一种基于关键点的时间序列线性表示方法

陈帅飞 吕 鑫 戚荣志 王龙宝 余 霖
(河海大学计算机与信息学院 南京 211100)

摘 要 时间序列数据具有规模大、维度高等特点,直接在原始序列上进行数据挖掘,其计算复杂度高且易受噪声影响,因此对原始时间序列进行预处理是必不可少的,而常用的线性表示方法大多存在对分段点的筛选准确度不高的问题。基于时间序列的变化特征,提出了一种基于时间序列关键点的线性表示方法。该方法综合考虑了时间跨度和振幅变化,能高效提取时间序列中的关键点,并防止过度除噪,实现简单。实验表明,该方法对不同领域的数据具有良好的普适性。

关键词 数据挖掘,时间序列,线性表示,关键点,过度除噪
中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.5.043

Linear Representation Method Based on Key Points for Time Series

CHEN Shuai-fei LV Xin QI Rong-zhi WANG Long-bao YU Lin
(College of Computer and Information, Hohai University, Nanjing 211100, China)

Abstract Time series data has the features of large scale and high latitude. It has high computational complexity and is susceptible to noise if doing data mining on the raw sequence directly, so the original time series pretreatment is essential, and most methods of commonly used linear representation have low accuracy in selection piecewise points. Based on the time series variation, we proposed a linear representation method based on key points for time series. The method takes into account the time span and amplitude changes and can efficiently extract key points in the time series, which can prevent excessive noise removal and is implemened simply. Experiments show that the method has good universality for data from different areas.

Keywords Data mining, Time series, Linear representation, Key points, Excessive noise removal

1 引言

在数据挖掘领域中,时间序列数据挖掘是一个热点研究方向[1]。时间序列是指按照时间顺序采集的一系列观测数据,并且采样的时间间隔相等。时间序列在股票、科学实验、气象、农业、医疗等领域都广泛存在。时间序列数据通常具有高维和海量的特点,若直接对原始数据进行数据挖掘(相似性搜索[2]、关联规则挖掘[3]、聚类[4]和分类[5]等),效率很低。解决该问题的方法是对原始的时间序列进行预处理操作。在时间序列的预处理方面,已有了很多成熟的方法,例如傅里叶变换[6](DFT)、离散小波变换[7](DWT)、分段聚合近似[8,9]、奇异值分解(SVD)[10]等等。以上几种算法虽然在某些方面优势明显,但是也存在如下不足:离散傅里叶变换、离散小波变换和分段聚合近似存在一个共同的缺点,即消除了局部极值点,从而引起时间序列重要信息缺失;离散小波变换只能应用于满足特定长度的时间序列,具有局限性;而奇异值分解法的时间复杂度比较大。针对以上几种算法存在的不足,时间序列线性表示[11,12]方法被提了出来,其主要目的是保留时间序列的主要特征,忽略局部细微的改变。

近年来,国内外许多学者对线性表示方法进行了深入的研究,提出了许多优秀的线性表示方法。时间序列线性表示方法的主要思想就是用一系列前后相互连接的线段来近似表示原始时间序列,其关键在于选取分段点[13]。线性表示方法在时间序列表示方面有如下优点,即实现简单、压缩率高,在保留时间序列主要趋势的前提下又保留了局部特征,同时还降低了噪声数据的干扰,有利于提升数据挖掘的精度。文献[11]介绍了一种基于重要点的线性表示(PLR-IP)方法,即如果一个点在局部区间内与区间端点的比值超过设定的阈值 R,则认为它是重要点。通过调节阈值 R 的值,可以获得不同精细程度的线性表示。文献[12]提出了基于特征点的线性表示方法(PLR-FP),该方法的思想是首先提取时间序列的极值点,然后根据每个极值点保持的时间跨度去除噪声点。文献[14]介绍了一种基于斜率提取边缘点的时间序列线性表示方法(PLR-SEEP),该方法的主要思想是首先设定阈值,然后根

到稿日期:2015-05-15 返修日期:2015-10-22 本文受国家自然科学基金面上项目(61272543),国家科技支撑计划(2013BAB06B04),国家自然科学基金委-广东联合项目(U1301252),江苏省博士后科研资助计划(1401001C)资助。
陈帅飞(1989-),男,硕士生,主要研究方向为时间序列的表示方法、时间序列的相似性搜索,E-mail:chenshuaifei163@163.com;吕 鑫(1983-),男,博士后,主要研究方向为密码学、网络信息安全。

据斜率变化来选取分段点。文献[8,9]各自独立地提出了分段聚集近似(PAA)的时间序列线性表示方法,该方法的主要思想是首先将时间序列按照相同的时间跨度进行划分,然后以每个时间序列子段的平均值来近似表示相应子段。Perng等人[15]提出的基于界标模型(Landmark Model)的时间序列线性表示方法以界标作为分段点对时间序列进行线性表示,界标的选取依据最小距离和百分比原则(MDPP)。以上几种方法虽然都能够保留原始时间序列的主要特征,但是在分段点的取舍方面依然还有很大的研究空间。本文从选取关键点过程中避免过度除噪这个角度出发,提出了一种基于关键点的线性表示(PLR-KP)算法,并且针对不同领域的数据集进行实验。实验结果表明,本文提出的线性拟合方法由于采用了时间跨度和振幅变化相结合的除噪声机制,提高了关键点的搜索准确度,并降低了拟合误差,且该方法对不同领域的数据适应性良好。

2 相关知识

2.1 基本概念

定义 1(时间序列) 时间序列是一组有序集合,集合中的每个元素由采集时间和采集值构成,记为 $X=\langle(x_1,t_1),(x_2,t_2),\cdots,(x_n,t_n)\rangle$,元素$(x_i,t_i)$表示时间序列在 t_i 时刻的采集值为 x_i,采集时间 t_i 是严格增加的,一般情况下,时间序列的采样周期也是固定的,其采样间隔是不变的,如果有 $t_1=0,\Delta t=1$,则时间序列 X 简记为 $X=\langle x_1,x_2,\cdots,x_n\rangle$。

定义 2(时间序列分段线性表示) 设有时间序 $X=\langle x_1,x_2,\cdots,x_n\rangle$,分段点的集合是 $X'=\langle x_1',x_2',\cdots,x_m'\rangle$,其中$x_1'=x_1,x_m'=x_n,m<n$。

X 的分段线性表示为:

$$X_L=\langle f_1(x_1',x_2'),f_2(x_2',x_3'),\cdots,f_{m-1}(x_{m-1}',x_m')\rangle \tag{1}$$

其中,$f_{m-1}(x_{m-1}',x_m')$表示在区间$[x_{m-1}',x_m']$内的线性拟合函数。

定义 3(拟合误差) 给定一组时间序列 $X=\langle x_1,x_2,\cdots,x_n\rangle$,通过对时间序列 X 进行分段取得分段点集合 X_{PLR},再利用线性插值的方法对 X_{PLR} 进行填充得到衍生的拟合时间序列,记为 $X^c=\langle x_1^c,x_2^c,\cdots,x_n^c\rangle$,则拟合序列与原始序列的拟合误差为:

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

拟合误差是衡量拟合时间序列与原始时间序列差异的一个重要指标,在同等压缩率情况下,拟合误差越小,拟合效果越好。

2.2 选取关键点

参照定义 2,分段线性表示就是从时间序列中筛选出一些分段点,然后用直线段将前后相邻的两个点依次连接起来近似表示原时间序列。通过该方法可以提取出对时间序列影响较大的点,删除一些无关紧要的点,从而有利于在后续研究中进行高效的数据挖掘。根据实际的情况分析,人们观察一个时序序列时,首先关心的是序列模式中的关键点,本文中研究的关键点分两类,一类是满足一定条件的极值点,另一类是非极值点中幅度变化较大的点。

定义 4(极值点) 如果时间序列 $X=\langle x_1,x_2,\cdots,x_n\rangle$上的某点满足下列条件之一:

(1)$x_i>x_{i-1}$且 $x_i\geqslant x_{i+1}$,或 $x_i\geqslant x_{i-1}$且 $x_i>x_{i+1}$,其中$1<i<n$;

(2)$x_i<x_{i-1}$且 $x_i\leqslant x_{i+1}$,或 $x_i\leqslant x_{i-1}$且 $x_i<x_{i+1}$,其中$1<i<n$。

定义 5(波动幅度大的点,即非极值点) 假定 x_{i-1}、x_i、x_{i+1}为时间序列相邻的 3 点,若

$$\frac{(|x_{i-1}|+|x_{i+1}|)/2-|x_i|}{||x_{i-1}|-|x_{i+1}||}>P \tag{3}$$

则 x_i 为波动幅度较大的点。

在相邻两个极值点之间的单调区间上,选取一定数量的波动幅度大的点作为关键点可以有效降低拟合误差。

定义 6 时间序列 X 的关键点是指满足下列条件的点:

(1)首先根据定义 3 和定义 4 选出候选关键点,然后根据条件(2)进行过滤;

(2)如果候选关键点 x_i'保持时间段的跨度大于某个阈值 C(阈值 C 由压缩率确定),那么该候选点是关键点,如果上述条件不满足,但是当 $x_i'>x_{i-1}'$时,$x_i'/x_{i-1}'>R$ 成立,或当 $x_i'<x_{i-1}'$时,$x_{i-1}'/x_i'>R$ 成立,那么该点为关键点。

3 基于关键点的时间序列线性表示

3.1 选取时间序列关键点

算法流程如下。

输入:时间序列 $X=\langle x_1,x_2,\cdots,x_n\rangle$和阈值 C、P、R

输出:时间序列关键点集合

算法步骤:

第 1 步:对原始时间序列进行规范化处理,找出序列中的最大值 x_{max}和最小值 x_{min},计算出它们差的绝对值 Δx,然后将原始序列中的每个点值 x_i 与 x_{min} 相减,得出的差值与 Δx 相除得到一个[0,1]之间的数值,由此产生一个新的衍生序列作为下一步输入。

第 2 步:对候选关键点集合进行初始化,将起点和终点加入候选关键点(KP)集合。

第 3 步:从第二个点开始依次将所有满足条件的点加入候选关键点集合,依据定义 5 中对关键点的表述,关键点从两类点中取得,一类是极值点,另一类是变化幅度大的非极值点。首先根据极值点的定义判断某个点是否为极值点,如果该点满足定义 4,那么该点是极值点,将其加入候选关键点集合;如果该点不是极值点,再根据定义 5,如果满足条件,将其加入候选关键点集合,否则舍去。

第 4 步:根据设定的阈值 C(关键点保持的跨度)和 R(关键点的变化的幅度)对上一步选取的候选关键点进行筛选:根据定义 6 中关键点的定义,对候选关键点集合进行过滤,先判断候选关键点保持的时间跨度,如果保持的时间跨度大于或等于阈值 C,则认为该点是关键点,保留此点;当跨度小于阈值 C 时,再根据候选关键点与前一关键点的波动幅度来判断,如果大于等于阈值 R,保留此点,否则删除。

第 5 步:输出关键点集合。

3.2 算法分析

PLR-KP 时间序列线性表示算法通过对时间序列进行两遍扫描选取关键点，并以关键点作为分段点对时间序列进行分段表示，其时间复杂度为 $O(n)$，n 为时间序列的长度。该时间序列线性表示算法综合考虑了时间跨度和振幅变化，因此较好地保留了时间序列的关键点，并且可以有效地避免过度除噪。

3.3 时间序列进行线性表示

用 PLR-KP 算法对时间序列进行关键点的筛选，从而得到一组有序的时间序列分段点集合，依据时间上的先后顺序将这些关键点依次连接起来，便可得到时间序列的近似分段线性表示。

假定有时间序列 X，其关键点集合为 $X'=\langle x_1', x_2', \cdots, x_m' \rangle$，根据定义 2，则时间序列基于关键点的线性表示如下：

$$X_L=\langle f_1(x_1', x_2'), f_2(x_2', x_3'), \cdots, f_{m-1}(x_{m-1}', x_m') \rangle$$

4 仿真实验

4.1 实验目的及实验环境

为了验证本文提出的基于关键点的线性表示算法的实际效果，选取了不同领域的 10 条时间序列数据集，以拟合误差为评价优劣的指标对各种线性表示方法进行对比。

4.2 实验数据

本文中的时间序列取自于 www. cs. ucr. edu/～eamonn/tutorials. html 公布的用于数据挖掘的通用时间序列数据集（本文简称为 KData），如表 1 所列。

表 1 KData 数据集

序列名称	序列长度	序列名称	序列长度
Burst	9382	Memory	6875
Chaotic	1800	Ocean	4096
Fluid_dynamics	10000	Powerplant	2400
Earthquake	4096	Speech	1020
Leleccum	4320	Tide	8746

4.3 实验方法

本文方法将与 Pratt 和 Fink 提出的基于重要点（PLR-IP）的线性表示方法[11]、Yi 和 Keogh 提出的分段聚集近似（PAA）线性表示方法[8]、Xiao 提出的基于特征点（PLR-FP）的分段线性表示方法[12] 3 个算法进行对比。

本文提出的基于关键点分段表示方法 PLR-KP 需要输入 3 个参数，参数 P 表示对于非极值点选为关键点所应满足的阈值下限，参数 C 表示极值点选为关键点应满足的最小跨度，参数 R 表示极值点选为关键点与已确定的前一重要点的最小比值。实验选取的时间序列来自不同领域，每个时间序列的取值范围差别很大，为了便于对各种线性表示方法进行对比，需要先对原始时间序列进行规范化操作，先将时间序列的取值范围规范化到[0,1]区间，之后再进行线性表示。时间序列规范化公式如下：

$$norm(x_i)=\frac{x_i-\min(X)}{\max(X)-\min(X)} \tag{4}$$

算法的性能优劣指标主要考虑各种 PLR 表示方法与原时间序列之间的拟合误差。由于两种分段线性表示算法的输入参数不同，为了保证算法的公平性，在压缩率相同的情况下，分别采用不同 PLR 线性表示方法计算与原时间序列之间的拟合误差。采用式(2)来计算拟合误差，在同一数据集、同一压缩率情况下，拟合误差越小，算法性能越好。

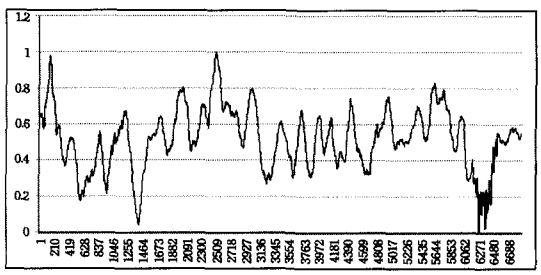
4.4 实验结果与分析

实验结果如表 2 所列，从中可以看出，在 10 个通用时间序列数据集合中，本文提出的基于关键点（PLR-KP）的线性表示算法在其中 6 个数据集上的拟合误差均为最小，在另外 4 条数据集上的效果与其它 3 种算法相当。实验结果表明，在采用了时间跨度和振幅变化相结合的选取关键点的机制后，对于短时间内波动频率较低的时间序列，该方法拟合效果良好，除了保留整体形态波动较大的极值点，还适当地保留了两个极值点之间波动幅度较大的非极值点，减小了一些重要点被粗暴舍弃的概率，从而降低了拟合误差。

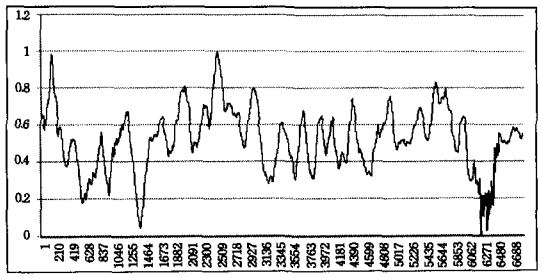
表 2 压缩率为 80%时几种线性表示算法的拟合误差对比

数据集值＼算法	PLR-IP	PAA	PLR-FP	PLR-KP
Burst	0.89	0.38	0.90	0.45
Chaotic	1.63	1.76	1.63	0.85
Earthquake	2.10	3.48	2.22	2.66
Fluid_dynamics	2.30	2.77	2.38	1.51
Leleccum	0.88	0.64	0.98	0.71
Memory	0.50	0.56	0.49	0.39
Ocean	0.38	0.31	0.31	0.30
Powerplant	2.23	1.07	1.11	1.05
Speech	1.52	2.48	3.22	1.20
Tide	2.29	3.22	2.48	3.41

同时实验结果表明：对于短时间内波动频率比较平缓的时间序列，PLR-KP 分段算法的实验结果比较良好（见图 1），但是对于短时间内波动频率剧烈的时间序列（见图 2），拟合效果一般。因为波动频率剧烈的时间序列的极值点数量较多，在高压缩率的情况下很大一部分极值点被粗暴舍弃，从而造成拟合误差较大；而对于波动频率较小的时间序列，极值点数量有限，除了重要的极值点被保留外，同时又加入了一些波动幅度大的非极值点作为补充，因此拟合效果较好。

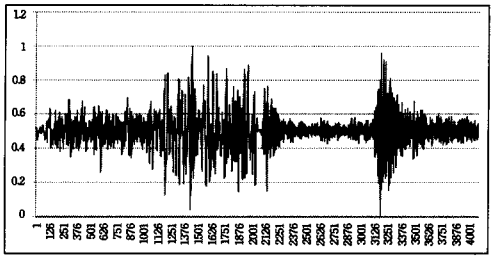


(a) Memory 原始序列

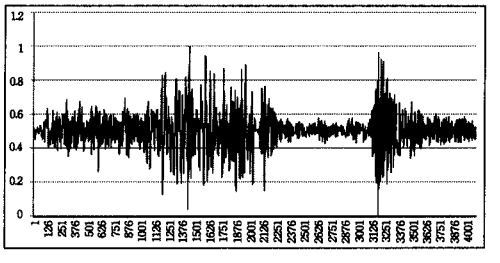


(b) Memory 拟合序列（压缩率 80%）

图 1



(a) EarthQuake 原始序列



(b) EarthQuake 拟合序列(压缩率 80%)

图 2

结束语 时间序列具有海量性、高纬性等特点,因此在进行时间序列数据挖掘之前需要选择一种合适的方法对原始序列进行预处理,从而降低其复杂度并提高其数据挖掘结果的准确性。预处理的主要目标有两个:降低维度和消除噪声。

本文基于综合考虑已有的时间序列选取分段点的思想,提出了一种结合时间跨度和振幅变化选取关键点的方法,并基于关键点对时间序列进行线性表示。该方法通过引入一种改进的筛选分段点的机制,在保留原始时间主要趋势的前提下降低了拟合误差。该方法对于短时间内变化平缓的时间序列拟合效果较好。下一步的研究重点是对于短时间内波动频率比较高的时间序列寻找有效的拟合方案。

参 考 文 献

[1] Pan Ding, Shen Jun-yi. Similarity Discovery Techniques in Temporal Data Mining[J]. Journal of Software, 2007, 18(2): 246-258 (in Chinese)
潘定,沈钧毅. 时态数据挖掘的相似性发现技术[J]. 软件学报, 2007, 18(2): 246-258

[2] Keogh E. Fast similarity search in the presence of longitudinal scaling in time series databases[C]// Proceedings of the International Conference on Tools with Artificial Intelligence, 1997. Washington: IEEE Computer Society, 1997: 578-584

[3] Das G, Lin K I, Mannila H, et al. Rule Discovery from Time Series[C]// KDD-98. New York: KDD, 1998: 16-22

[4] Debrégeas A, Hébrail G. Interactive Interpretation of Kohonen Maps Applied to Curves[C]// KDD-98. New York: KDD, 1998: 179-183

[5] Hellerstein J M, Koutsoupias E, Papadimitriou C H. On the analysis of indexing schemes[C]// Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems-PODS, 1997. Tucson: ACM, 1997: 249-256

[6] Agrawal R, Faloutsos C, Swami A. Efficient similarity search in sequence databases[C]// Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms, 1993. London: Springer Berlin Heidelberg, 1993: 69-84

[7] Chan K P, Fu A W C. Efficient time series matching by wavelets [C]// Proceedings International Conference on Data Engineering, 1999. Sydney: IEEE, 1999: 126-133

[8] 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

[9] Yi B K, Faloutsos C. Fast time sequence indexing for arbitrary Lp norms[C]// Proceedings of the 26th VLDB Conference, 2000. Cairo: VLDB, 2000: 385-394

[10] Wu D, Singh A, Agrawal D, et al. Efficient retrieval for browsing large image databases[C]// International Conference on Information and Knowledge Management, 1996. Rockville: ACM, 1996: 11-18

[11] Pratt K B, Fink E. Search for patterns in compressed time series [J]. International Journal of Image and Graphics, 2002, 2(1): 89-106

[12] Xiao H, Feng X F, Hu Y F. A new segmented time warping distance for data mining in time series database[C]// Proceedings of 2004 International Conference on Machine Learning and Cybernetics, 2004. Shanghai: IEEE, 2004: 1277-1281

[13] Liu Shi-yuan, Jiang Hao. A Review on Time Series Representation for Similarity-based Pattern Search[J]. Computer Engineering and Applications, 2004, 40(27): 53-59 (in Chinese)
刘世元,江浩. 面向相似性搜索的时间序列表示方法述评[J]. 计算机工程与应用, 2004, 40(27): 53-59

[14] Zhan Yan-yan, Xu Rong-cong, Chen Xiao-yun. Time Series Piecewise Linear Representation Based on Slope Extract Edge Point [J]. Computer Science, 2006, 33(11): 139-142 (in Chinese)
詹艳艳,徐荣聪,陈晓云. 基于斜率提取边缘点的时间序列分段线性表示方法[J]. 计算机科学, 2006, 33(11): 139-142

[15] Perng C S, Wang H, Zhang S R, et al. Landmarks: a new model for similarity-based pattern querying in time series databases[C]// Proceedings International Conference on Data Engineering, 2000. San Diego: IEEE, 2000: 33-42

(上接第 222 页)

[9] Candès E J, Tao T. The power of convex relaxation: Near-optimal matrix completion[J]. IEEE Transaction on information, 2010, 56(5): 2053-2080

[10] Candès E J, Recht B. Exact matrix completion via convex optimization[J]. Foundation of Computational Mathematics, 2008, 9: 717-772

[11] Chen Cai-hua, He Bing-sheng. Matrix completion via alternating direction method[J]. IMA Journal of Numerical Analysis, 2012, 32: 227-245

[12] Konstan J A, Miller B N, Mahz D, et al. GroupLens: Applying collaborative filtering to usenet news[J]. Communications of the ACM, 1997, 40(3): 77-87

[13] Toh K C, Yun S. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problem[J]. Pacific Journal of Optimization, 2010, 6: 615-640

[14] Gang Wu, Swaminathan, Mitra, et al. Online video session progress prediction using Low-Rank matrix completion[C]// IEEE International Conference on Multimedia and Expo Workshops (ICMEW). 2014: 14-18

[15] Anupriya G, Angshul M. SVD free matrix completion with online bias correction for Recommender Systems[C]// International Conference on Advances in Pattern Recognition (ICAPR). 2015: 4-7