segmentations of time series T using OLS algorithm

表 1 OLS 算法对时间序列 T 作分割的结果

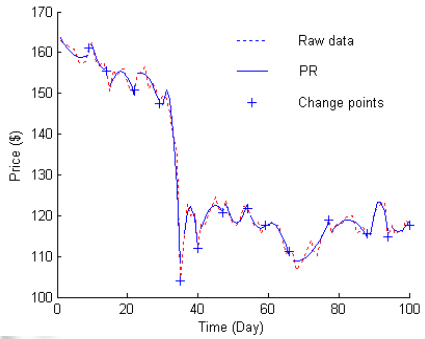
p	δ	L	Number of segmentation	J
1	2	1 948.5	6	0.118 0
2	1	233.6	15	0.156 9
2	2	237.4	14	0.147 1



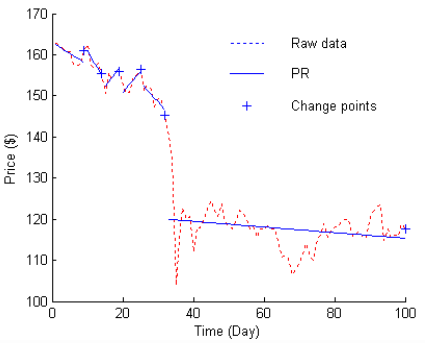
(a) OLS p=1, δ=2



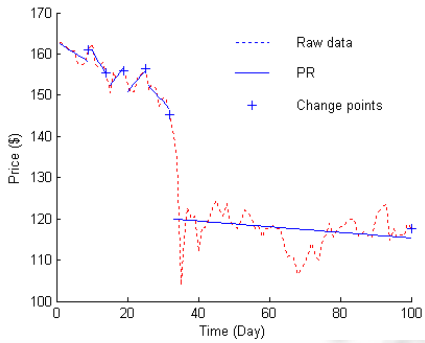
(b) OLS p=2, δ=1



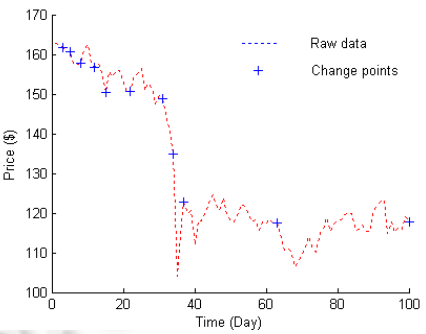
(c) OLS p=2, δ=2



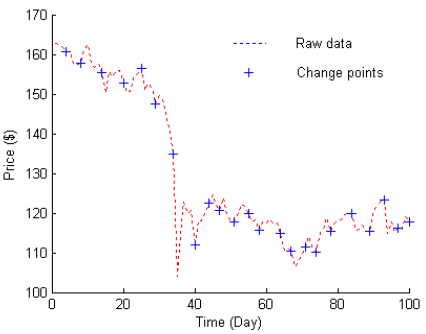
(d) SW



(e) BU



(f) ROS p=1



(g) ROS p=2

Fig.2 Segmentations of time series T

图 2 对时间序列 T 作分割的结果

Table 2 Comparison of the segmentations of time series T using different algorithms

表 2 对比不同算法对时间序列 T 的分割结果

Algorithms	L	Number of change-point	J
SW	517.5	24	0.255 4
BU	370.7	31	0.321 0
ROS (p=1)	541.2	11	0.126 1
ROS (p=2)	41.4	23	0.231 2

比较表 1 和表 2 中每种情况的 J 值可见,本文提出的在线分割算法 OLS 的分割结果相对来说最好,ROS 算法的分割结果居中,而 SW 算法和 BU 算法的分割结果最差.

图 2(a)~图 2(g)各图中的虚线表示原始时间序列 T 的曲线,实线段表示由模型产生的近似曲线;"+"表示变化点的位置.从图中可见,原始数据明显地分为两个大的阶段:高位震荡下滑阶段和低位徘徊阶段.这两个阶段的分界线大致出现在 31~33 的时间范围内.从统计学的变化点检测的角度来看,这是一个重要的时间变化点,因为这是两个大阶段的分水岭.我们看到,ROS 算法准确地检测出了这个变化点.而其他算法未能检测出这个变化点.从数据挖掘的角度来看,这个变化点则不是最重要的,因为此时已经无法规避风险了,最重要的变化点应在 28~30 的时间范围内,因为这个变化点预示着第 1 个大阶段结束,第 2 个大阶段即将开始.我们看到,只有图 2(b)、图 2(c)和图 2(g)准确地检测出了这个变化点,而图 2 中的其余各图只检测出了与之接近的时间点.说明 OLS 算法和 ROS 算法的性能相对较好.另外,需要指出的是,图 2(g)中的变化点数目要比图 2(b)和图 2(c)中的多很多,这显示 ROS 算法是靠"过拟合"的办法检测到 28~30 的时间范围内的这个关键变化点的.

表 3 给出了 OLS 算法、SW 算法和 BU 算法对整个 IBM 股票价格序列的分割结果,可见 OLS 算法的 J 值最小.

Table 3 Segmentations of daily closing price time series of IBM stock using different algorithms

表 3 不同算法对 IBM 每日股票收盘价格时间序列分割的结果

Algorithms	δ/α	L	Number of segmentation	J
ROS (p=1)	2	7 094.7	329	0.101 3
ROS (p=2)	2	3 302.6	283	0.086 1
SW	0.7	3 082.6	466	0.140 9
BU	0.7	1 250.6	615	0.185 0

4 结束语

时间序列分割是时间序列数据挖掘研究和应用中的主要任务之一[1].研究时间序列分割算法具有重要的理论意义和实际应用价值.时间序列分割主要有两个应用:系统模型变化检测和创建时间序列的高级数据表示,以便对时间序列进行索引、聚类和分类.在实时时间序列数据挖掘中,需要对实时得到的时间序列数据进行在线分割,以便于发现和预测时态模式.但是就我们所知,关于这方面的研究还有所欠缺.本文在对时间序列分割问题进行形式化描述的基础上,研究了评估时间序列的分割结果以及分割算法的评价指标,并提出了一种在线分割时间序列数据的递推算法(OLS).对比实验结果说明,OLS 算法能够有效地在线检测出数据挖掘应用中感兴趣的关键变化点,而且"过拟合"程度低.

References:

[1] Keogh E, Kasetty S. On the need for time series data mining benchmarks: A survey and empirical demonstration. In: Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2002. 102~111.

[2] Li AG, Qin Z, He SP. Extracting similar patterns in time series data. Journal of Xi'an Jiaotong University, 2002,36(12):1275~1278 (in Chinese with English abstract).

[3] Firoiu L. Segmenting time series with a hybrid neural networks—hidden markov model. In: Proc. of the 18th National Conf. on Artificial Intelligence (AAAI). Menlo Park: AAAI Press, 2002. 247~252.

[4] Keogh E, Chu S, Hart D, Pazzani M. An online algorithm for segmenting time series. In: Proc. of IEEE Int'l Conf. on Data Mining. Los Alamitos: IEEE Computer Society Press, 2001. 289~296.

[5] Guralnik V, Srivastava J. Event detection from time series data. In: Proc. of the 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 1999. 33~42.

[6] Qin Z, Li AG. Rubost optimization segment for time series data. Journal of Xi'an Jiaotong University, 2003,37(4):338~342 (in Chinese with English abstract).

[7] Li AG, He SP, Qin Z. Real-Time segmenting time series data. In: Zhou X, Zhang Y, Orlowska ME, eds. Proc. of the 5th Asia-Pacific Web Conf. 2003. LNCS 2642, Heidelberg: Springer-Verlag, 2003. 178~186.

[8] Li AG, Qin Z. Prediction time series using product unit neural networks with FIR synapses. Journal of Computer Research and Development, 2004,41(4):577~581 (in Chinese with English abstract).

[9] Kantz H, Schreiber T. Nonlinear Time Series Analysis. Cambridge: Cambridge University Press, 1997.

[10] Hawkins DM. Fitting multiple change-point models to data. Computational Statistic & Data Analysis, 2001,37(3):323~341.

附中文参考文献:

[2] 李爱国,覃征,贺升平.时间序列数据的相似模式抽取.西安交通大学学报,2002,36(12):1275~1278.

[6] 覃征,李爱国.时间序列数据的稳健最优分割.西安交通大学学报,2003,37(4):338~342.

[8] 李爱国,覃征.具有 FIR 突触的积单元神经网络预测时间序列.计算机研究与发展,2004,41(4):577~581.