第35卷 第22期 计 算 机 工 程 2009年11月
Vol.35 No.22 Computer Engineering November 2009
·博士论文· 文章编号：1000—3428(2009)22—0032—03 文献标识码：A 中图分类号：TP311.13

时间序列周期模式挖掘的周期检测方法

王 阅，高学东，武 森，陈 敏
(北京科技大学经济管理学院，北京 100083)

摘 要：周期是时间序列的重要特征之一，用于精确描述时间序列并预测其发展趋势。在现有周期模式挖掘算法中，周期长度由用户事先定义，忽略了噪声的存在。在ERP度量和时间弯曲算法的基础上，提出一种新的周期长度检测方法。该方法可以在时间轴上实现弯曲，包括延伸和平移。它受噪声干扰的影响较小，实验结果表明其性能优于原有周期检测算法。

关键词：时间序列；数据挖掘；周期检测；动态时间弯曲

Periodicity Detection Method of Periodic Pattern Mining in Time Series

WANG Yue, GAO Xue-dong, WU Sen, CHEN Min
(School of Economics and Management, University of Science and Technology Beijing, Beijing 100083)

【Abstract】 Periodicity is an important feature for time series that can be used for describing time series exactly and predicting its development trends. In existing mining algorithms for periodic patterns, the periodicity length is user-specified in andvanc, and the presence of noise is not taken into account. Based on ERP(Edit distance with Real Penalty) measurement and time warping algorithm, this paper proposes a novel algorithm for periodicity length detection, which can realize warp on the time axis including extending and translation. It is less affected by noise interference. Experimental results show that the performance of this algorithm is better than existing periodicity detection algorithms.

【Key words】 time series; data mining; periodicity detection; Dynamic Time Warping(DTW)

1 概述

周期模式挖掘是时态数据挖掘领域的一个研究热点。在现实世界中，严格周期定义下的准周期模式很少，多数模式并非在每个时间点上都呈现周期性。当模式呈现部分周期特性时，可以认定模式是满足一定置信度的周期模式。周期检测的目的是发现时间序列的周期长度，周期模式挖掘过程中存在2个问题，即如何确定周期长度和去除噪声。文献[1]介绍了部分周期模式和最大子模式匹配集的概念，文献[2]在文献[1]工作的基础上研究部分周期模式的增量式在线合并算法。算法挖掘部分周期模式时，需要用户事先定义周期长度，在不了解背景知识的情况下，挖掘结果是错误的，甚至没有意义。也有学者用所有可能的周期长度来运行算法，其计算代价极大。因此，文献[3]提出基于卷积的CONV方法，文献[4]提出WARP方法,文献[5]提出基于小波的AWSOM方法。CONV方法只需扫描数据库一次，时间复杂度为 $O(n\log n)$，其中，n 为时间序列的长度。WARP方法的抗噪声能力很强，但时间复杂度较高，为 $O(n^2)$。AWSOM方法受小波函数的限制只能发现长度为 2^l 的周期模式，其中，l 为小波系数。

事先确定周期长度对大型数据库来说很困难，噪声的存在影响了算法准确性，且周期检测主要是为后续挖掘工作提供有利帮助，对检测算法的准确性和复杂度要求很高。本文在相似性匹配和时间弯曲ERP(Edit Distance with real Penalty)度量的基础上，提出一种新的周期检测方法，以解决周期挖掘中如何确定周期长度的问题。

2 时间序列与周期检测

2.1 时间序列的周期

给定时间序列 $V=e_1,e_2,\cdots,e_t$，将 V 的特征值离散化为若干等级(如高、中、低)，每个等级用一个字母表示(如 a, b, c)，字母集为 $\Sigma=\{a,b,c,\cdots\}$，则 V 是有限集 Σ 上的符号序列。如果 $V=S'S^P$，则称 S 是 V 的周期模式，其中，S' 是 S 的前缀，$p(\ p\geq 2\ )$是任意正整数。即 V 是模式 S' 的 p 次重复连接。如 aac 是序列 $aacaacaacaacaa$ 的周期模式。由于噪声的影响，模式的周期是不严格的。

符号序列的周期定义如下[3]：给定时间序列 V，将 V 分割成长为 p 的多个片段，每段都近似相同，则称 p 为 V 的周期。例如，$S_1=abcabcabc$ 的周期 p=3，$S_2=abcabdabc$ 的周期 p=3，abc 与 abc 近似相同。

2.2 汉明距离和ERP距离

周期检测的关键问题是相似性度量，时间序列符号化后的字符匹配问题可以用汉明距离来度量序列间的相似度，但汉明距离度量的符号序列必须是等长的，长度不等、时间间隔不完全一致的序列不适合进行相似性度量。例如，直观地看，$S_1=s_1s_2s_3s_4s_5s_6$ 和 $S_2=s_1s_2s_7s_3s_4s_5$ 很相似，但它们的汉明距离为4，相似度函数 $D=1-4/6=0.33$，因此，判断它们为非相似的。可见，需要其他能处理噪声影响的相似性度量方法。

编辑距离和动态时间弯曲(Dynamic Time Warping, DTW)是字符匹配的常用度量[4]，它们不满足三角不等式，其原因

基金项目：国家自然科学基金资助项目(11260011)；教育部新世纪优秀人才支持计划基金资助项目(NCET-05-0097)

作者简介：王 阅(1981－)，女，博士研究生，主研方向：数据挖掘，管理信息系统；高学东，教授、博士生导师；武 森，副教授、博士；陈 敏，博士研究生

收稿日期：2009-05-18 **E-mail：**ycwangyue@163.com

是当被比较序列在时间轴上某处出现空缺时，用空缺的前一个位置上的元素进行填充，填充元素与空缺之间的差异会因为前位元素的不同而不同。

ERP 距离[6]可以进行动态计算。给定序列 $R=(r_1,r_2,\cdots,r_n)$ 和 $S=(s_1,s_2,\cdots s_n)$ (序列长度可以相等或不等)，则 ERP 为

$$ERP=\begin{cases}\sum_{i=1}^{n}|s_i-g| & m=0\\ \sum_{i=1}^{m}|s_i-g| & n=0\\ \min\{ERP(Rest(R),Rest(S))+dist_{erp}(r_1,s_1),ERP(Rest(R),S)+\\ \quad dist_{erp}(r_1,gap),ERP(R,Rest(S))+dist_{erp}(s_1,gap)\} & \text{else}\end{cases} \tag{1}$$

其中，$dist_{erp}$ 是 r_i 与 s_i 的 ERP 距离；$Rest(R)=(r_2,r_3,\cdots,r_n)$。

$$dist_{erp}(r_i,s_i)=\begin{cases}|r_i-s_i| & \text{当}r_i,s_i\text{不是空缺}\\ |r_i-g| & \text{当}s_i\text{是空缺}\\ |s_i-g| & \text{当}r_i\text{是空缺}\end{cases} \tag{2}$$

其中，g 是常数。

ERP 结合了编辑距离与动态时间弯曲的优点，能处理局部时间位移并满足三角不等式。但 ERP 与编辑距离在计算处理方法的不同之处在于，ERP 方法不需要事先定义距离阈值。ERP 与 DTW 的不同之处在于不使用之前元素作为空缺的填充值，而用常数 g 来填充。文献[6]已证明当 g=0 时，填补了空缺后的序列与原序列相同，用填补序列进行相似性比较对结果没有影响，且距离度量满足三角不等式。

ERP 的计算过程如下：先建立一个 $n\times n$ 的矩阵 $\boldsymbol{D}$，$\boldsymbol{D}$ 中的元素 $cell(i,j)$ 为 r_i 与 s_j 的 $dist_{erp}(r_i,s_j)$ 值。从元素 $cell(1,1)$ 到 $cell(n,n)$ 之间的路径 $M=m_1,m_2,\cdots,m_k$ 称为弯曲路径，其中，m_k 对应 $cell(i_k,j_k)$，即 $m_k=dist_{erp}(i_k,j_k)$。由 ERP 距离矩阵 $\boldsymbol{D}$ 可知，弯曲路径有许多条，ERP 计算的目标是寻找弯曲路径总长度最小的路径(序列最相似)，即

$$D(R,S)=\min\left\{\sum_{k=1}^{K}m_k\Big/K\right\} \tag{3}$$

其中，分母 K(最长序列的长度)保证了比较不同长度的路径时有统一的标准。

由动态规划理论可知，如果点 $cell(i,j)$ 在最佳路径上，那么从点 $cell(1,1)$ 到 $cell(i,j)$ 的子路径也是局部最优解，即从点 $cell(1,1)$ 到点 $cell(n,n)$ 的最佳路径可以通过递归搜索从起点 $cell(1,1)$ 到终点 $cell(n,n)$ 之间的局部最优解获得，即

$$\gamma(i,j)=\min\begin{Bmatrix}d(r_i,s_j)+\gamma(i-1,j-1),\\ \gamma(i-1,j)+d(r_i,g),\\ \gamma(i,j-1)+d(s_j,g)\end{Bmatrix} \tag{4}$$

最终的时间序列弯曲路径为最小累加值 $\gamma(n,m)$，从 $\gamma(n,m)$ 起沿弯曲路径按最小累加值倒退，直到起始点 $\gamma(1,1)$，就能找到整个弯曲路径。

ERP 的计算过程与 DTW 的计算过程基本相同，时间和空间复杂度都为 $O(n^2)$。汉明距离可以视为 DTW 的特例，汉明距离的弯曲路径是矩阵的对角线，即 $i_k=j_k=k$。

3 周期检测算法

基于 ERP 的周期检测算法(ERP-based Period Detection Algorithm, ERPP)主要思想如下：先建立时间序列 $S=(s_1,s_2,\cdots,s_n)$ 的 ERP 矩阵，矩阵中的元素是序列与其自身序列之间的 ERP 距离，然后从矩阵的所有次对角线上寻找满足周期阈值的周期模式，该模式的长度即为周期。ERP 矩阵 $\boldsymbol{D}$ 描述如下：

$$\boldsymbol{D}=\begin{pmatrix}0 & d(s_1,s_2) & \cdots & d(s_1,s_n)\\ d(s_2,s_1) & 0 & \cdots & d(s_2,s_n)\\ \vdots & \vdots & & \vdots\\ d(s_n,s_1) & d(s_n,s_2) & \cdots & 0\end{pmatrix}$$

$\boldsymbol{D}$ 是对称方阵，所有主对角线上的元素都为 0，时间序列与它自身的 ERP 距离为 0。矩阵 $\boldsymbol{D}$ 表示序列中所有元素之间的比较关系以及每个次对角线包含的序列与自身平移序列之间的比较关系。检测周期即计算序列 S 与它自身平移序列之间的 ERP 距离。例如，第 1 条次对角线是从 $cell(1,2)$ 到 $cell(n-2,n-1)$、对应 S 与 S^1 的距离，即 $d(s_1,s_2),d(s_2,s_3),\cdots,d(s_{n-2},s_{n-1})$，第 2 条次对角线是从 $cell(1,3)$ 到 $cell(n-3,n-1)$、对应 S 与 S^2 的距离。依此类推，第 i 条次对角线是从 $cell(1,i)$ 到 $cell(n-i-1,n-1)$、对应 S 与 S^i 的距离。

对于给定长度为 n 的时间序列，可能的周期值为 $p=1,2,\cdots,n/2$。计算从 $cell(1,p)$ 开始的最小弯曲路径 M_p，对应弯曲路径 M_p 的路径累加值为序列 S 与 $S^{(p)}$ 之间的 ERP 距离，用动态规划方法计算累积距离为

$$ERP(S,S^{(p)})=\gamma(n-p-1,n-1) \tag{5}$$

当 $ERP(S,S^{(p)})$ 取最小值时，表示 S 与 $S^{(p)}$ 之间高度相似，因此，p 是序列 S 的最优候选周期值，$ERP(S,S^{(p)})$ 取最大值 n-p 时，表示矩阵中比对相应位置上的符号，所有符号都不相同，即符号序列不具备周期性。因此，定义周期长度 p 的置信度为

$$\frac{n-p-ERP(S,S^{(p)})}{n-p} \tag{6}$$

候选周期值的置信度应大于等于给定的周期阈值。

周期置信度说明一个模式必须连续出现一定次数才认为该段时间序列具有周期性。在呈现周期性的时间序列中存在一些噪音，但噪音出现的次数不能超过一定范围，若超出一定范围，则认为周期不再持续。在 ERP 矩阵中，当某一周期模式频繁出现时，$ERP(S,S^{(p)})$ 趋向于 0，此时置信度趋向于 1，说明此时周期值 p 最可信。由于噪声的影响而不再呈现周期性时，$ERP(S,S^{(p)})$ 会增大，置信度趋向于 0，因此此时序列无周期性。

上述方法存在 2 个问题：(1)主对角线上的零对最小弯曲路径的影响。如果一个特定的周期值 p 的最小弯曲路径恰好和主对角线重合，则 $S^{(p)}$ 中平移的 p 个点在 S 和 $S^{(p)}$ 的比对中被忽略。例如，对于图 1 中序列 $S=s_1s_2s_1s_3s_4s_5$ 的 ERP 矩阵，当 p=2 时，最小弯曲路径用虚线圆表示，由于主对角线上序列值的距离都为 0，因此将弯曲路径拉到对角线上而与实际路径偏离。为了解决该问题，ERPP 算法将主对角线上的零值设置为无穷大，无穷大把最小弯曲距离推离主对角线，如图 1 中的实线圆所示，它表示变换后的最小弯曲路径。(2)如果 p 是一个候选周期值，那么第 p 条次对角线上会有很多零，因此，会牵引邻近的弯曲距离。第(2)个问题与第(1)个问题类似，例如，序列 $S=s_1s_2s_3s_4s_1s_2s_3s_4$ 的最佳周期 p=4，而 ERP 矩阵会显示 p=3 也为候选周期，候选周期的弯曲路径和周期形成如图 2 所示的形状。ERPP 算法将局部最小弯曲路径对应的周期长度作为候选周期长度。

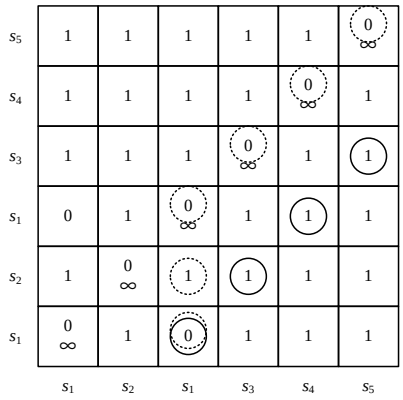

图 1　S=$s_1s_2s_3s_4s_5$ 的 ERP 弯曲路径求解

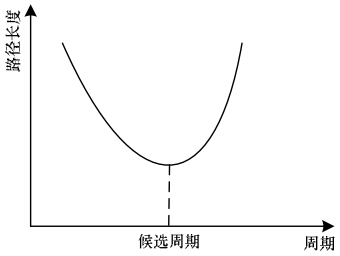

图 2　路径长度与周期

4　实验分析

使用 Cylinder-Bell-Funnel 数据集验证算法效率，并与 CONV, WARP 和 AWSOM 方法进行比较。数据集中包含 cylinder(C), bell(B)和 funnel(F) 3 种事件。先重复预先定义的模式生成一定长度的周期模式，如序列 FFBCFFBC…。序列服从正态分布(用 N 表示)和均匀分布(用 U 表示)，序列中的噪声用其他符号表示，在随机选取的位置上进行插入、删除和替换操作生成带噪声的时间序列。

取不同周期阈值，算法的计算精度不同。一般来说，周期阈值越大候选周期的可信度越高。因此，定义候选周期 p 的置信度为检测到 p 的最小阈值，用所有候选周期的平均置信度评价算法精度。图 3 给出了算法处理不同嵌入式周期的精度，其中，U.P 表示嵌入有异常值的周期序列；N.P 表示嵌入的是无异常值的周期序列。在 85%的置信水平下，几乎能发现所有嵌入式周期。图 4 比较了 4 种算法的精度，ERPP 算法比 CONV 算法更精确，且对于所有的周期值，该算法的精度是均衡的。CONV 和 AWSOM 方法分别适用于长周期和短周期的检测。

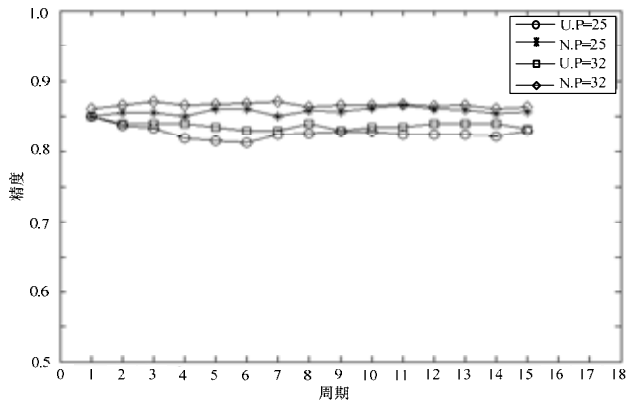

图 3　ERP 算法的精度

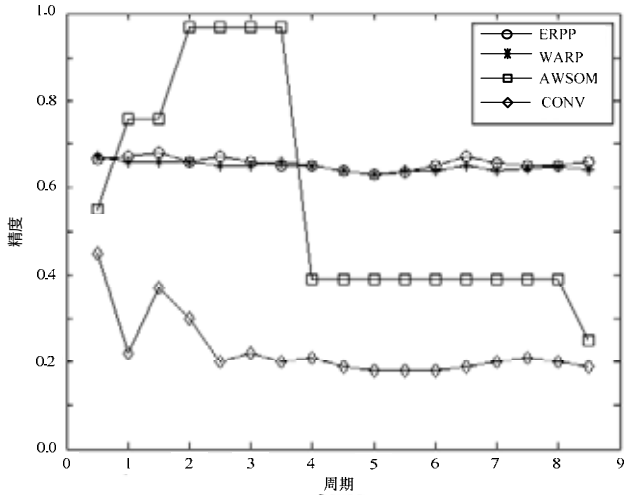

图 4　4 种算法的准确度

从图 5 中可以看出，ERPP 的抗噪声能力和 WARP 基本相同，比 AWSOM 和 CONV 方法更好。但 RELAX 算法的精度随着噪声率的增加单调减少，而 ERPP 算法的精度随着噪声率的增加而增加。

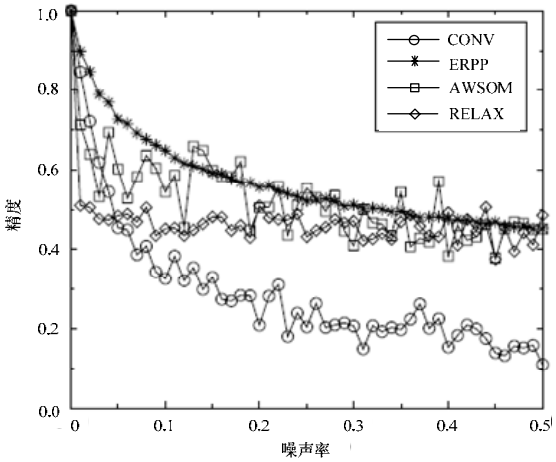

图 5　4 种算法的抗噪能力

ERPP 的精度和抗噪声能力较好，但 ERPP 比 CONV 的计算复杂度高，这是因为 ERP 算法是 DTW 算法的扩展，复杂度为 $O(n^2)$，而 CONV 的复杂度为 $O(n\log n)$。AWSOM 算法的复杂度为 $O(n\log^2 n)$。

5　结束语

本文提出的 ERPP 算法利用时间弯曲度量的原理，在不同位置上延伸和平移时间轴，以去除噪声，适合对有噪声干扰的符号时间序列进行周期检测。与现有周期检测方法相比，ERPP 方法抗噪能力很强但运算时间较长，它在运算精度和运行效率之间采取了折中策略。ERP 度量有很好的度量性质，可以利用下边界定理和三角不等式进行索引。下一步工作的重点是降低算法的时间复杂度，并研究处理在线时间序列的方法。

参考文献

[1] Han Jiawei, Dong Guozhu, Yin Yiwen. Efficient Mining of Partial Periodic Patterns in Time Series Databases[C]//Proc. of the 15th Int'l Conf. on Data Engineering. Sydney, Australia: [s. n.], 1999: 106-115.

(下转第 37 页)