一种无限长时间序列的分段线性拟合算法

闫秋艳,夏士雄
(中国矿业大学计算机科学与技术学院,江苏徐州 221116)

摘 要: 本文提出了一种无限长时间序列的分段线性拟合(Infinite Time Series_Piecewice Linear Fitting,简称 ITS_PLF)算法,该算法根据关键点保持时间段的统计特性,确定选择关键点的区间范围;若极值点的保持时间段不在区间范围,则根据包含极值点的连续三个时间数据之间的夹角与筛选角度之间的关系,判断该极值点成为关键点的可能性.实验表明,ITS_PLF 算法的执行不依赖于时间序列长度及领域知识,可以有效识别关键点,并可根据数据压缩率的变化实现自适应拟合.

关键词: 时间序列; 分段线性拟合; 压缩率
中图分类号: TP311.13 **文献标识码:** A **文章编号:** 0372-2112 (2010) 02-0443-06

An Piecewise Linear Fitting Algorithm for Infinite Time Series

YAN Qiu-yan, XIA Shi-xiong
(*The School of Computer Science and Technology, China University of Mining Technology, Xuzhou, Jiangsu* 221116, *China*)

Abstract: In order to resolving the problem of depending on the length of time series and domain knowledge of traditional PLF algorithm, we proposed a Piecewise Linear Fitting algorithm for Infinite Time Series (ITS_PLF). To determine the interval of Key Points selecting, the statistical attributes of maintaining time of these Key Points was considered. If the maintaining time of a Extreme Point beyond the selection interval, the relation between the threshold angle and the angle of three consecutive data points containing the Extreme Point was selected to determine whether the Extreme Point was a Key Point or not. The experimental results show that ITS_PLF algorithm does not depend on the length of time series and domain knowledge, can effectively identify the Key Point and adaptively fit the time series according to the changing of the data compression ratio.

Key words: time series; piecewise linear fitting; compression ratio

1 引言

时间序列的分段线性拟合(Piecewise Linear Fitting 简称 PLF)是时间序列的模式表示方法中研究最早和最多的方法之一.PLF 是指用 K 条首尾相邻的线段近似表示一条长度为 L 的时间序列[1].

在时间序列的 PLF 方法中,线段的数目决定了对原始序列的近似粒度,线段越多,线段的平均长度就越短,反映了时间序列的短期波动情况;线段越少,线段的平均长度就越长,反映了时间序列的中长期趋势[2],通常用数据的压缩率[3]来表征这个参数,这里的压缩率为从数据序列中删除的数据点所占的比例,如 80% 的压缩率即为选择 20% 个数据点并删除剩余的 80%.一种好的时间序列的模式表示方法必须能够准确识别噪音数据,并对噪音数据进行有效过滤,从而保证较高的数据压缩率.

关于时间序列分段线性拟合算法的研究自论文[1]以来,得到了广泛关注[2~6],这种简单直观的线性拟合表示方法采用首尾相邻的一系列线段近似表示时间序列,压缩原始序列,换取更小的存储和计算代价;保留时间序列主要形态的同时去除了细节干扰,更能反映时间序列的变化模式.通过对已有的 PLF 算法进行分析,发现这些 PLF 算法存在两个缺点:

(1)极值点选择的阈值依赖于相关的领域知识,不同的时间序列选择的阈值不同,因此算法不具有普遍的适用性;

(2)极值点选择的阈值依赖于时间序列的长度 L,当 L 为无穷大时,传统的 PLF 算法就不再适用.

为解决上述问题,本文提出了一种无限长时间序列的分段线性拟合算法(Infinite Time Series_Piecewice Linear Fitting,简称 ITS_PLF 算法),该算法根据已有关键点保持时间段的统计特性,判断极值点选择的区间范围;若某

收稿日期:2009-04-03;修回日期:2009-06-16
基金项目:国家自然科学基金(No.50674086);中国矿业大学青年科研基金(No.2008A041)

点的保持时间段不在区间范围,选择包含极值点的连续三个数据点,并根据三点构成的夹角与筛选角度之间的关系判断其成为关键点的可能性,从而解决了 PLF 算法依赖于时间序列长度 L 及领域知识的问题.实验表明,ITS_PLF 算法的执行不依赖于 L 及领域知识,可以有效识别关键点,并可根据数据压缩率的变化实现自适应拟合.

2 相关工作

本节对三种主要的时间序列分段线性拟合算法(IPSegmentation[4], FPSegmentation[6] 和 KPSegmentation[2])进行比较分析,说明现有 PLF 算法存在的问题和不足.

2.1 符号说明

定义本文使用的一些符号如下:

(1) $T=\langle(x_1,t_1),\cdots(x_i,t_i),\cdots(x_\infty,t_\infty)\cdots\rangle(0<i<\infty)$:采样时间间隔相同的时间序列,其中$(x_i,t_i)$表示采样时间 t_i 时刻的数值为 x_i;

(2) $X=\langle X_1(t_1,x_1),\cdots X_i(t_i,x_i),\cdots X_\infty(t_\infty,x_\infty)\cdots\rangle,0<i<\infty$:将 T 经过归一化处理后用直角坐标系表示的点序列,横坐标为时间轴,纵坐标为数值轴;

(3) $|X_i-X_j|$:表示时间序列中 $X_i(t_i,x_i)$和 $X_j(t_j,x_j)$在坐标平面内的欧氏距离;

(4) EP(*ExtremePoint*):极值点,T 的单调性在极值点发生改变;

(5) KP(*KeyPoint*):关键点,满足筛选条件的极值点;

(6) $KPS=<KP_1,\cdots,KP_n>$:关键点集

(7) α_0:筛选角度

2.2 相关算法比较

本文选取 Quarterly S&P 500 index, 1900 – 1996. Source: Makridakis, Wheelwright and Hyndman (1998)[7] 的前 100 条数据,对三种算法的拟合效果进行说明:

(1)极值点拟合法(IPSegmentation).该算法利用序列数据的单调变化属性识别极值点 EP,通过依次连接 EP 点实现序列的分段线性拟合.这种拟合算法尽管操作简单,运算效率高,较好地保留了原始时间序列的变化模式,但不能有效地去除噪音,过多地保留了细节变化,降低了压缩率.

(2)特征点拟合法(FPSegmentation).可以看作是极值点拟合法的改进算法.其实现思路为:选择原始序列中对序列形态影响最大的点作为特征点(Feature Point),通过依次连接这些特征点实现序列的线段化.特征点需要同时满足以下条件:①该数据点必须是序列的极值点;②该极值点保持极值的时间段(即该点与前后极值点的时间段)与该序列长度的比值必须大于某个阈值 C.FPSegmentation 的优点是:通过阈值 C 对转折点变化幅度的控制,可以较好地过滤变化短暂的噪音数据;缺点是:由于限定了极值点的变化幅度,对于变化时长小于 C 的转折点则无法有效识别,如图 1(选取 C = 0.04,X16 和 X32 之间的点)因为保持极值的时间段与 L 的比值小于 0.04,则这些数据被认为是噪音数据而删除;但同时,对于短暂变化的尖峰数据,则有可能被认为是噪音数据而被忽略,比较图 1 和图 2,X60 点保持极值的时间段与 L 的比值 $=\frac{3}{100}=0.03$,在 $C=0.03$ 时为一特征点,但在 $C=0.04$ 时该点被认为是噪音数据被忽略.从分析中可知,阈值 C 是特征点判断的影响因子,其取值和领域知识、序列长度以及用户关注角度有关,因此不同的 C 值会得到不同的拟合结果,直接影响拟合的质量;同时,当时间序列的长度 L 为无穷大时,C 为无穷小,则 FPSegmentation 算法不再适用.

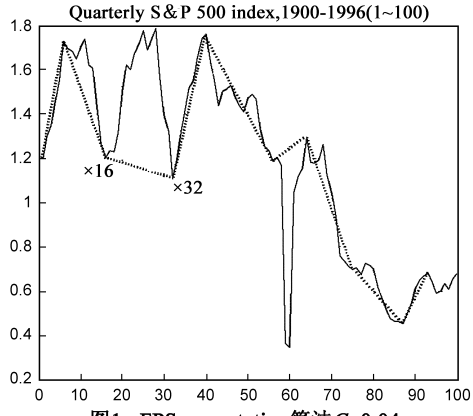

图1 FPSegmentation算法C=0.04

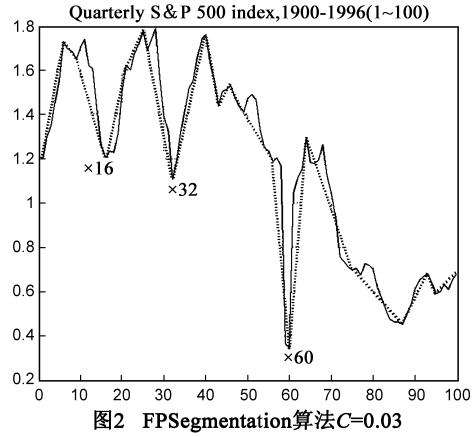

图2 FPSegmentation算法C=0.03

(3)关键点拟合法(*KPSegmentation*).该算法提出:包含极值点 X_i 的三个数据点构成的最小序列模式$\langle X_{i-1},X_i,X_{i+1}\rangle$中,如果三点连线形成的夹角越小,则中间点 X_i 为关键点的可能性越大.为便于进行在线运算,提出了基于三角形中线距离的关键转折点选择算法(IKP 算法),将计算三点夹角转换为计算距离 $\left|x_i-\frac{x_{i+1}-x_{i-1}}{2}\right|$,若 $\left|x_i-\frac{x_{i+1}-x_{i-1}}{2}\right|>\varepsilon$,($x_i$ 为极值

点纵坐标,$\varepsilon > 0$,为自定义的单调序列中线距离阈值). KPSegmentation 算法采用 FPSegmentation 算法和 IKP 算法保存数据序列中的特征点与突变序列中的关键点,然后利用特征点保持时间段阈值 C 过滤数据序列中的噪音干扰,利用关键点发现短暂变化的尖峰数据.其划分效果如图 3.为方便计算,取$\langle X_{i-1}, X_i, X_{i+1}\rangle$的筛选夹角为 45°, $C = 0.03$. KPSegmentation 算法在 ε 取的适当时,可以部分地发现时间序列中的关键点,如图 3 中点 X11 和 X68 保持极值的时间段与 L 的比值为 0.02,在图 2 中不是特征点,但运用 KPSegmentation 算法,其包含极值点的夹角小于筛选角度,因此成为关键点.存在的问题:因为该算法在判断关键点时,基于 FPSegmentation 的距离阈值 C 过滤噪音数据,也不可避免地遗传了 FPSegmentation 算法的所有缺点,且 ε 的取值依赖于领域知识.

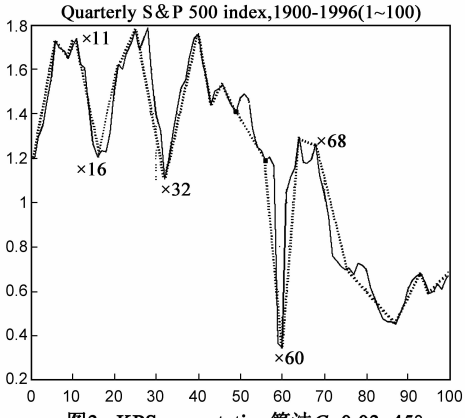

图3　KPSegmentation算法C=0.03, 45°

综上,尽管 IPSegmentaion 算法的序列拟合效果较好,但无法有效过滤噪音和细节干扰,压缩率较低;FPSegmentaion 算法在较高压缩率的情况下仍能较好地过滤噪音,保持原数据序列的形态,但该算法无法及时定位突变状态的起点和终点,也不能拟合原始序列中的尖峰状态;KPSegmentation 算法则在较高压缩率的情况下过滤了细节干扰,能较好地线性拟合原始序列,但该算法过分依赖两个经验阈值,对无线大时间序列的拟合设置了较大障碍.基于以上分析,本文提出一种无限长时间序列的分段线性拟合(Infinite Time Series Piecewice Linear Fitting,简称 ITS_PLF)算法,该算法可以不依赖于领域知识,并在时间序列长度为无穷大时准确识别噪音数据并保证较高的数据压缩率.

3　算法思想

基于以上分析可知,极值点 X_i 成为关键点(KP)的条件为:

条件 1　X_i 保持极值的时间段与该序列长度的比值必须大于某个阈值 C;

条件 2　若条件 1 不满足,则包含 X_i 的最小序列模式$\langle X_{i-1}, X_i, X_{i+1}\rangle$中三点连线形成的夹角小与筛选角度 α_0.

为了方便对极值点进行判断,得到如下定理 1~3:

定理 1　设时间序列 X 长度 L,令 Δt_i 表示 KP_i 点保持极值的时间段,则 Δt_i 满足 $N[\mu, \sigma^2]$的正态分布,其中 $\mu = \frac{1}{n}\sum_{i=1}^{n}\Delta t_i$, $\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\Delta t_i - \mu)^2}$, n 为关键点集 KPS 中元素的个数.

由中心极限定理,任意随机事件在样本数趋于无穷时,随机变量均服从 $N[\mu, \sigma^2]$的正态分布.时间序列本身是一个随机过程,极值点保持极值的时间段 Δt_i 的变化幅度是一个随机事件,因此 Δt_i 在序列长度趋于无穷时必满足正态分布.

推论 1　设时间序列的数据压缩率为 p,由定理 1 可得 p 的度量公式:

$$2\Phi(x) - 1 \leqslant 1 - p \tag{1}$$

证明　若时间序列的数据压缩率为 p,则被保留的关键点占时间序列数据总和的比例应 $\leqslant 1 - p$,即关键点的存在概率应 $\leqslant 1 - p$;

同时,由定理 1 可知,某个数据被保留,则其极值点保持时间段 Δt_i 满足 $N[\mu, \sigma^2]$的正态分布,即被保留的极值点的 Δt_i 以 μ 为中心对称分布,越靠近 μ,存在的概率越大,反之概率越小.因此,被保留的极值点的 Δt_i 应分布在$[\mu - x\sigma, \mu + x\sigma]$的范围内($x$ 代表偏离 μ 的程度),概率 $\leqslant 1 - p$,令 Y 表示该随机事件,则有:

$$P\{\mu - x\sigma < Y < \mu + x\sigma\} \leqslant 1 - p \tag{2}$$

因为 $\frac{Y - \mu}{\sigma} \sim N(0,1)$

则对式(2)变换后可得

$$\Phi\left(\frac{(\mu - x\sigma) - \mu}{\sigma}\right) - \Phi\left(\frac{(\mu - x\sigma) - \mu}{\sigma}\right) \leqslant 1 - p$$

即 $2\Phi(x) - 1 \leqslant 1 - p$ 得证.

例如,要得到大于 80% 的数据压缩率,则由式(1)得:

$$2\Phi(x) - 1 \leqslant 0.2$$
$$2\Phi(x) \leqslant 1.2$$
$$\Phi(x) \leqslant 0.6$$

查表得 $x = 0.25$,即若极值点 X_i 为关键点,则该点的 Δt_i 应分布在$[\mu - 0.25\sigma, \mu + 0.25\sigma)$范围内.

推论 1 得出了通过预先设定的数据压缩率,确定选择关键点区间范围的方法.

定理 2　若 X_i 不满足定理 1,则 $\frac{2\left|x_i - x_{i+1}\right|}{\left|x_i - x_{i+1}\right|^2 - 1} \leqslant \mathrm{tg}\alpha_0$.设 α_0 为筛选角度且 $\left|x_{i+1} - x_i\right| \geqslant \left|x_i - x_{i-1}\right|$),是 X_i 为关键点的充分条件.

证明 以图 4 为例,我们注意到在图 4 所示的角度变化中,由于时间序列$\langle X_{i-1}, X_i, X_{i+1}\rangle$是等间隔的时间点,因此 X_{i-1}, X_i, X_{i+1}三点的取值只能够在与时间轴垂直的三条直线上(图 4 中的 L_1, L_2, L_3),若 X_i 点固定,比较$|x_i - x_{i-1}|$和$|x_{i+1} - x_i|$的大小,取其中较大者(设较大者的端点为 X_{i+1});通过 X_{i+1}画一条水平线,与直线 L_1 交于点 $P_{i-1}(t_{i-1}, x_{i+1})$,与直线 L_2 交于 $P_i(t_i, x_{i+1})$,则一定存在$\angle X_{i-1}X_iX_{i+1} > \angle P_{i-1}X_iX_{i+1}$(证略),因此,只需要对$\angle P_{i-1}X_iX_{i+1}$进行考察,若$\angle P_{i-1}X_iX_{i+1} > \alpha_0$,则 X_i 点一定不是极值点.令$\angle P_{i-1}X_iX_{i+1} = 2\theta$,$\because \mathrm{tg}\theta = \frac{1}{|x_i - x_{i+1}|}$($|x_i - x_{i+1}|$为两点纵坐标的差值),$\mathrm{tg}2\theta = 1 - \frac{2\mathrm{tg}\theta}{1 - \mathrm{tg}^2\theta}$,因此当定理 1 不满足时,$X_i$ 为极值点必满足$\frac{2|x_i - x_{i+1}|}{|x_i - x_{i+1}|^2 - 1} < \mathrm{tg}\alpha_0$,反之不成立,充分性得证.

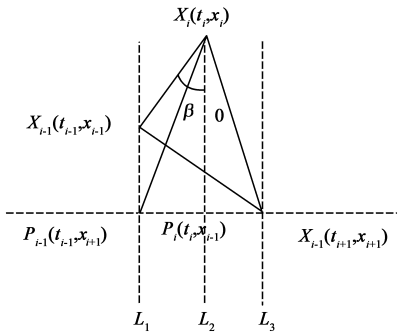

图4 序列(X_{i-1}, X_i, X_{i+1})间的夹角关系

定理 3 若 X_i 不满足定理 1,$\frac{2|x_i - x_{i+1}|}{|x_i - x_{i+1}|^2 - 1} \leqslant \mathrm{tg}\alpha_0$且$\frac{1}{|x_i - x_{i-1}|} \leqslant \frac{|x_{i+1} - x_i|\,\mathrm{tg}(\alpha_0) - 1}{|x_{i+1} - x_i| + \mathrm{tg}(\alpha_0)}$是 X_i 为关键点的充要条件.

证明 如图 4,当 $\theta \leqslant \frac{\alpha_0}{2}$,即$\frac{2|x_i - x_{i+1}|}{|x_i - x_{i+1}|^2 - 1} \leqslant \mathrm{tg}\alpha_0$时,还必须要求$\angle X_{i-1}X_iX_{i+1} \leqslant \alpha_0$,则$\angle X_{i-1}X_iX_{i+1} - \theta \leqslant \alpha_0 - \theta$,令 $\beta = \angle X_{i-1}X_iX_{i+1} - \theta$

$$\because \mathrm{tg}\beta = \frac{1}{|x_i - x_{i-1}|}$$

$$\mathrm{tg}(\alpha_0 - \theta) = \frac{\mathrm{tg}(\alpha_0) - \mathrm{tg}(\theta)}{1 + \mathrm{tg}(\alpha_0)\mathrm{tg}(\theta)} \tag{3}$$

$$\therefore \frac{1}{|x_i - x_{i-1}|} \leqslant \frac{\mathrm{tg}(\alpha_0) - tg(\theta)}{1 + \mathrm{tg}(\alpha_0)\,tg(\theta)} = \frac{\mathrm{tg}(\alpha_0) - \frac{1}{|x_{i+1} - x_i|}}{1 + \mathrm{tg}(\alpha_0)\frac{1}{|x_{i+1} - x_i|}}$$

$$= \frac{|x_{i+1} - x_i|\,\mathrm{tg}(\alpha_0) - 1}{|x_{i+1} - x_i| + \mathrm{tg}(\alpha_0)} \tag{4}$$

证毕

4 算法步骤

输入:时间序列 $T = \langle (x_1, t_1), (x_2, t_2), \cdots, (x_i, t_i), \cdots (x_\infty, t_\infty)\cdots\rangle$,$0 < i < \infty$,筛选夹角 α_0,预设数据压缩率 p.

输出:关键点集合 $KPS = \langle KP_1, \cdots, KP_n\rangle$.

step1 根据推论 1,由 p 计算系数 x

step2 初始化,$KP_1 = X_1(t_1, x_1)$,$\mu = 1$,$\sigma = 0$;

step3 从 X_1 开始判断时间序列的单调性,获得包含三个极值点 $X_{i-p}(t_{i-p}, x_{i-p})$ $X_i(t_i, x_i)$ $X_{i+q}(t_{i+q}, x_{i+q})$的局部时间序列 $X = \langle X_{i-p}, \cdots, X_{i-1}, X_i, X_{i+1}, \cdots, X_{i+q}\rangle$,待考察的极值点为 $X_1(t_i, x_i)$,包含该点的最短时间序列为$\langle X_{i-1}, X_i, X_{i+1}\rangle$;

step4 计算点 X_i 保持极值的时间段 Δt_i,若 $\Delta t_i \in [\mu - x\sigma, \mu + x\sigma]$,则 X_i 是关键点,将 X_i 点并入集合 KPS,对下一个极值点进行判断,否则继续;

step5 计算 $\max(|x_i - x_{i-1}|, |x_{i-1} - x_i|)$,设返回 $X_{i+1}(t_{i+1}, x_{i+1})$;

step6 若$\frac{2|x_i - x_{i+1}|}{|x_i - x_{i+1}|^2 - 1} > \mathrm{tg}\alpha_0$,则 X_i 一定不是关键点,返回 step4,对下一个极值点进行判断,否则继续;

step7 若$\frac{1}{|x_i - x_{i-1}|} > \frac{|x_{i+1} - x_i|\,\mathrm{tg}(\alpha_0) - 1}{|x_{i+1} - x_i| + \mathrm{tg}(\alpha_0)}$,则 X_i 一定不是关键点,返回 step4,对下一个极值点进行判断,否则继续;

step8 将 X_i 点并入集合 KPS,更新区间$[\mu - x\sigma, \mu + x\sigma]$;返回 step4,对下一个极值点进行判断.

5 实验结果及分析

5.1 实验环境及实验数据

本实验的运行环境为 CPU2.51hz,2G 内存,160G 硬盘,操作系统为 WindowsXP,开发工具为 Visual C++.

考虑到序列形态(振幅及角度)变化对数据压缩率的影响,实验选取形态变化较大和形态变化均匀的时间序列 1、2[7],预设筛选夹角 α_0 及数据压缩率 p,运用 ITS_PLF 算法进行分段线性拟合,实验结果表明,ITS_PLF 算法可以通过对比实际获得的数据压缩率与预设数据压缩率,自适应地调整筛选夹角 α_0 及关键点保持时间段的选择区间,从而使 ITS_PLF 算法不依赖于任何参数及领域知识,使 ITS_PLF 算法适用于无限长时间序列的线性拟合,具体拟合结果和分析见 5.2 节,实验数据[7]如下:

(1) Quarterly S&P 500 index, 1900 – 1996. Source: Makridakis, Wheelwright and Hyndman (1998).9 – 17b.DAT

(2)Simulated series $Z(T)=0.9.Z(T-1)+A(T)\sim IN(0,1)$. Source: O. D. Anderson (1976). ANDERSON5. DAT

5.2 实验结果

实验 1 时间序列 1 的形态变化较大,考察不同的筛选角度对拟合效果及数据压缩率的影响.为了图像的清晰,选取时间序列 1 前 100 条数据进行拟合,预设筛选角度 $\alpha_0=45°$,数据压缩率为 0.8,由推论 1 得 $x=0.25$,拟合效果如图 5.

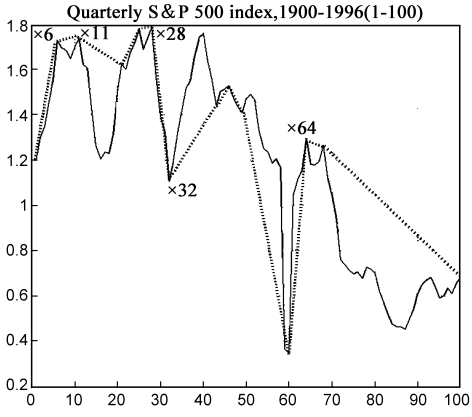

图5 时间序列1, ITS_PLF算法在 α_0=45°时的拟合效果

将图 5 与图 3 进行比较,二者均能较好反映时间序列形态变化的趋势.KPSegmentation 算法的数据压缩率 $1-\frac{18}{100}=0.82$,ITS_PLF 算法的数据压缩率为 $1-\frac{13}{100}=0.87$.区间$[\mu-0.25\sigma,\mu+0.25\sigma]$的变化情况如表 1,当筛选角度相同时,KPSegmentation 算法的阈值 $C=0.03$,ITS_PLF 算法的 $\Delta t_i\approx 3,\frac{3}{100}=0.03$,与 KPSegmentation 算法的阈值一致,但由于 ITS_PLF 算法是一个在线算法,因此二者只能具有相似的数据压缩率.

表 1 ITS_PLF 算法在 $\alpha_0=45°$时 Δt_i 区间的变化情况

坐标点	Δt_i	μ	σ	$[\mu-0.25\sigma,\mu+0.25\sigma]$
X_1	1	1	0	[1,1]
X_6	5	3	2.83	[2.79,4.21]≈[3,4]
X_{11}	2	2.67	2.08	[2.48,3.52]≈[3,3]
X_{28}	2	2.5	1.73	[2.07,2.93]≈[2,2]
X_{32}	4	2.8	1.64	[2.39,3.21]≈[3,3]
X_{64}	4	3	2.4	[2.4,3.6]≈[3,3]

从以上分析也可看出,$\alpha_0=45°$时 ITS_PLF 算法的压缩率为 0.87 与预期的压缩率≈0.8 相近,说明筛选角度选择的较为合适.

改变筛选角度 $\alpha_0=30°$,ITS_PLF 算法的拟合效果如图 6,图中可以看出,对于时间序列 1 随着筛选角度的减小,ITS_PLF 算法的数据压缩率明显增加,达到 0.92,但此时拟合的效果欠佳,如 X60 点,$\Delta t_{60}=3$,而此时 $\Delta t_i\approx 2$ 才会成为关键点,同时 X60 所在的夹角大于 30°,因此,该点没有被保留.分析原因,在 $\alpha_0=30°$时 ITS_PLF 算法的压缩率 0.92 与预期的压缩率≈0.8 相差较大,说明选择的筛选角度不合适.可见,在序列形态变化较大时,ITS_PLF 算法可将实际数据压缩率与预期数据压缩率相对照,判断筛选角度选取的是否合适,从而去除了算法对于初始参数 α_0 的依赖性.

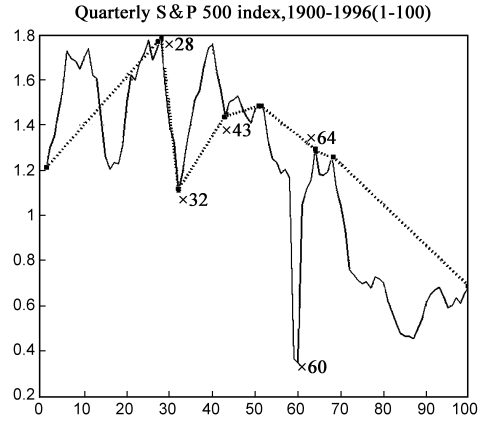

图6 时间序列1, ITS_PLF算法在α_0=30° 时的拟合效果

表 2 ITS_PLF 算法在 $\alpha_0=30°$时 Δt_i 区间的变化情况

坐标点	Δt_i	μ	σ	$[\mu-0.25\sigma,\mu+0.25\sigma]$
X_1	1	1	0	[1,1]
X_{28}	2	1.5	0.71	[1.32,1.68]≈[1,1]
X_{32}	4	2.3	1.52	[1.95,2.71]≈[2,2]
X_{43}	3	2.5	1.29	[2.18,2.82]≈[2,2]
X_{64}	4	2.8	1.98	[2.3,3.3]≈[3,3]

实验 2 时间序列 2 的形态变化均匀,考察此时不同的筛选角度对拟合效果及数据压缩率的影响,预设筛选角度 $\alpha_0=60°$,数据压缩率 0.8,前 50 条数据的拟合效果如图 7.在 X_2 点将筛选区间更新为[1,3],此后基本维持不变,数据压缩率为 $1-\frac{36}{50}=0.28$.将筛选角度 α_0 调整为 $\alpha_0=15°$,发现数据的压缩率维持在 0.28 附近基本保持不变,从图 7 可看出,该时间序列数据的振幅

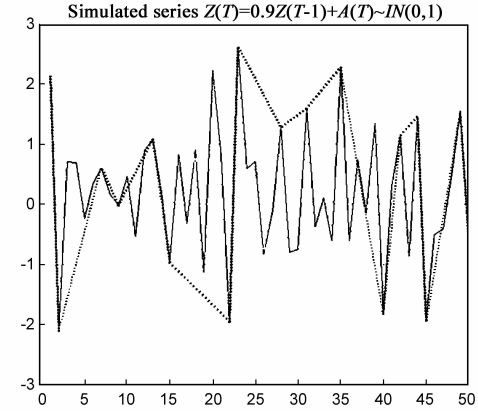

图7 时间序列2,前50条数据的拟合效果,压缩率0.28

和角度变化比较均匀,Δt_i 的分布区间为[1,3],因此,大部分数据都符合关键点的标准而被保留,这种结果也同时验证了定理1确定关键点选择区间的有效性.在这种情况下,如果需要继续对数据进行压缩,则可动态调整区间至[2,3],拟合效果如图8,数据压缩率为:$1-\frac{17}{50}=0.66$.可见,对于形态变化均匀的时间序列,筛选角度无法有效判断关键点,此时,ITS_PLF 算法可以随着时间数据的不断到来,根据以往实际获得的数据压缩率动态调整关键点选择区间,实现对未来数据压缩率的控制,从而实现了在无任何领域知识的条件下,对无限长时间序列的自适应拟合.

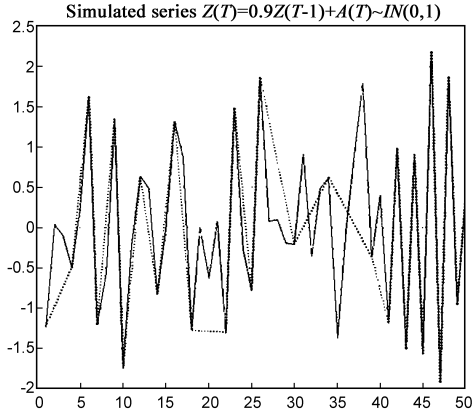

图8 时间序列2,50~100条数据的拟合效果,压缩率0.66

6 结束语

本文在对现有时间序列分段线性拟合(PLF)算法进行分析的基础上,总结现有算法在对无限长时间序列进行拟合时的不足和不适用性,提出了在时间序列无限长时判定某极值点为关键点的三条定理,并在此基础上提出了 ITS_PLF 算法,实验表明,在不同变化形态的时间序列上,ITS_PLF 算法的性能不依赖于序列长度和领域知识,可以有效识别关键点,并可根据数据压缩率的变化实现自适应拟合.

参考文献:

[1] T Pavlidis, S L Horowitz. Segmentation of plane curves[J]. IEEE Transactions on Computers, 1974, 23(8): 860 – 870.

[2] 杜奕.时间序列挖掘相关算法研究及应用[D].合肥:中国科学技术大学博士论文,2007.

[3] Kevin B Pratt, Eugene Fink. Search for patterns in compressed time series[J]. International Journal of Image and Graphics. 2002, 2(1): 89 – 106.

[4] Sanghyun Park, Sang-wook Kim, Wesley W. Chu. Segment – based approach for subsequence searches in sequence databases [A]. Proceedings of the 16th ACM Symposium on Applied Computing[C]. New York: ACM Press, 2000. 248 – 252.

[5] Sanghyun Park, Dongwon Lee, Wesley W Chu. Fast retrieval of similar subsequences in long sequence databases[A]. Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange[C]. Washington: IEEE. Computer Society, 1999. 60 – 67.

[6] 肖辉,胡运发.基于分段时间弯曲距离的时间序列挖掘[J].计算机研究与发展.2005,42(1):72 – 78.

[7] Hyndman, R J(n d). Time Series Data Library(DB/OL), http://www.robhyndman.info/TSDL, 2009-5.

作者简介:

闫秋艳 女,博士研究生,主要研究方向为数据流技术,时间序列处理,数据挖掘.
E-mail: yanqy@cumt.edu.cn

夏士雄 男,博士生导师,教授,主要研究方向为数字化矿山,智能信息处理.