基于 EEMD 的固定分段数分段线性表示方法

刘学彬[1],梁智飞[2],朱卫平[2],祝　凯[1*]
(1. 青岛理工大学 信息与控制工程学院,山东 青岛 266000;
2. 中石油煤层气有限责任公司,北京 102200)

摘　要:针对采用单一启发式规则的分段线性表示方法存在局部最优化和无法准确预计分段数目的问题,提出了基于集合经验模态分解(EEMD)的固定分段数分段线性表示方法。该方法通过将集合经验模态分解和重构思想引入分段线性表示方法研究中,同时将自底向上算法的拟合误差阈值改进为分段数阈值来解决上述两个问题。首先,通过模态重构思想过滤掉细节信息,提取到全局性分段点;然后,根据各初始分段子序列的波动程度,确定子序列段内分段点数量分布;最后,采用基于分段数阈值的自底向上方法将子序列合并到要求的分段数。该方法不仅继承了自底向上方法拟合误差小的优点,同时克服了局部最优化以及不能预计分段数的缺点。通过仿真实验证明了该方法克服了局部性的缺点,并有效减弱了噪声的干扰。相比现有方法,在压缩率相同的情况下,该方法的拟合误差更小。最终,在压裂施工时序数据趋势提取的应用中也验证了其有效性。

关键词:时间序列;分段线性表示;集合经验模态分解;模态重构;符号化;自底向上
中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1673-629X(2023)11-0202-07
doi:10.3969/j.issn.1673-629X.2023.11.030

Piecewise Linear Representation Algorithm of Fixed Section Number Based on EEMD

LIU Xue-bin[1],LIANG Zhi-fei[2],ZHU Wei-ping[2],ZHU Kai[1*]
(1. School of Information and Control Engineering,Qingdao University of Technology,Qingdao 266000,China;
2. Petrochina Coalbed Methane Company Limited,Beijing 102200,China)

Abstract:Aiming at the problems of local optimization and inability to accurately predict the number of segments in the piecewise linear representation method using a single heuristic rule,a piecewise linear representation method with a fixed number of segments based on Ensemble Empirical Mode Decomposition (EEMD) was proposed. This method introduces the idea of ensemble empirical mode decomposition and reconstruction into the research of piecewise linear representation method,and at the same time improves the fitting error threshold of the bottom-up algorithm to the threshold of piecewise number to solve the above two problems. Firstly,the detail information is filtered out by the idea of modal reconstruction,and the global segmentation point is extracted. Then,the distribution of the number of segmentation points in the subsequence is determined according to the fluctuation degree of each initial segmentation subsequence. Finally,a bottom-up method based on the number of segments threshold is used to merge the subsequences into the required number of segments. This method not only inherits the advantages of small fitting error of the bottom-up method,but also overcomes the shortcomings of local optimization and unpredictable number of segments. The simulation experiment proves that the proposed method overcomes the shortcoming of locality and effectively weakens the interference of noise. Compared with existing methods,the fitting error of the proposed method is smaller when the compression rate is the same. Finally,its effectiveness is also verified in the application of time series data trend extraction of fracturing construction.

Key words:time series;piecewise linear representation;ensemble empirical mode decomposition;mode reconstruction;symbolization;bottom-up

收稿日期:2023-01-08　　**修回日期**:2023-05-10
基金项目:山东省自然科学基金资助项目(ZR2019PEE013)
作者简介:刘学彬(1997-),男,硕士研究生,CCF 会员(O1349G),研究方向为时间序列数据挖掘;通讯作者:祝　凯(1988-),男,博士,副教授,研究方向为信号处理、机器学习、多模态数据融合。

0　引　言

由于时间序列是高维且存在大量噪音的,直接在原始序列上进行预测、模式发现和分类等挖掘任务的效率较低,同时也会影响挖掘结果的精度和可信度。因此,使用特征表示方法将时间序列从高维度转换到低维度,这种方法可以在降低时间序列复杂度的同时,保留时间序列的主要信息,为进一步深入研究时间序列奠定基础[1]。

目前国内外有不少学者致力于时间序列特征表示方法的研究,时间序列特征表示方法的主要代表有:基于域变换的表示方法(离散傅里叶变换[2]和离散小波变换[3];符号化表示方法,其中应用最广泛的是符号聚合近似方法[4-5];分段累计近似方法[6]和分段线性表示(Piecewise Linear Representation, PLR)[7]。其中 PLR 具有简单、直观的特点,能够有效保留原序列的形态信息以减少拟合误差,是一种应用广泛的时间序列特征表示方法。因此,该文着眼于分段线性表示方法的研究和改进。

目前,PLR 的研究主要集中于解决分段数和分段点的选择问题上。为了解决这些问题,时序的分段表示方法可以分为以下几种:(1)限制分段数:主要代表是分段累计近似方法,但该方法没有考虑实际序列形态,不能很好地保留原始序列特征;(2)限制分段误差:主要代表性算法有自顶向下[8]、自底向上[9]、滑动窗口[10]。限制分段误差方法对一些状态变化的拐点不敏感,不能保证每一分段只具有一种基本趋势。针对上述问题,近年来不少学者提出了一些改进方法。例如,尚福华[11]和廖俊[12]提出基于趋势转折点的分段线性表示方法;陈帅飞[13]提出基于关键点的分段线性表示方法;刘意杨[14]提出基于转折点和趋势段的分段线性表示方法等。但是,这些方法使用单一的启发式规则,难以适用于数据分布复杂的时间序列,进而导致算法出现局部最优化问题,而且不能灵活控制压缩率,不能适应后期要求分段数一定的应用[15]。

针对上述方法存在局部最优化和不能预计分段数的问题,提出了基于 EEMD 的固定分段数分段线性表示方法。首先,通过模态重构思想过滤掉细节信息,提取到全局性分段点;然后,根据各初始分段子序列的波动程度,确定子序列段内分段点数量分布;最后,采用基于分段数阈值的自底向上方法将子序列合并到要求的分段数。

1　分段线性表示相关概念及问题描述

1.1　分段线性表示相关概念

定义 1(拟合误差):时间序列 $X=\{x_1,x_2,\cdots,x_n\}$ 经过分段线性表示方法得到时间序列的分段线性表示为 X_{PLR}。X_{PLR} 经过线性插值得到的时间序列记为 $X^C=\{x_1^c,x_2^c,\cdots,x_n^c\}$,分段线性表示与原始时间序列之间的拟合误差为:

$$E=\sqrt{\sum_{i=1}^{n}(x_i-x_i^c)^2} \tag{1}$$

定义 2(压缩率):原始时间序列 $X=\{x_1,x_2,\cdots,x_n\}$,给定其分段线性表示 $X_{\mathrm{PLR}}=\{x_1^p,x_2^p,\cdots,x_d^p\}$,其中 $x_1=x_1^p,x_n=x_d^p$。则时间序列分段线性表示后的压缩率 Cr 可以表示为:

$$\mathrm{Cr}=(1-\frac{d}{n})\times 100\% \tag{2}$$

定义 3(重要点序列):给定时间序列 X,定义 X 的第 q 个重要点为 $x_q^Z=x_{p_q}$,其中 $p_q\in\{1,2,\cdots,n\}$ 表示第 q 个重要点在时间序列 X 中的位置,x_{p_q} 满足以下关系[16]:

$$\{x_{p_q-1}\leqslant x_{p_q}\}\cap\{x_{p_q+1}<x_{p_q}\}\cup\{x_{p_q-1}<x_{p_q}\}\cap\{x_{p_q+1}\leqslant x_{p_q}\}$$

或

$$\{x_{p_q-1}\geqslant x_{p_q}\}\cap\{x_{p_q+1}>x_{p_q}\}\cup\{x_{p_q-1}>x_{p_q}\}\cap\{x_{p_q+1}\geqslant x_{p_q}\} \tag{3}$$

此外,规定一个有限长度的时间序列起点和终点为重要点。由式(3)得到 m 个重要点,则重要点序列表示为:

$$X^Z=\{x_q^z\}_{q=1}^m \tag{4}$$

1.2　问题描述

传统的算法采用单一的启发式规则提取局部特征点,当原始时间序列波动频率较为剧烈且集中时,容易出现多个点的斜率变化近似。时间序列如图 1 所示。

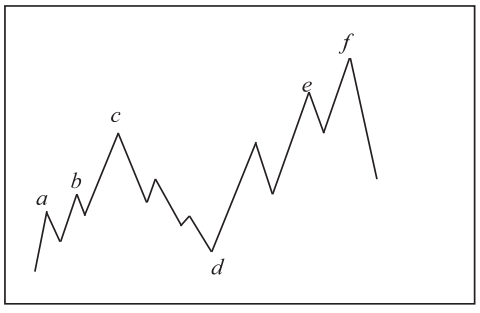


图 1　斜率波动频繁剧烈的情况

图 1 中序列点 a,b,c,d,e,f 点斜率变化近似,当通过调节斜率变化阈值 d 使得达到要求的压缩率时,会出现临界阈值,如下:

$$|\tan a_{\mathrm{L}}-\tan a_{\mathrm{R}}|>d,\ |\tan b_{\mathrm{L}}-\tan b_{\mathrm{R}}|>d$$

$$|\tan c_{\mathrm{L}}-\tan c_{\mathrm{R}}|<d,\ |\tan d_{\mathrm{L}}-\tan d_{\mathrm{R}}|<d$$

$$|\tan e_{\mathrm{L}}-\tan e_{\mathrm{R}}|<d,\ |\tan f_{\mathrm{L}}-\tan f_{\mathrm{R}}|>d$$

其中,下标 L 表示左,R 表示右。由上述公式和图 1 知,c,d 两点作为反映序列整体趋势的特征点因斜率

变化小而“漏提取”,即分段方法的结果遗漏掉能够反映整体特征的数据点;由此可认为, b,e 两点为“过提取”,即分段方法的结果提取到不能反映整体特征的数据点,导致算法陷入局部最优化。

2 集合经验模态分解和改进的自底向上分段

2.1 集合经验模态分解

Huang[17] 提出了经验模态分解(Empirical Mode Decomposition,EMD)。该方法的核心思想是将复杂的信号分解为有限个频率从高到低的本征模态函数(Intrinsic Mode Functions, IMF),对于某时间序列 $\{x(t)\}$ 经验模态分解的具体步骤如下:

(1)求出 $\{x(t)\}$ 中所有的极值。

(2)采用 3 次样条函数进行插值拟合上包络线 $b_{\max}(t)$ 和下包络线 $b_{\min}(t)$。

(3)计算上下包络线平均值 $m(t)$:

$$m(t) = [b_{\max}(t) + b_{\min}(t)] \tag{5}$$

(4)从时间序列中提取均值并将 $x(t)$ 和 $m(t)$ 的差定义为:

$$d(t) = x(t) - m(t) \tag{6}$$

(5)检查 $d(t)$ 的属性:如果满足 IMF 分量条件,则将 $d(t)$ 表示为第 k 个 IMF,并将 $x(t)$ 替换为残差 $r_1(t) = x(t) - d(t)$。第 k 个 IMF 分量通常表示为 $c_k(t)$;如果不满足,则将 $x(t)$ 替换为 $d(t)$。

(6)重复步骤(1)~(5)直到残差为单调函数为止。

原始时间序列可以表示为若干个 IMF 和一个残差的线性组合:

$$x(t) = \sum_{k=1}^{N} c_k(t) + r(t) \tag{7}$$

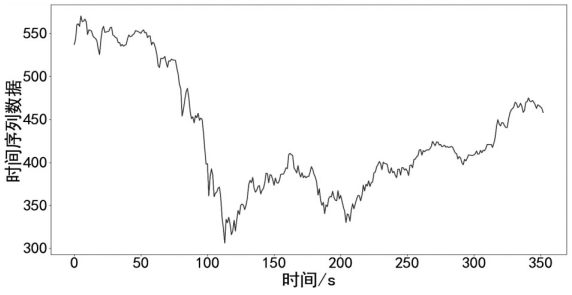
其中, $x(t)$ 表示 1 维信号; $c_k(t)$ 表示第 k 个 IMF 分量; $r(t)$ 表示残余。

当时间序列的时间尺度呈现跳跃性时,采用 EMD 对其进行分解,将会产生一个 IMF 分量包含不同时间尺度特征成分的情况,这种现象被称为模态混叠[18],它使得 EMD 得到的分解结果的可靠性和可解性受到影响。Wu[18] 提出了集合经验模态分解(Ensemble Empirical Mode Decomposition,EEMD)解决这一问题。基本思想是将不同白噪声多次加入原始时间序列以消除模态混叠现象。

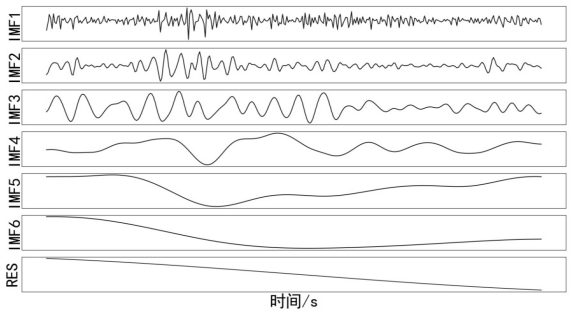
如图 2(a)所示,对 1 组示例时间序列进行 EEMD 分解,得到了 6 个 IMF 分量和 1 个 RES 残余,如图 2(b)所示。

2.2 IMF 重构

Zhang 等人[19] 采用 EEMD 技术来分析石油价格



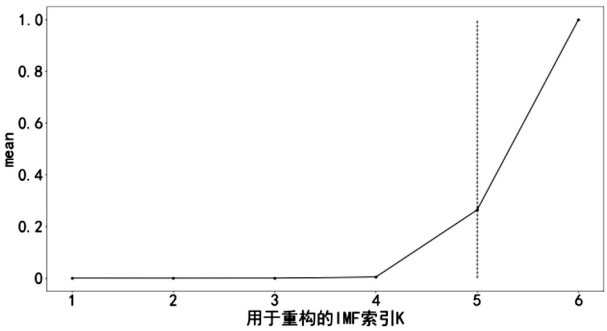
(a)示例时间序列



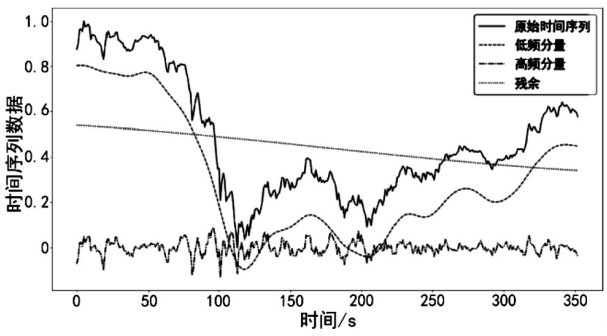
(b)EEMD 分解结果

图 2 示例时间序列集合经验模态分解

变化。他们发现,经本征模态函数重构后的序列可以很好地反映序列的关键转折点和整体趋势变化。基于这项研究,该文使用 EEMD 技术对时间序列进行分解,并将分解得到的 IMF 分为高频部分、低频部分和残余。前两个成分能够揭示时间序列所蕴含的物理意义,并发现时序的一些新特征。对 EEMD 分解得到的 N 个 IMF,求出每个 IMF 的平均值,得到用于分解高频和低频分量的 K 函数。以图 2(a)的时间序列为例,构建的 K 函数及高、低频和残余分量如图 3 所示。



(a)分解高频和低频分量的 K 函数



(b)原始时序和 3 个分量

图 3 K 函数及对应的 3 个分量

由图 3(a)知,在 IMF5 处,平均值开始偏离零点,因此使用 IMF1 ~ IMF4 的部分重构表示高频分量,使用 IMF5 和 IMF6 的部分重构表示低频分量,残余单独处理。图 3(b)显示了原始时间序列和 3 个分量。残余反映时间序列长期缓慢变化;低频分量的每次急剧上升或下降可能对应 1 个物理事件或是某种程度上的噪声表征;而高频分量通过去除大量的小幅波动使得可以反映时间序列的整体变化趋势。下面给出模态重构序列的定义。

定义 4(模态重构序列):对于某时间序列 X,对 X 进行 EEMD 分解得到 N 个 IMF,定义参与重构的起始 IMF 索引为 s,终止索引为 e,重构序列 X^{R} 表示为:

$$X^{\mathrm{R}} = \{x_i^{\mathrm{r}}\}_{i=1}^{n} = \sum_{k=s}^{e} \widetilde{\mathrm{IMF}_k} \tag{8}$$

在高频分量基础上,提取全局特征点,实现时间序列的初始分段。

定义 5(全局特征点序列):对于某时间序列 X,得到重构序列 $X^{\mathrm{R}} = \{x_i^{\mathrm{R}}\}_{i=1}^{n}$,对 X^{R} 使用式(3)得到 M 个全局特征点,则全局特征点序列可以表示为:

$$X^{\mathrm{RZ}} = \{x_w^{\mathrm{rz}}\}_{w=1}^{M} \tag{9}$$

根据上式,对图 3(b)中的高频分量提取全局特征点,实现时间序列的初始分段。

由图 4 知,原始时间序列被全局特征点分割为 12 段子序列,每段子序列都保持整体上升、下降、保持三种基本趋势,有效去除大量小幅波动,反映时间序列整体变化趋势。

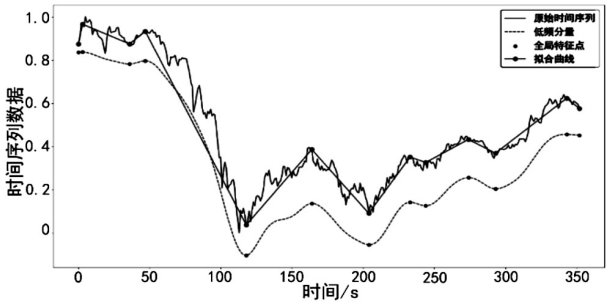


图 4　全局特征点初始分段

2.3　时间序列符号化

假设在序列中需要查找 N 个分段点,上节已提取了 M 个全局特征点,并将原时间序列分成了 $M+1$ 个初始段。接下来,采用廖俊[12]提到的时间序列点间的模式变化提取剩下的 $N-M$ 个分段点,如图 5 所示。

为了反映时间序列内的模式变化,将所有时间序列数据点符号化[20]。在时间序列 X 中,给定某一序列点 x_j,然后分别用前一点 x_i 和后一点 x_k 与该点做差分,即 $x_k - x_j = Q$ 和 $x_j - x_i = P$。具体步骤如下[1]:

(1)当符合模式 7 和 8($P * Q < 0$)时,相邻的左右点位于 x_j 同一端,如果符合条件:$\left|\frac{P}{j-i}\right| > \delta$ 或 $\left|\frac{Q}{k-j}\right| > \delta$,用“1”表示该序列点。

(2)当符合模式 1 到模式 6($P * Q \geqslant 0$)时,相邻的左右点位于 x_j 不同端,如果符合条件:$\left|\frac{Q}{k-j} - \frac{P}{j-i}\right| > \delta$,用“1”表示该数据点。

(3)将不符合上述条件的点用“0”表示。

(4)遍历整个序列,得到符号化序列。

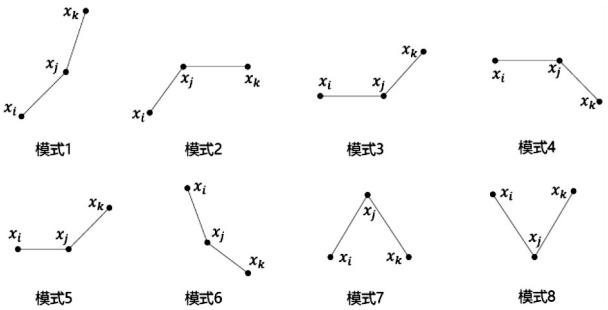


图 5　时间序列 3 点之间的模式变化

其中 δ 为自定义阈值,将所有符号化的子序列分别求和,存入 H_i 中,得到长度为 $M+1$ 的序列:$H = \{H_1, H_2, \cdots, H_{M+1}\}$,通过以下公式:

$$N_i = (N - M) * \frac{H_i}{\sum_{i=1}^{M+1} H_i} \tag{10}$$

得到 $M+1$ 个子序列内分段点的分布数量:

$$C = \{C_1, C_2, \cdots, C_{M+1}\} \tag{11}$$

2.4　固定分段数的自底向上分段

经典的自底向上方法由 Keogh 等人[7]提出,该方法的基本思想是通过循环地合并误差最小的相邻分段,直到所有的拟合误差均不小于分段阈值为止。该算法存在偶数限制的不足,为了解决该问题,孙焕良在其[21]研究中提出了优化的 PLR_BU 算法,但是仍无法准确地预测时间序列的分段数。

针对这一缺陷,该文在优化的 PLR_BU 算法基础上进行了改进,提出了固定的 PLR_BU 算法。该算法的基本思想:首先将长度为 n 的时间序列依次相连前后两点,然后给定分段数阈值,循环地执行下述过程:(1)计算相邻的分段合并后的拟合误差;(2)查找拟合误差最小的相邻分段 $\{x_i, x_j, x_k\}$,移除此相邻分段的中心点 x_j,序列长度减 1;(3)计算新生成的段与前后分段的拟合误差。重复上述过程,直到合并到满足设定的分段数为止。固定分段数的 PLR_BU 算法伪代码如表 1 所示。

2.5　时间序列分段线性表示方法

该文提出基于 EEMD 的固定分段数分段线性表示方法,具体算法步骤如下:

给定时间序列 $X = \{x_1, x_2, \cdots, x_n\}$,斜率变化阈值 δ,分段数 N。

(1)获取重要点序列。根据式(3),筛选出时间序列中的所有局部极值点,得到重要点序列 $X^Z=\{x_1^Z,x_2^Z,\cdots,x_m^Z\}$。

(2)获取模态重构序列和全局特征点序列。根据集合经验模态分解对时间序列分解,利用图 3 的 K 函数确定分解高频和低频分量的 IMF 起始索引 s 和结束索引 e,根据式(8)得到重构序列 $X^Z=\{x_1^R,x_2^R,\cdots,x_m^R\}$,然后根据式(3)筛选出高频分量中的所有局部极值点,得到全局特征点序列 $X^{RZ}=\{x_1^{RZ},x_2^{RZ},\cdots,x_M^{RZ}\}$,完成时间序列初始分段。

(3)时间序列符号化和确定初始分段段内分段点分布。根据斜率变化将时间序列转换成由“0”和“1”组成的符号化序列。计算符号化后各初始分段内数据和,得到 $H=\{H_1,H_2,\cdots,H_{M+1}\}$,根据式(10)和(11),得到最终子序列段内分段点分布数量序列 $C=\{C_1,C_2,\cdots,C_{M+1}\}$。

(4)固定分段数的 PLR_BU 算法确定最终分段点。根据改进的 PLR_BU 算法对子序列继续分段,直到分段数为 N,最终分段点序列:$X=\{x_1,x_{i_1},\cdots,x_{i_{N-2}},x_n\}$。

表 1 固定分段数的 PLR_BU 算法

输入:时间序列 X,分段数 N
输出:分段结果序列 Seg_TS

```
for  i = 1 to length( X ) //初始化
     Seg_TS = concat( Seg_TS, create_segment( X[i:i + 1] ));
for  j = 1 to length( Seg_TS )
     merge_cost (j) = calculate_error( [ merge( Seg_TS (i), Seg_
TS( i +1 ) ) ] );
     while length( Seg_TS ) > N   //不断融合直到预先设定的
分段数
     Index = min ( merge_cost ); //找到拟合误差最小的相邻
分段
Seg_TS ( index ) = merge ( Seg_TS ( index ), Seg_TS ( index +
1 ) );//更新分段序列
     delete( Seg_TS( index+1 ) );//删除相邻分段中的中间点
     merge_cost ( index - 1 ) = calculate_error ( merge ( Seg_TS
( index-1 ), Seg_TS( index ) ) );//更新新的合并分段与前一分
段拟合误差
     merge_cost ( index ) = calculate_error ( merge ( Seg_TS
( index ), Seg_TS( index+1 ) ) );//更新新的合并分段与后一分
段拟合误差
```

参数说明:在文中方法中,斜率变化阈值 δ 是主要的参数。设置阈值 δ 的目的是按照斜率变化过滤数据点,δ 值过小时,会将斜率变化相对较小的数据点也转换为“1”,导致相对平缓的序列段分段点的分布数量也较多;δ 值过大时,会将斜率变化相对较大的数据点转换为“1”,导致只有斜率变化相对较大的序列段分段点的分布数量才会较多。由于 P 和 Q 表示相邻点的差分,所以 δ 取值范围为 $0<\delta<\max(\Delta x)$,$\max(\Delta x)$ 表示序列相邻点最大差分。

3 实验与分析

3.1 实验对比方法

该文选择以下 3 种时间序列分段线性表示方法作为比较对象。

(1)PAA[6]。

(2)PLR_ITTP[12]。

(3)PLR_TRIP[14]。

3.2 仿真数据验证

$$X=\begin{cases}40 & t\in[1,100]\cup[501,600]\\ 40-(t-100)*2 & t\in[101,140]\\ -40-(t-141)/2 & t\in[141,190]\\ -65 & t\in[191,240]\cup[361,410]\\ -65+(t-191)*3 & t\in[241,250]\\ -35+(t-251)/2 & t\in[251,270]\\ -25 & t\in[271,330]\\ -25-(t-331)/2 & t\in[331,350]\\ -35-(t-351)*3 & t\in[351,360]\\ -65+(t-441)/2 & t\in[411,460]\\ -40+(t-461)*2 & t\in[461,500]\end{cases}\tag{12}$$

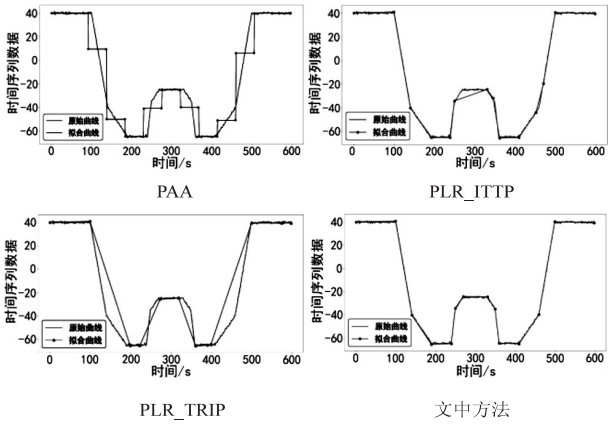
式中,t 为整数,共 600 个数据。选择该仿真序列作实验,是由于这个序列的重要点比较清晰,且没有噪声的干扰,因此,重要点更容易被发现。对序列 X 加上均值($\mu=0$)、方差(σ 设为 0.5 ~ 3)、步长为 0.5 的随机误差,对比在不同噪声的情况下,各个分段方法的抗噪声能力,检验是否可以准确提取全局特征点的有效性,对于对比分析时间序列消除局部最优化问题具有参考意义。

根据仿真数据的实际趋势,将序列分为 13 段,将拟合误差作为评价标准,显然,提取的分段点越接近原始序列趋势分段点,拟合误差越小。实验结果如图 6 和表 2 所示。

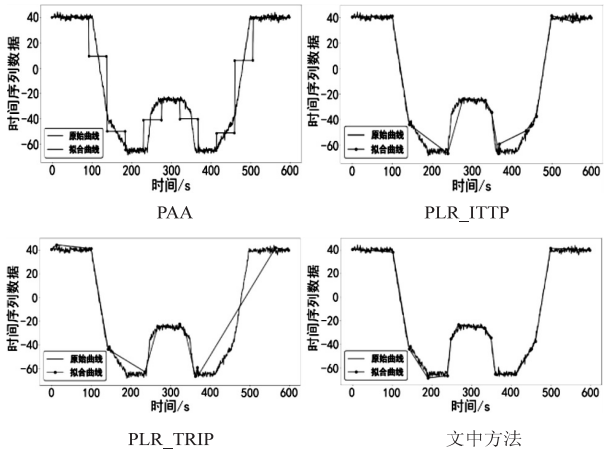
表 2 不同噪声下不同 PLR 的拟合误差

σ	PAA	PLR_ITTP	PLR_TRIP	文中方法
0.5	266.09	43.81	326.99	17.42
1.0	268.69	85.16	197.48	28.54
1.5	267.62	127.34	272.28	52.01
2.0	267.89	124.64	321.55	67.49
2.5	271.83	301.35	230.34	77.98
3.0	277.01	363.93	226.88	84.86

图 6 是不同噪声情况下,不同分段方法的分段拟合结果。



(a) $\mu=0,\ \sigma=0.5$ 时不同 PLR 结果



(b) $\mu=0,\sigma=1.5$ 时不同 PLR 结果

图 6　不同噪声情况下的不同 PLR 结果

由图 6 可知,随着 σ 的逐渐增大,分段算法对能够反映整体趋势的分段点的识别越来越困难,而文中方法相比于其他 3 种方法,对该序列中能够反映整体趋势的分段点的识别较为准确,尤其在高噪声情况下更为明显,而其他 3 种方法均提取了错误的分段点。

由表 2 知,文中方法虽受噪声干扰,但总的来说,抗噪声干扰的能力比其他 3 种方法有所加强,可以非常准确地提取反映整体趋势的分段点,而其他方法则极易受到噪声的干扰,导致陷入局部最优状态。

由图 6 和表 2 可知,尽管噪声的增加对 PAA 的拟合误差影响较小,但其整体的拟合误差相对较大,这是由于 PAA 采用等长分段,分段点的选取不会受到噪声的影响。而 PLR_ITTP、PLR_TRIP 都通过某种抗噪机制削弱了噪声的干扰,使得拟合误差相对较小。PLR_ITTP 在噪声较低时,拟合误差较小,但在噪声较高时,拟合误差随之变大,这是因为 PLR_ITTP 只重视时间序列的局部特征,而忽略了全局意义下的时间转折点,这将导致在高噪声的时候,算法会错误地提取关键点。PLR_TRIP 在不同噪声下拟合误差都较大,并且出现了震荡上升的情况,由于 PLR_TRIP 的角度阈值和趋势段阈值的组合选取是复杂的,不同的阈值组合会造成拟合误差的大幅变化,同时 PLR_TRIP 提出趋势段的概念来削弱噪声的干扰也是只关注局部信息。而文中方法先通过模态重构方法得到全局分段点,使得文中方法有效克服上述方法存在的局部最优化缺陷,之后使用自底向上方法进行融合,保留了基于分段误差的分段算法拟合误差较小的优点。

3.3　工业实例应用

压裂施工过程中,通过记录不同时间段的施工压力、泵注排量和加砂体积分数获得压裂施工曲线,有效利用压裂施工曲线并提取有效信息,不仅能够对储层及裂缝参数再认识,而且对于指导压裂施工以及调整压裂设计方案、提高压裂技术水平和施工效果有重要的借鉴作用。因此,对压裂施工曲线的挖掘和分析有着重要的工程意义和应用价值[22]。但压裂施工曲线是一种高维数据,为了方便后续存储和挖掘需要对其进行压缩表示。该文选用某区块的压裂施工数据,时间间隔为 1 s,共 2 500 个数据值。

为了比较文中算法与其他算法的优劣,并考虑到实际压裂施工曲线的高维性特点,将考察所有方法在压缩率分别为 90%、92%、94%、96%、98% 时的拟合误差,实验结果如图 7 所示。

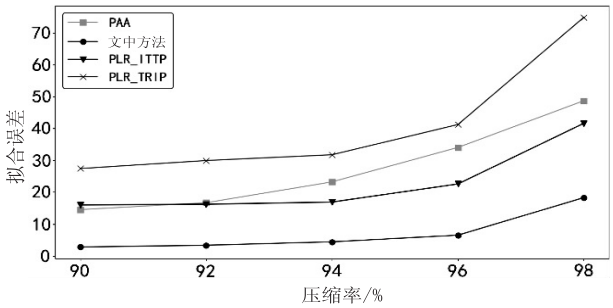
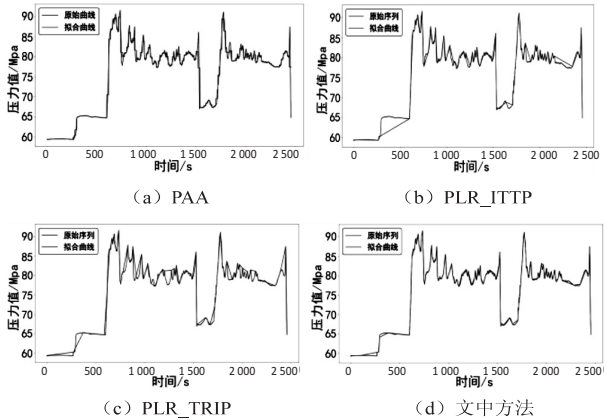


图 7　压裂施工曲线不同压缩率下的拟合误差

由图 7 可知,文中方法在不同压缩率下的拟合误差都是最小的,优于其他 3 种方法。



(a) PAA　(b) PLR_ITTP　(c) PLR_TRIP　(d) 文中方法

图 8　96% 的压缩率下不同方法分段拟合效果对比

由图 8 可知,对压裂施工时间序列这种高噪声且分布复杂的序列进行趋势提取时,PLR_ITTP、PLR_TRIP 因存在局部最优化的问题而导致“漏提取”“过提取”现象较为严重,丢失了反映整体趋势的重要点;

而文中算法能够有效地去除时间序列中的噪声,准确提取反映时间序列整体趋势的分段点。

4 结束语

针对现有方法的不足,该文提出一种基于 EEMD 的固定分段数表示方法。仿真实验结果表明:该方法的拟合误差分别比 PAA、PLR_ITTP、PLR_TRIP 平均减小了 80%、59%、78%,其有效地解决了现有分段方法存在的问题,极大地削弱了噪声干扰,从而能够准确地找到反映整体趋势的分段点。最后将该方法应用于压裂施工数据,其拟合误差分别比 PAA、PLR_ITTP、PLR_TRIP 减小了 86%、84%、89%,再次证明了所提方法对趋势提取的有效性和准确性。

参考文献:

[1] 林 意,孔斌强. 基于多尺度的时间序列固定分段数线性表示[J]. 计算机工程与应用,2016,52(21):81-87.

[2] AGRAWAL R,FALOUTSOS C,SWAMI A. Efficient similarity search in sequence databases[C]//Foundations of data organization and algorithms: 4th international conference. Washington DC:IEEE,1993:69-84.

[3] STRUZIK Z R,SIEBES A. Wavelet transform in similarity paradigm[C]//Research and development in knowledge discovery and data mining: second Pacific-Asia conference. Melbourne:Springer,2005:295-309.

[4] LIN J,KEOGH E,LONARDI S,et al. A symbolic representation of time series,with implications for streaming algorithms[C]//Proceedings of the 8th ACM SIGMOD workshop on research issues in data mining and knowledge discovery. San Diego:ACM,2003:2-11.

[5] LIN J,KEOGH E,WEI L,et al. Experiencing SAX:a novel symbolic representation of time series[J]. Data Mining and Knowledge Discovery,2007,15:107-144.

[6] KEOGH E,CHAKRABARTI K,PAZZANI M,et al. Dimensionality reduction for fast similarity search in large time series databases[J]. Knowledge and Information Systems,2001,3(3):263-286.

[7] KEOGH E,CHU S,HART D,et al. An online algorithm for segmenting time series[C]//Proceedings 2001 IEEE international conference on data mining. San Jose:IEEE,2001:289-296.

[8] KEOGH E J,PAZZANI M J. An enhanced representation of time series which allows fast and accurate classification,clustering and relevance feedback[C]//Proceedings of ACM international conference on knowledge discovery and data mining. New York:ACM,1998:239-243.

[9] PARK S,LEE D,CHU W W. Fast retrieval of similar subsequences in long sequence databases[C]//Proceedings 1999 workshop on knowledge and data engineering exchange. Piscataway:IEEE,1999:60-67.

[10] QU Y,WANG C,WANG X S. Supporting fast search in time series for movement patterns in multiple scales[C]//Proceedings of the seventh international conference on Information and knowledge management. New York: ACM,1998:251-258.

[11] 尚福华,孙达辰. 基于时间序列趋势转折点的分段线性表示[J]. 计算机应用研究,2010,27(6):2075-2077.

[12] 廖 俊,于 雷,罗 寰,等. 基于趋势转折点的时间序列分段线性表示[J]. 计算机工程与应用,2010,46(30):50-53.

[13] 陈帅飞,吕 鑫,戚荣志,等. 一种基于关键点的时间序列线性表示方法[J]. 计算机科学,2016,43(5):234-237.

[14] 刘意杨,李俊朋,白洪飞,等. 基于转折点和趋势段的时间序列趋势特征提取[J]. 计算机应用,2020,40(S1):92-97.

[15] 陈 然,戴 齐. 基于重要点的时间序列固定分段数分段算法[J]. 计算机技术与发展,2011,21(9):103-106.

[16] 周 黔,吴铁军. 基于重要点的时间序列趋势特征提取方法[J]. 浙江大学学报:工学版,2007,41(11):1782-1787.

[17] HUANG N E, SHEN Z,LONG S R,et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[J]. Proceedings of the Royal Society of London. Series A:Mathematical,Physical and Engineering Sciences,1998,454(1971):903-995.

[18] WU Z,HUANG N E. Ensemble empirical mode decomposition:a noise-assisted data analysis method[J]. Advances in Adaptive Data Analysis,2009,1(1):1-41.

[19] ZHANG X,LAI K K,WANG S Y. A new approach for crude oil price analysis based on empirical mode decomposition[J]. Energy Economics,2008,30(3):905-918.

[20] 赵建秀,王洪国,邵增珍,等. 一种基于信息熵的时间序列分段线性表示方法[J]. 计算机应用研究,2013,30(8):2391-2394.

[21] 孙焕良,邱邦华,魏溯华. 一种优化的自底向上时间序列分段算法[J]. 沈阳建筑大学学报:自然科学版,2007,23(6):1049-1052.

[22] 卞晓冰,蒋延学,贾长贵,等. 基于施工曲线的页岩气井压后评估新方法[J]. 天然气工业,2016,36(2):60-65.