doi:10.16018/j.cnki.cn32-1650/n.202104008

# 改进的时间序列关键点表示算法

刘永志[1,2],林 峰[1]

(1.福州职业技术学院 阿里巴巴大数据学院,福建 福州 350108;
2.南京航空航天大学 计算机科学与技术学院,江苏 南京 210016)

**摘要**:针对时间序列数据表示存在压缩率不高的问题,首先给出了时间序列定义和时间序列的9种基本形态;然后通过对极值点的优化处理,提出了关键极值点、剔除了轻微变化的极值点,并给出了结合转折点算法的IRAKPTS算法;最后通过实验验证,通过IRAKPTS算法的时间序列数据,较好地保留了时间序列的外形轮廓,并提高了压缩效率。

**关键词**:时间序列;关键点;转折点

**中图分类号**:TN402 **文献标志码**: **文章编号**:1671-5322(2021)04-0044-05

在物联网广泛应用的今天,如何表示采集到的时间序列数据是一个极具有挑战性的问题。由于这些海量时间序列具有瞬时波动性、持续性、噪声扰动严重等特点,如直接对采集的原始数据进行挖掘,不但传输、存储和挖掘的效率不高,而且还影响挖掘算法的准确性和可靠性,难以在相似性度量[1-3]、分类和聚类[4-6]、模式识别[7-10]等工作中获得令人满意的结果。因此,为了研究时间序列特点以方便数据处理,许多学者提出了时间序列的表示方法。因为时间序列的表示有3方面好处:一是对时间序列进行压缩处理,会带来更小的传输、存储和计算代价;二是只保留更能反映时间序列自身轮廓特征的重要关键点,而忽略其余点,因此在去除噪声干扰后的时间序列,更有利于提高挖掘的效率和准确性;三是研究者主要关心部分时段的数据演变规律,而不是整个原始数据或数据中某个元素的值。因此,合理的时间序列表示,更能满足这些领域的需求。

对时间序列表示的研究,许多学者提出了各种各样的方法,其中文献[11]将极值点原封不动地保留下来,而没有进行相应的处理,导致压缩比较低。针对已有的时间序列表示方法的不足,本文提出一种改进的时间序列关键点表示算法,即IRAKPTS(Improved Representation Algorithm of Key Points in Time Series)算法。

## 1 时间序列

对时间序列进行观察和研究,发现每个序列都可以表示为一个二元组$(T,D)$,其中$T$代表时间变量,$D$代表数据变量。由此,对时间序列给出如下定义:

**定义1** 时间序列$S$是一个长度为$n$的有限集,记$S=\{(T_1,D_1),(T_2,D_2),\cdots,(T_i,D_i),\cdots,(T_n,D_n)\}$,式中$T_i<T_{i+1}$ $(i=1,2,\cdots,n)$。如果$T_2-T_1=T_3-T_2=\cdots=T_i-T_{i-1}=\Delta t$,则称时间序列$S$为等时间间隔序列;否则,称为非等时间间隔序列。

以下时间序列如未经声明,则专指等间隔序列,可以简单记为$S=\{D_1,D_2,\cdots,D_i,\cdots,D_n\}$。

**定义2** 从时间序列$S$中选取能描绘$S$形态的点,称为关键点,关键点序列记为$S_{kp}=\{d_1,d_2,\cdots,d_i,\cdots,d_m\}$,式中$m<n$,$S_{kp}\subset S$。

**定义3** $\forall s\in S,\exists P\in M$,($M$为候选的模型集合,且$M\neq\phi$),使得$\hat{s}=P(s)$($\hat{s}$为原始序列$s$经由模型$P$产生的时间序列)。

**定义4** 给定阈值$\varepsilon>0$、距离度量$d$,如果$\forall s\in S,\exists P\in M$,有$d(\hat{s},s)\leqslant\varepsilon$成立,则称模型$P$与$s$近似,简记为$P\cong s$。

**收稿日期**:2021-03-09

**作者简介**:刘永志(1973— ),男,河南杞县人,教授,博士生,主要研究方向为时间序列数据挖掘。

## 2　时间序列基本形态

时间序列的基本形态单元可以划分为平稳、上升和下降，如图1所示。由这3类基本形态单元，可以组成9种变化形态，如图2所示。从图2可以发现除形态1没有转折点，其余形态都有转折点；形态6、9的转折点不可以用极大或极小值方法求出，而其余的形态都容易求出转折点。


**图1　时间序列的基本形态单元**

**Fig 1　Basic morphological units of time series**

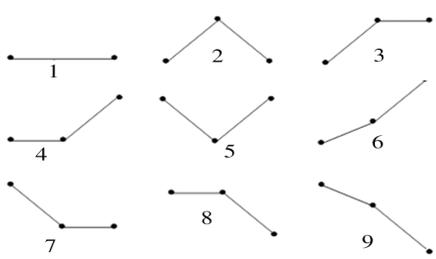



**图2　九种变化形态**

**Fig 2　Nine forms of change**

## 3　关键点表示

### 3.1　极值点表示

极值点是时间序列外观轮廓的特征表示。如何选择极值点表示时间序列，学者提出了极值点表示法、局域极值点表示法和特征点表示法，从不同侧面给出了如何选择极值点的问题。极值点表示法解决了一直上升、一直下降、上升后下降、下降后上升的极值点表示，如图2。对于轻微的变化，极值点也会表示出来，这有可能对时间序列的数据压缩表示不利，特别是对变化剧烈的序列。局域极值点表示法用比值的方式求极值点，克服了极值点表示的缺点，但对时间域没有约束，可能在指定的时间窗口内求不出极值点，也可能求出多个极值点，给研究带来不便。特征点表示法为了克服局域极值点表示法的缺点，提出了保持极值的时间长度与该序列长度的比值必须不小于某个比值$C$。这是从整个时间域来考虑的，对研究局域时间序列或者是在线表示时无能为力。基于以上分析本章提出新的极值点表示法。

**定义5**　基本极值点。如果时间序列中$D_{i-1}$、$D_i$和$D_{i+1}$满足式(1)的关系，则$D_i$被认为是基本极值点，记为$S_{bp}$(Basic Extreme Point)。

$$D_{i-1} < D_i > D_{i+1} \text{ OR } D_{i-1} < D_i = D_{i+1} \text{ OR}$$
$$D_{i-1} = D_i < D_{i+1} \text{ OR } D_{i-1} > D_i < D_{i+1} \text{ OR}$$
$$D_{i-1} > D_i = D_{i+1} \text{ OR } D_{i-1} = D_i > D_{i+1} \quad (1)$$

以上不等式分别对应图2中的2、3、4、5、7、8形态，利用定义5能识别出上面几种形态的全部极值点。

由于时间序列可能包含很多细微变化的波动极值点，因此需要对基本极值点进行再度鉴别，以获得关键极值点$S_{ke}$(Key Extremum Point)。

**定义6**　关键极值点。对于$S_{bp}$中的基本极值点$D_m$和$D_n$，如果满足式(2)的关系，且$D_m$为关键极值点，则$D_n$被认为是关键极值点。其中，$D_{max}$和$D_{min}$是时间序列的最大值和最小值，$\mu$是序列值阈值。

$$\frac{|D_n - D_m|}{|D_{max}| + |D_{min}|} \geqslant \mu \quad (2)$$

关键极值点$D_m$的识别如图3所示。该关键极值点的识别从值域进行了约束，在关键极值点$D_n$求出之后，按照时间顺序分别有极值点1、2、3、4和$D_m$，按照定义6分别将基本极值点与$D_m$进行比较并计算，求出$D_n$之后的关键极值点$D_m$，以此类推求出整个序列的关键极值点序列。

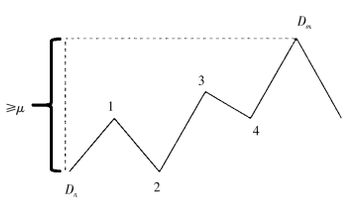



**图3　关键极值点$D_m$的判断**

**Fig 3　Judgment of key extreme point $D_m$**

**定义7**　在线关键极值点。对于$S_{bp}$中的基本极值点$D_m$和$D_n$，如果满足式(3)的关系，且$D_m$为关键极值点，则$D_n$被认为是在线关键极值点，$\sigma$是序列值阈值。

$$\frac{|D_n - D_m|}{|D_n| + |D_m|} \geqslant \sigma \quad (3)$$

如果想处理在线时间序列可以用此定义。

关键极值点算法1描述如下：

输入：时间序列集$S=\{D_1, D_2, \cdots, D_i, \cdots, D_n\}$，$\mu$

输出：关键极值点序列集$X^{ke}$

BEGIN

1. $D_1 \rightarrow X^{ke}$，$D_n = D_1$

2. for i=2 to n-1

3. if($D_{i-1} < D_i > D_{i+1}$ OR $D_{i-1} < D_i = D_{i+1}$ OR
$D_{i-1} = D_i < D_{i+1}$ OR $D_{i-1} > D_i < D_{i+1}$ OR
$D_{i-1} > D_i = D_{i+1}$ OR $D_{i-1} = D_i > D_{i+1}$) then
4. $D_i$->$X^{bp}$, $D_m$=$D_i$
5. if( $\frac{|D_n - D_m|}{|D_{max}| + |D_{min}|} \geqslant \mu$)
6. $D_i$-> $X^{ke}$, $D_n$=$D_i$
7. end if
8. end if
9. Next i
10. $D_n$-> $X^{ke}$
11. Return($X^{ke}$)
END

### 3.2 转折点表示

具体相关理论知识和算法见参考文献[11-13]。转折点算法2表示如下:

输入:时间序列集 $S=\{D_1, D_2, \cdots, D_i, \cdots, D_n\}$, $\varepsilon$
输出:转折点序列集 $X^{tp}$
BEGIN
1. $D_1$->$X^{tp}$
2. for i=2 to n-1
3. if $|D_{i+1}-2D_i+D_{i-1}|>\min(|D_{i-1}|,|D_i|,|D_{i+1}|)*\varepsilon$ then
4. $D_i$-> $X^{tp}$
5. end if
6. Next i
7. $D_n$-> $X^{tp}$
8. Return($X^{tp}$)
END

### 3.3 关键点表示

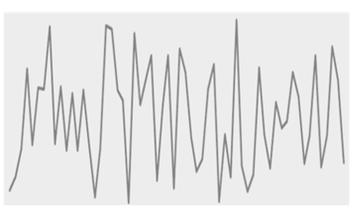
时间序列的外观轮廓由关键极值点和转折点组成,这些点本文统称为关键点,是时间序列的压缩和拟合的重要点。算法1和2分别给出了关键极值点 $X^{ke}$ 和转折点 $X^{tp}$ 的求法,根据上面的规定,即原始序列的关键点由关键极值点和转折点组成,从而提出了改进的时间序列关键点表示算法,即IRAKPTS算法。关键极值点算法如下:

输入:时间序列集 $S=\{D_1, D_2, \cdots, D_i, \cdots, D_n\}$, $\varepsilon$, $\mu$
输出:关键点序列集 $X^{kp}$
BEGIN
1. j=2, k=2, m=2, $D_{max}$=max(S), $D_{min}$=min(S)
2. $D_1$->$X^{kp}$, $D_1$->$X^{ke}$, $D_1$->$X^{tp}$
3. for i=2 to n-1
4. if($D_{i-1} < D_i > D_{i+1}$ OR $D_{i-1} < D_i = D_{i+1}$ OR
$D_{i-1} = D_i < D_{i+1}$ OR $D_{i-1} > D_i < D_{i+1}$ OR
$D_{i-1} > D_i = D_{i+1}$ OR $D_{i-1} = D_i > D_{i+1}$) then
5. $D_i$->$X^{bp}$(D), i-> $X^{bp}$(T) // $X^{bp}$(D)、$X^{bp}$(T)分别保存极值点、序列号
6. if( $\frac{|X^{ke}(D) - X^{bp}(D)|}{|D_{max}| + |D_{min}|} \geqslant \mu$) then
7. $X^{bp}$(D)-> $X^{ke}$(D), i-> $X^{ke}$(T)
8. end if
9. else if $|D_{i+1}-2D_i+D_{i-1}|>\min(|D_{i-1}|,|D_i|,|D_{i+1}|)*\varepsilon$ then
10. $D_i$-> $X^{tp}$(D), i-> $X^{tp}$(T)
11. end if
12. Next i
13. $D_n$->$X^{kp}$, $D_n$->$X^{ke}$, $D_n$->$X^{tp}$
14. for i=2 to len($X^{kp}$)+len($X^{tp}$) // 合并关键点序列
15. if($X^{ke}$(T)< $X^{tp}$(T)) then
16. $X^{ke}$-> $X^{kp}$, $X^{ke}$(T)++
17. else
18. $X^{tp}$-> $X^{kp}$, $X^{tp}$(T)++
19. end if
20. Next i
21. Return $X^{kp}$
END

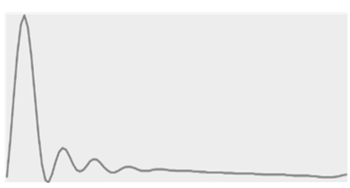
## 4 实验

### 4.1 实验数据

实验数据来源于Eamonn Keogh维护的经典的时间序列,网址为https://www.cs.ucr.edu/~eamonn/。选择的数据分别代表了周期性、变化平稳和变化剧烈3个方面。原始数据如图4所示。



a Synthetic_Control



b MedicalImages

c DistalPhalanxTW

**图4 原始数据图**

**Fig 4 Raw data diagram**

图4中Synthetic_Control为模拟控制数据,包

含6类数据，共300个训练集和300个测试集，每个数据集长度为60；MedicalImages为医疗图像，包含10类数据，共381个训练集和760个测试集，每个数据集长度为99；DistalPhalanxTW为图像数据，包含6类数据，共399个训练集和154个测试集，每个数据集长度为80。

### 4.2　实验设置

选择有代表性的Synthetic_Control数据，在Intel i7-8565U@1.8 GHz四核CPU、16 GB内存、1 TB硬盘、Microsoft Windows 10操作系统、Excel VBA环境下，研究IRAKPTS算法中$\varepsilon$和$\mu$对时间序列数据压缩的影响，结果图5、图6所示。其中，图5表示$\mu$不变，$\varepsilon$在一定范围内变化的情况；图6表示$\varepsilon$不变，$\mu$在一定范围内变化的情况。

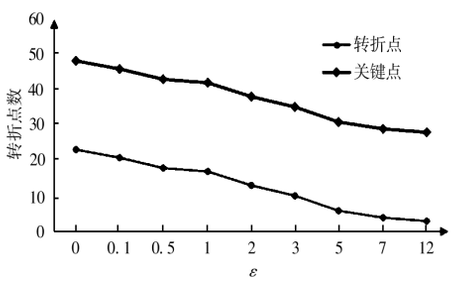



**图5　关键点随$\varepsilon$的变化情况**

**Fig 5　change of key points with $\varepsilon$**

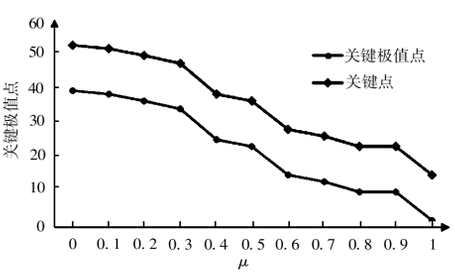



**图6　关键点随$\mu$的变化情况**

**Fig 6　Changes of key points with $\mu$**

从图5可以看出，转折点数和关键点数均随$\varepsilon$的增大而减少，即压缩率提高。从图6可以看出，关键极值点数和关键点数均随$\mu$的增大而减少，压缩率也提高了。但是，不论$\varepsilon$还是$\mu$，当他们增大到一定程度时，关键点的个数都不会再继续减少，即压缩率有一个极限值。

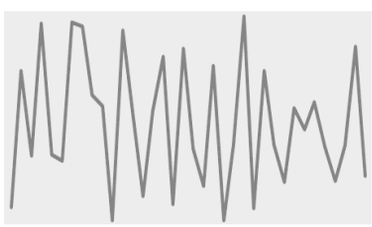
在$\varepsilon$和$\mu$取一定值的情况下，利用IRAKPTS算法对图4中的原始数据Synthetic_Control、MedicalImages和DistalPhalanxTW的任一数据集进行压缩，得到相关数据如表1所示。

**表1　利用IRAKPTS对3类数据的压缩**

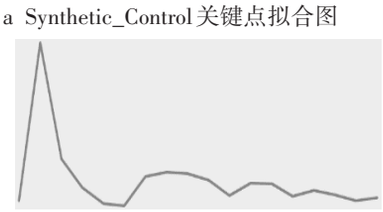
**Table 1　Compression of 3 types of data using IRAKPTS**

| 类别 | 数量 | $\varepsilon$ | $\mu$ | 关键点数 | 压缩率/% |
|---|---|---|---|---|---|
| Synthetic_Control | 60 | 1.9 | 0.35 | 36 | 40.0 |
| Medical Images | 99 | 0.6 | 0.02 | 18 | 81.8 |
| Distal PhalanxTW | 80 | 0.3 | 0.05 | 19 | 76.3 |

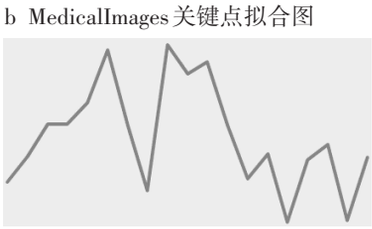
图7是利用表1的压缩进行线性拟合后的图形。与图4的原始数据图相比，可以发现图7拟合后的图形较好地保留了原始数据的外观轮廓，以及较好的压缩比例。



a Synthetic_Control关键点拟合图



b MedicalImages关键点拟合图



c DistalPhalanxTW关键点拟合图

**图7　利用IRAKPTS压缩后拟合的序列图**

**Fig 7　Sequence diagram fitted by IRAKPTS compression**

## 5　结语

针对时间序列数据表示存在压缩率不高的问题，通过对极值点的优化处理，提出了关键极值点、剔除了轻微变化的极值点；通过转折点算法，较好地保留了时间序列的外形轮廓，提高了压缩效率，并得到了实验的验证。

**参考文献:**

[1] YI B K, JAGADISH H V, FALOUTSOS C. Efficient retrieval of similar time sequences under time warping[C]//Proceedings 14th International Conference on Data Engineering. February 23-27, 1998, Orlando, FL, USA. IEEE, 1998: 201-208.

[2] 武天鸿, 翁小清, 单中南. 基于LDA符号表示的时间序列分类算法[J]. 计算机应用与软件, 2020, 37(2): 259-265,307.

[3] 沈培璐, 汪朝海, 钱源来, 等. 基于特征的时间序列信号表示方法[J]. 中国测试, 2020, 46(5): 13-18.

[4] 孙冬璞, 曲丽. 时间序列特征表示与相似性度量研究综述[J]. 计算机科学与探索, 2021, 15(2): 195-205.

[5] 郑诚, 王鹏, 汪卫. LS-Cluster: 大规模多变量时间序列聚类方法[J]. 计算机应用与软件, 2017, 34(5): 205-210,246.

[6] YIN Y, SHANG P J. Forecasting traffic time series with multivariate predicting method[J]. Applied Mathematics and Computation, 2016, 291: 266-278.

[7] YIN Y, SHANG P J. Multivariate multiscale sample entropy of traffic time series[J]. Nonlinear Dynamics, 2016, 86: 479-488.

[8] ALE J M, ROSSI G H. An approach to discovering temporal association rules[C]//Proceedings of the 2000 ACM symposium on Applied computing, 2000,1:294-300. https://doi. org/10. 1145/335603. 335770

[9] YANG K, WANG D H, LI H. Threshold autoregression analysis for finite-range time series of counts with an application on measles data[J]. Journal of Statistical Computation and Simulation, 2018, 88(3): 597-614.

[10] WANNER F, JENTNER W, SCHRECK T, et al. Integrated visual analysis of patterns in time series and text data: Workflow and application to financial data analysis[J]. Information Visualization, 2016, 15(1): 75-90.

[11] 刘永志, 皮德常, 贾学萍. 基于三点的时间序列关键点研究[J]. 微电子学与计算机,2015,32(1): 45-47,53.

[12] 刘永志, 皮德常, 陈传明. 基于关键点的不同长度时间序列相似性度量[J]. 计算机工程与应用, 2014, 50(20): 1-4,19.

[13] LIU Y Z, JIA X P,PI D C . Research on the application of the segmentation based on key points in the power consumption of wireless sensor[J],Journal of Applied Science and Engineering, 2016,19(1) :109-112.

# Improved Representation Alogrithm of Key Points in Times Series

LIU Yongzhi[1,2], LIN Feng[1]

(1. Alibaba Big Data School, Fuzhou Polytechnic, Fuzhou Fujian 350108, China;
2. College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing Jiangsu 210016, China)

**Abtract:** Aiming at the problem of low compression rate in time series data representation, firstly, the definition of time series and nine basic forms of time series are given. Then, through the optimization of extreme points, the key extreme points and the extreme points with slight changes are proposed, and the IRAKPTS algorithm combined with the turning point algorithm is given. Finally, it is verified by experiments that the time series data of the IRAPKTS algorithm retains the contours of the time series well and improves the compression efficiency.

**Keywords:** time series; key points; turning points

(责任编辑:李华云)