

基于三点的时间序列关键点研究

刘永志^{1,2}, 皮德常¹, 贾学萍³

(1 南京航空航天大学 计算机科学与技术学院, 江苏 南京 210016; 2 宣城职业技术学院 信息工程系,
安徽 宣城 242000; 3 宣城职业技术学院 建筑艺术系, 安徽 宣城 24200)

摘 要: 关键点(转折点及极值点)是刻画时间序列的最重要点, 由此提出了一种新的基于三点时间序列关键点算法. 首先, 分析了三点组成的序列形态; 其次, 提出了转折点定理和三点转折定理, 并给予了证明; 最后, 给出了算法及实验分析. 实验结果表明该算法能精确定位序列中的关键点.

关键词: 时间序列; 关键点; 数据挖掘; 三点

中图分类号: TN402

文献标识码: A

文章编号: 1000-7180(2015)01-0045-03

Study on Time Series Based on the Key Point of Three Points

LIU Yong-zhi^{1,2} PI De-chang¹, JIA Xue-ping³

(1 College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics,
Nanjing 210016, China; 2 Department of Information Engineering, Xuancheng Vocational &
Technical College, Xuancheng 242000, China; 3 Department of Architecture and Art,
Xuancheng Vocational & Technical College, Xuancheng 242000, China)

Abstract: The key point (turning point and extreme point) is the most important point characterization of time series. This paper proposes a novel algorithm of time series based on the three key points. First, It analyzes the time series form of three points; Secondly, it presents the turning point theorem and the three turning point theorem, And proved them; Finally, it gives the algorithm and experiment analysis. The experimental results show that the algorithm can accurately positioning key points in time series.

Key words: time series; key point; data mining; three points

1 引言

按照时间先后顺序排列的数据记录称为时间序列数据, 它广泛存在于工业、农业、经济、科学等众多领域^[1]. 如何分析和处理这些海量的时间序列数据, 并从中挖掘出一些潜在的、未知的、有价值的知识信息, 正受到越来越多的数据挖掘者的研究和关注^[2-3]. 由于时间序列固有的短期波动性、噪声干扰、非稳态及长期性等特点, 如果直接对原始时间序列进行研究, 存储量巨大而且效率低下, 难以获得令人满意的结果^[4].

自上世纪 70 年代以来^[5], 时间序列数据分段线

性的表示 PLR (piecewise linear representation) 的相关研究得到了极大热捧^[6-7]. PLR 是采用首尾相邻的系列线段去近似表示原始的时间序列. 但也存储了部分未过滤的细节变化, 降低了压缩精度^[8]. XIAO H 等人较好地处理了噪音细节, 但对单调序列中的转折点, 没有给出具体的解决方案^[9], 这些转折点恰恰是时间序列分段表示的关键. 喻高瞻等人改进了 PLR 算法, 提高了 PLR 描绘序列的功能, 将其简称为 NPLR 算法, 能捕获到极值点, 但遗失了较多的转折点^[10]. 本文针对以上算法的不足, 提出一种新的基于三点时间序列的关键点算法.

本研究详细分析了三点组成的基本形态, 提出

收稿日期: 2014-03-19; 修回日期: 2014-05-18

基金项目: 安徽省质量工程项目(20101452); 安徽高校基金重点课题(KJ2014A285)

一种新的关键点选择算法 TPS (Three points segmentation). 利用非单调序列极值点(关键点)及基于前两点的单位延长点与第三点的距离阈值比较求出转折点(关键点). TPS 算法仅需一次扫描序列数据库, 就可以保留非单调序列中重要的极值点, 并根据 TPS 算法对中间数据点是否需要保留转折点做出判断. TPS 算法仅保留数据序列变化的主要关键点, 极大减少了数据存储量, 提高了计算速度. 实验结果表明, 与以往研究方法相比, TPS 算法能够更加高效地选择关键点, 在保持原数据序列的变化趋势的同时还具有较高的压缩率, 能精确定位时间序列中的转折点.

2 时间序列

对时间序列进行观察和研究, 每个序列都可以抽象为一系列数据单元. 数据单元为一个二元组 (T, D) , 其中, T 为时间变量; D 为数据变量, 反映数据单元的具体含义, 如温度随时间的变化、股票价格的波动等. 由此, 对时间序列可以给出如下定义.

定义 1 时间序列 S 是一个有限集 $S = \{ (T_1, D_1), (T_2, D_2), \dots, (T_i, D_i), \dots, (T_n, D_n) \}$, 满足 $T_i < T_{i+1} (i = 1, 2, \dots, n)$. 如果 $T_2 - T_1 = T_3 - T_2 \dots = T_i - T_{i-1} = \Delta t$ 称为等时间间隔序列, 否则称为非等时间间隔序列.

如未特殊说明, 时间序列特指等时间间隔序列, 可以简记为 $S = \{ D_1, D_2 \dots D_i \dots D_n \}$.

定义 2 从时间序列 S 中抽取能刻画 S 形态的点, 称为关键点, 其他的点称为非关键点, 记关键点序列为 $S_{kp} = \{ d_1, d_2 \dots d_i \dots d_m \} m < n, S_{kp} \subset S$.

3 时间序列基本形态

时间序列的基本形态单元划分为平稳、上升和下降, 如图 1 所示, 由这三个基本形态单元可以构成九种变化形态, 即三点组成的形态如图 2 所示. 从图中可以看出 1 没有转折点, 其余都有转折点, 怎样求出转折点是研究者关心的问题, 经观察, 5、8 可以用极大和极小值方法求出, 其余的就不容易求出.



图 1 时间序列的基本形态单元

定理 1 由三点组成的简单时间序列, 在一定的约束条件下, 夹角越小, 角处的顶点成为转折点的

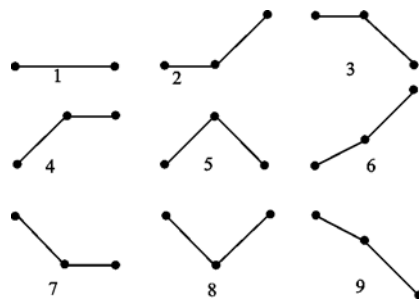


图 2 九种变化形态

可能性就越大, 反之成立, 称为转折点定理. 这里的约束为不超过序列有意义下条件可能的最大值和最小值.

证明: 为了讨论方便, 以图 2 中的 6 三点变化趋势为例, 并作图 3, 从图 3 中可以看出: $\angle AOC = 180^\circ$.

假设 AO 不动, OC 分别逆时针旋转到 B 点、 B' 点, 则 $\angle AOB' < \angle AOB < 180^\circ$.

当 AOC 是一条直线时, O 不是转折点, 但随着 OC 逆时针旋转, $\angle AOB$ 角度逐渐变小, 顶点 O 处的转折凸显, 直到旋转至 Max 最大值时, 超出约束条件, 角度变小也无意义, 定理成立.

同理可证, 当 OC 顺时针旋转时, 定理亦成立. 其他情况同理可证.

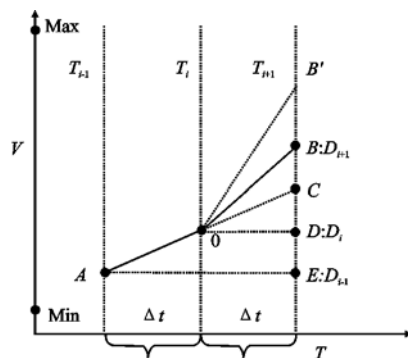


图 3 基本序列三点关系图

注: 图 3 横轴表示时间 T , 纵轴表示值域 $V[\text{Min Max}]$, 超出约束条件部分不是有效值. D_{i-1} 是 A 点在 T_{i-1} 时刻的值, D_i 是经过 Δt 时 T_i 时刻 O 点的值, D_{i+1} 是经过 Δt 时 T_{i+1} 时刻 B 点的值, C 是延长向量 AO 交于 T_{i+1} 时刻的交点, D 是通过 O 点平行于时间轴 T 与 T_{i+1} 时刻的交点.

转折点定理在实际操作中存在一定局限性, 因角度不仅与序列的值有关, 还与采样时间间隔 Δt 有关, 经研究, 得出如下定理.

定理 2 由三点组成的基本时间序列, 其值为 D_{i-1}, D_i, D_{i+1} ; D_i 处是否为转折点与 $|D_{i+1} - 2D_i +$

D_{i-1} | 的值有关,值越大,成为转折点的可能性就越大,反之亦成立,称为三点转折定理。

证明:以图3为例,

∵ 在 $\triangle ACE$ 中 $\overrightarrow{OD} \parallel \overrightarrow{AE}$, 采样时间间隔 Δt 相同

$$\therefore \frac{\overrightarrow{OD}}{\overrightarrow{AE}} = \frac{1}{2}, \frac{\overrightarrow{CD}}{\overrightarrow{CE}} = \frac{1}{2}$$

又 D 和 C 处值已知,由欧氏几何可知 C 处的值为 $2D_i - D_{i-1}$ 。

由转折点定理可知,距离 C 点越远, O 点为转折点的可能性越大,即: $|D_{i+1} - 2D_i + D_{i-1}|$ 值越大,距离点 C 越远, O 成为转折点的可能性越大。证毕。

4 算法描述

规定关键点由极值点和转折点组成,时间序列求关键点的算法如下。

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

输出:关键点序列集 $S_{kp} = \{d_1, d_2 \cdots d_i \cdots d_m\}$

步骤:

1 $d_1 = D_1$

2 for $i = 2$ to $n-1$

3 if $((D_i > D_{i-1} \text{ and } D_i > D_{i+1}) \text{ or } (D_i < D_{i-1} \text{ and } D_i < D_{i+1}))$

4 $\{d_j = D_i$

5 $j = j + 1\}$ // 极值求关键点

6 else if $|D_{i+1} - 2D_i + D_{i-1}| > \min(D_{i-1}, D_i, D_{i+1}) * \epsilon$

7 $\{d_j = D_i$

8 $j = j + 1\}$ // 利用三点转折定理求关键点

9 end if

10 end for

11 $d_m = D_n$

5 性能分析

实验环境为 Pentium IV 双核 CPU, 2 GB 内存和 160 GB 硬盘, Microsoft Windows XP 操作系统, 开发工具为 Java。利用 NPLR 和 TPS 两种算法求出关键点, 然后对原始数据序列的分段拟合效果进行对比分析。所用测试数据为 1700 年至 1987 年的太阳黑子数, 此数据序列包含 288 个数据点。压缩比如表 1 所示, 拟合效果如图 4 所示, 阈值 $\epsilon = 0.8$ 。

由实验结果发现, 尽管利用 NPLR 算法求出关键点进行拟合的序列压缩效果较好, 也能过滤噪音和细节干扰, 但无法描绘一些重要的转折点, 存在部分失真。TPS 算法没有 NPLR 压缩效率高, 但保留了关键的转折点, 故能较好的拟合原始序列。

表 1 两种算法的压缩比

算法	拟合点数	压缩比/%
NPLR	67	76.7
TPS	118	59

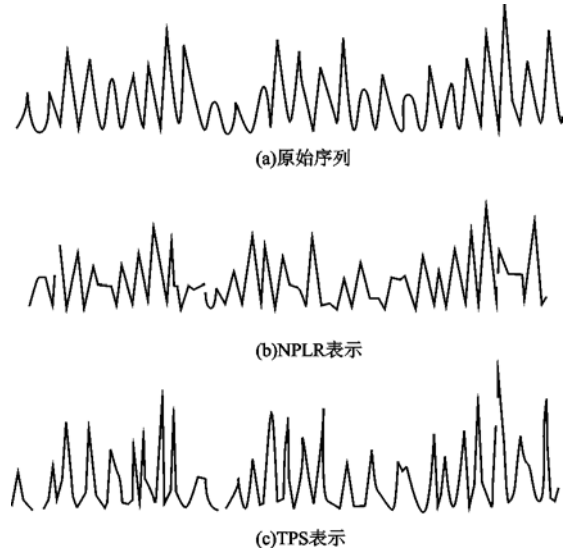


图 4 利用关键点拟合效果图

6 结束语

提出了一种新的基于三点的关键点算法 TPS, 利用非单调序列极值点及基于前两点的单位延长点与第三点的距离阈值大小比较求关键点。TPS 算法具有以下特点: 扫描一次序列数据库, 并确定数据点是否为求的关键点; 时间复杂度为 $O(n)$ 。实验结果与理论分析表明, TPS 算法能够实时有效地找出原始时间序列的关键点, 在去除了噪音干扰的同时保留了时间序列主要形态。

参考文献:

- [1] 贾澎涛, 何华灿, 刘丽, 等. 时间序列数据挖掘综述[J]. 计算机应用研究, 2007, 24(11): 15-17.
- [2] 林子雨, 江弋, 赖永炫, 等. 一种新的时间序列延迟相关性分析算法—预测探查法[J]. 计算机研究与进展, 2012, 49(12): 2645-2655.
- [3] 李正欣, 张凤鸣, 李克武. 多元时间序列模式匹配方法研究[J]. 控制与决策, 2011, 36(4): 565-570.
- [4] 李海林, 郭崇慧. 基于多维形态特征表示的时间序列相似性度量[J]. 系统工程理论与实践, 2013, 33(4): 1024-1034.
- [5] Pavlidis T, Horwitz S. Segmentation of plane curves [J]. IEEE Transactions on Computer, 1974, 23(8): 860-870.

(下转第 53 页)

北京:清华大学出版社,2008.

- [2] 张思扬,匡芳君,徐蔚鸿.混沌动态模糊神经网络及变压器故障诊断的应用[J].湖南科技大学学报:自然科学版,2012,27(1):35-39.
- [3] Bergh F. Ananalysis of particle swarm optimizers [D]. South Africa, Department of Computer Science, University of Pretoria, 2006: 118-123.
- [4] Gheitanchi S, Ali F, Stipidis E. Particle swarm optimization for adaptive resource allocation in communication networks [J]. EURASIP Journal on Wireless Communications and Networking, 2010(4): 1-13.
- [5] 刘道华,原思聪,兰洋,等.混沌映射的粒子群优化方法[J].西安电子科技大学学报,2010,37(4):764-768.
- [6] 唐贤伦.混沌粒子群优化算法理论及应用研究[D].重庆:重庆大学,2007.
- [7] 范会联,陶维安.一种改进的自适应进化粒子群优化算法[J].微电子学与计算机,2011,28(5):11-14.
- [8] 肖文显,刘震.一种融合反向学习和量子优化的粒子群算法[J].微电子学与计算机,2013,30(6):126-130.
- [9] 任小波,杨忠秀.一种动态扩散粒子群算法[J].计算机应用,2010,30(1):159-161.
- [10] 李祥飞,邹莉华,彭可.一类模糊神经网络结构的混沌

优化设计[J].计算机应用,2007,26(6):1411-1414.

- [11] University of California. UCI Machine Learning Repository: Data Sets [EB/OL]. (2012-06-10). [2014-04-05]. <http://archive.ics.uci.edu/ml/datasets.html>.
- [12] Wang Chao, Wang En Yuan, Xu Jiankun, et al. Bayesian discriminant analysis for prediction of coal and gas outbursts and application [J]. Mining science and technology, 2009(3): 520-523.
- [13] 陈全秋.遗传算法与神经网络相结合在煤与瓦斯突出危险预测中的应用研究[D].太原:太原理工大学,2010.

作者简介:

匡芳君 女,(1976-),博士,副教授.研究方向为模式识别、智能信息处理技术及应用、信息安全等. E-mail: kfjztb@126.com.

张思扬 男,(1973-),硕士,高级工程师.研究方向为人工智能、故障诊断、电子技术等.

徐蔚鸿 男,(1963-),博士,教授,博士生导师.研究方向为人工智能、模式识别、软件工程等.

(上接第47页)

- [6] 李爱国,覃征.在线分割时间序列数据[J].软件学报,2004,15(11):1671-1678.
- [7] 朱天,白似雪,王柏,等.基于时间段的时序规则发现[J].通信学报,2009,30(8):112-115.
- [8] Park S, Kim S, Chu W. Segment-based approach for subsequence searches in sequence databases[C] // Proceedings of the 16th ACM Symposium on Applied Computing. New York: ACM Press, 2001: 248-252.
- [9] Xiao H, Hu Y F Data mining based on segmented time warping distance in time series database[J]. Journal of Computer Research and Development, 2005, 42(1):

72-78.

- [10] 喻高瞻,彭宏,胡劲松,等.时间序列数据的分段线性表示[J].计算机应用与软件,2007,24(12):17-18.

作者简介:

刘永志 男,(1973-),博士研究生,副教授,高级工程师.研究方向为数据挖掘. E-mail: y_zliu@163.com.

皮德常 男,(1971-),博士,教授,博士生导师.研究方向为数据挖掘.

贾学萍 女,(1977-),硕士,讲师.研究方向为农业数据处理.