一种时间序列相似性的快速搜索算法

张 军[1,2], 陈汉武[1], 马志民[1]

(1.东南大学 计算机科学与工程系, 江苏 南京 210096;
2.江苏海事职业技术学院 信息工程系, 江苏 南京 211170)

[摘要] 时间序列数据库中相似子序列的搜索, 常用滑动窗口、分形插值逼近等方法将时间序列分割成各子序列, 线性拟合各分段子序列, 计算查询序列与各子序列的欧氏距离, 满足距离阈值条件的为相似子序列. 这些方法忽略了时间序列本身的位置和连贯特性. 为此提出时间序列变化关键点的概念, 以检索出的关键点为边界分割时间序列, 线性拟合各分割的子序列, 计算查询序列和各子序列的形态距离, 快速搜索出相似子序列.

[关键词] 时间序列, 数据挖掘, 相似子序列, 形态距离

[中图分类号] TP311.1, [文献标识码] A, [文章编号] 1672-1292(2005)03-0050-04

An Algorithm for Similar Sub-patterns Discovery From Time Series

ZHANG Jun[1,2], CHEN Hanwu[1], MA Zhimin[1]

(1. Department of Computer Science and Engineering, Southeast University, Jiangsu Nanjing 210096, China;
2. Department of Information Engineering, Jiangsu Maritime Institute, Jiangsu Nanjing 211170, China)

Abstract General method of similar sequence mining based on time series is to transform time series into discrete character series and cluster them into different sets; then compute the Euclidean distance between querying series and these sets to measure their similarity. These methods ignore the position and holistic characteristic of time series and work with high time complexity, according to which this paper proposes an algorithm of searching the key points which divides the time series into line segments. After checking the fitness of each line segments, we can quickly mine the similar sub-sequence with pattern distance measurement and quick pruning method.

Key words time series, data mining, similar sub-sequence, pattern distance

0 引言

时间序列是一种在金融、水文、气象等许多领域都普遍使用的重要数据, 时间序列中相似子序列的搜索是时间序列数据挖掘中的一项很有意义的工作. 该问题可描述为给定某个时间序列, 从一个大型时间序列数据库中找出与之相似的子序列, 需要研究相似性的度量标准和检索过程中的时间复杂性. 时间序列的连续数值形式表示的结构不便于挖掘过程的描述和计算, 需对这种连续数值形式作等时间间隔采样, 将连续数值形式表示成离散的序列, 根据研究问题的实际需要决定采样时间间隔大小. 将连续数值形式的时间序列 X 表示成离散的时间序列 $\{x_1, x_2, \cdots, x_i, \cdots, x_n\}$, i是均匀采样时刻点, x_i 为 i时刻的振幅值, 均匀采样点数 n 的选择以能够比较完整地表达出连续数值形式为标准. 为分析方便, 以下所讨论的时间序列都认为已经离散化成等时间间隔的序列.

时间序列分段线性拟合已有一些方法, 例如: Gautam Das等人[1] 提出将一个时间序列经过固定宽度窗口分割, 得到等长的时间序列片段构成的分段序列, 以各段采样点振幅的平均值来代替各段振幅, 既能够压缩时间维, 又能够用欧氏公式计算各分段的相似距离. 如: 连续数值形式的时间序列 X 和 Y可以离散成 $X = \{x_1, x_2, \cdots, x_i, \cdots, x_m\}$ 和 $Y = \{y_1, y_2, \cdots, y_i, \cdots, y_m\}$, 其中 x_i, y_i 分别为各个时

收稿日期: 2005-03-18
基金项目: 国家自然科学基金资助项目(90412014).
作者简介: 张军(1973-), 讲师, 硕士研究生, 主要从事计算机应用等方面的学习和研究. E-mail: njhxzhj@163.com
通信联系人: 陈汉武(1955-), 博士, 教授, 主要从事现代信息化和数据处理等方面的教学与研究. E-mail: hw_chen@seu.edu.cn

刻点的振幅, 两时序 X, Y之间的欧氏距离表示为:

$$D(X, Y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

给定阈值 $\varepsilon > 0$, 若$D(X, Y) < \varepsilon$, 则两时序X, Y相似. 该方法存在着一些问题: (1) 对具体的原始数据, 如何确定合理的分隔窗口宽度才能满足实际研究需要; (2) 时间序列不经过去噪声、特征提取和变换等预处理, 造成挖掘时计算量大, 分割效果不理想; (3) 以平均值来代替各段振幅, 可能造成序列的某些重要特征 (极大值、极小值) 丢失. 又如李斌等人[2] 在分形插值逼近理论的基础上提出, 在时间序列上依次取相邻 3点, 拟合成线段. 在允许累积的误差范围内将其合并成更长的直线段, 以标识符号序列代替各直线段, 形成字符串集, 用比较字符串相似的方法查找相似子序列. 这种方法的出发点是时间序列最基本的变化形态, 能够解决前一种方法的部分问题, 但是时间序列数据内在的整体特征被忽略了, 而且没有保留各分段在时间序列中的位置信息, 挖掘效果受影响.

因此, 研究一种高效的时间序列相似性搜索方法很有实际意义. 直观观察时间序列曲线图, 会发现序列变化中有相对重要影响的点通常是局部极大值和极小值点, 这些点能够反映时间序列整体的变化趋势和主要特征模式, 则称这些点为关键点.

本文首先提出一种检索关键点的算法, 以关键点为边界划分成各子序列. 各子序列考虑实际采样点的振幅值的分布, 以最大似然函数和最小二乘法求出各分段线性拟合函数 $y = \hat{a} + \hat{b}x$, 以线性拟合函数式中 $\hat{b}$的值为形态相似比较的基本单元, 经过快速处理, 搜索出相似子序列.

如一段时间序列 s(图 1(a)), 局部放大其中的子序列 s_1、s_2 (图 1(b)、图 1(c)). 如采用欧氏距离计算, 两者差距值很大, s_1, s_2 不是相似序列; 但用下文中的形态相似计算方法, s_1, s_2 是相似模式, 这种方法允许时间序列中适当的噪声和扰动, 以及适度的时间轴相位平移和振幅伸缩.

1 检索关键点

假设以下分段函数模型能够拟合时间序列 X[3]:

$$X = \begin{cases} \boldsymbol{f}_1(t, \boldsymbol{w}_1) + e_1(t), & 1 \leqslant t < \alpha_1 \\ \quad\vdots \\ \boldsymbol{f}_k(t, \boldsymbol{w}_k) + e_k(t), & \alpha_{k-1} \leqslant t < \alpha_k = N \end{cases}$$

其中 $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_k)$ 是时间序列 X 的关键分段点的集合, 关键点是时间序列趋势上升或下降的变化的分界点. $e_1(t)$, $e_2(t)$, …, $e_k(t)$是第 i段的绝对误差项, $e_i(t)$ 为满足均值为零的高斯白色噪声分布的函数. $\boldsymbol{f}_i(t, \boldsymbol{w}_i)$ 为时间序列第 i段的拟合多项式函数 ($1 \leqslant i \leqslant k$), $\boldsymbol{w}_i$ 是系数向量, $\boldsymbol{f}_i(t, \boldsymbol{w}_i) \in M$, M 为线性模型.

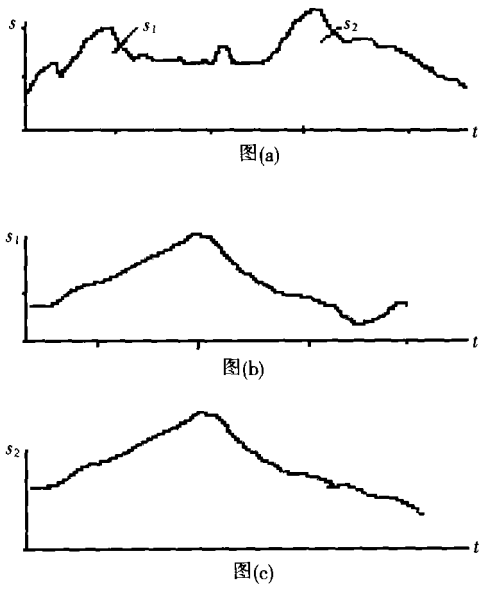

图 1 形态相似计算方法下两相似子序列

检索分段关键点的目的是使整个拟合误差 G 值最小, G值越小, 说明总体拟合越接近原时间序列, 随着关键点增多, G 值会减少, 计算量增大, 时间序列分割零碎, 不便于研究时间序列的整体趋势特征, 因此最优方法是在拟合误差 G 值给定的条件下, 关键分割点的总数量最少.

检索关键点的算法如下:

输入: 时间序列 X; 增幅比阈值 δ

输出: 关键点集合 cp.

算法描述:

(1) 扫描时间序列数据库, 找出时间序列库中振幅最大值 x_{max}, 最小值 x_{min}.

(2) 时间序列进行规范化预处理, 对各点的振幅 x_i 作如下变换:

$$x_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

振幅 x_i的值限制在 $[0, 1]$ 之间, 可以消除振幅平移和时间缩放对相似性计算的影响.

(3) 以时间为序计算 $|(x_{i+1} - x_i)/x_i|$的值, 且依次与给定的增幅比阈值 δ($\delta > 0$), 如大于增幅比阈值 δ, 在集合 cp中记录时刻 t_i 值和振幅值 x_i.

(4) 检索出满足阈值 δ条件的关键点, 根据实际研究需要, 可适当调整 δ的值, 重新查找时间序列变化的关键点.

cp 集合中每一点是时间序列趋势上升或下降变化的关键点, 时间序列上升或下降过程中增幅小于给定 δ的可忽略, 如图 2中的 $\theta(\theta < \delta)$ 变化可忽略, 实现时间序列理想化处理, 在一定精度条件下可得到全部变化关键点.

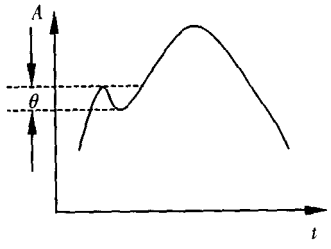

图 2　幅度变化小于阈值 δ 的忽略

cp 集合中每一点为分界点, 将时间序列分割为各段子序列, 考虑时间序列实际复杂性, 不能直接将各关键点的连接线代替各子序列, 需对每段子序列作一元线性回归拟合, 线性拟合方程如下:

$$\begin{cases} y_i = a + bx_i + \varepsilon_i \\ \varepsilon \sim N(0, \sigma^2) \end{cases}$$

式中, ε为随机误差. 用最小二乘法求 a, b的最大似然估计[4]:

(1) 构作似然函数.

因为 $y_i \sim N(a + bx_i, \sigma^2)(i = 1, 2, \cdots, n)$, 作似然函数:

$$L = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - a - bx_i)^2}{2\sigma^2}} =$$

$$(2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{\sum_{i=1}^{n}(y_i - a - bx_i)^2}{2\sigma^2}}$$

(2) 求 a, b的最大似然估计.

令函数 $Q(a, b) = \sum_{i=1}^{n}(y_i - a - bx_i)^2$, 要使 L 为最大, 根据函数的极值性质, $Q(a, b)$对 a, b求偏导, 即求出 $Q(a, b)$ 最小值, 联立方程组得:

$$\begin{cases} \dfrac{\partial Q}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \\ \dfrac{\partial Q}{\partial b} = -2\sum_{i=1}^{n}(y_i - a - bx_i)x_i = 0 \end{cases}$$

用最小二乘法求 a, b的最大似然估计, 解方程组得:

$$\begin{cases} \hat{a} = \bar{y} - \hat{b}\bar{x} \\ \hat{b} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{cases}$$

式中 $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$, $\bar{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$, 即各子序列中每个采样点振幅的平均值. 将求解出的 $\hat{a}$、$\hat{b}$代入线性拟合方程, 得近似经验回归方程式 $\hat{y} = \hat{a} + \hat{b}x$, 依次计算出各分段的拟合方程中的 $\hat{b}_i$.

设某工厂 100个月实际用电负荷数据时间序列原始图如图 3所示, 使用检索关键点算法, 查找出 25个变化关键点, 如图 4中矩形框标出部分, 直观看出关键点往往也是极值点. 以关键点为边界分段线性拟合, 得到 24个线性子段合成的曲线 S, 如图 6所示.

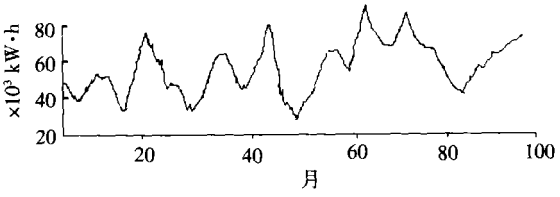

图 3　工厂 100个月用电负荷原始曲线

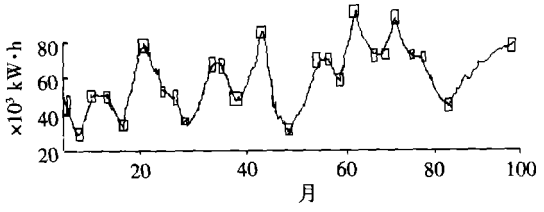

图 4　用电负荷原始曲线关键点分布

用固定窗口算法作为对比, 如图 5所示, 设窗口宽度为 5个月, 共检索出 20个点作为边界分段, 这些分界点 (如图 5中虚线标出) 往往不是曲线变化关键点, 原始时间序列数据中的部分关键点没有发现, 以每个窗口中振幅平均值代替各分段, 时间维能够压缩, 但极值被平均值平滑丢失.

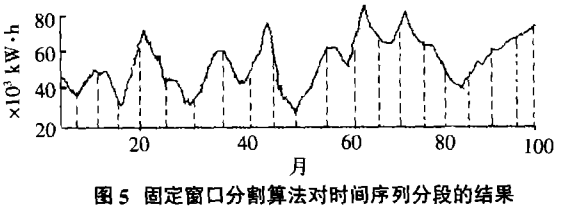

图 5　固定窗口分割算法对时间序列分段的结果

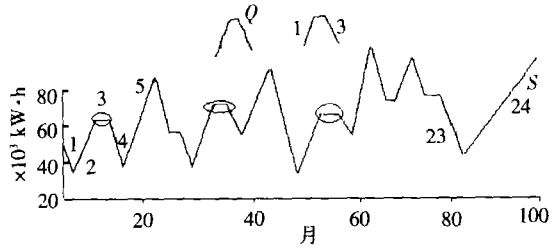

图 6　用电负荷曲线分段线性拟合

2 形态相似性计算

设有时间序列 S, 查询时间序列 Q, (S的长度远大于 Q 的长度), 在 S 中查询与 Q 相似的子序列集, 将 Q 采用与 S 完全相同 (增幅比阈值 δ相同) 的预处理, 如分割成 $(q_1, q_2, \cdots, q_i, \cdots, q_j)$, q_i 为关键点集, mp 为关键点集合. 同样计算出各分段 b_i, 根据如下的快速剪除比较算法, 检索出相似子序列.

形态相似子序列的查找算法如下:

输入: 查询时间序列 Q; 时间序列 S; 相似阈值累积误差 ε

输出: 在 S 中找到与 Q 相似的子序列集合 R.

算法描述:

(1) 范化预处理时间序列 Q、S, 按上文中算法各自检索关键点, 求出关键点集合 mp、cp, 记录 Q 关键分割点的个数 n, Q、S 将各自分段拟合成线性方程.

(2) 分别求出 Q、S 各分段线性拟合方程中的斜率 b_i, b_i 的集合, 并转换成直线倾斜的角度的 q_i、s_i 的集合. 记录 Q 段中最左、最右段拟合直线倾斜角度 q_l、q_r.

(3) 由给定相似阈值累积误差 ε, 计算每段 $\varepsilon_i = \varepsilon/(n-1)$, 近似折算成直线倾斜角度允许误差. 并根据倾斜角度的分布, 估计出倾斜角度分组数 m.

(4) 建立三行 m 列二维表格, 将 q_i、s_i 倾斜角度依次置于表格中, 在直线倾斜角度允许误差范围内, 调整 s_i 在表格中的分布.

(5) 在 s_i 中找出与最左 (L), 最右 (R) 拟合直线的倾斜角度 q_l、q_r 在同一列的记录, 保留完整连续 ($n-1$个) 段的记录, 在保留的记录中依次检索余下对应的连续列, 继续剪去与 q_i 不在对应列的记录, 最终保留的记录为与 Q 相似子序列.

同样以上文中用电负荷的线性拟合 S 为研究对象, 使用形态相似性搜索算法, 查找与 Q 相似的子序列. 将 Q 作同样的分段线性拟合, S 中每段拟合直线的倾斜角度集合为 S', 查询时间序列 Q 的直线倾斜角度集合为 Q', 设允许倾斜角度误差 $\pm 5°$.

$S' = \{s_1: -81, s_2: 78, s_3: 0, s_4: -82, s_5: 85, s_6: -84, s_7: 0, s_8: -72, s_9: 86, s_{10}: 0, s_{11}: -83, s_{12}: 85, s_{13}: -85, s_{14}: 86, s_{15}: 0, s_{16}: -84, s_{17}: 89, s_{18}: -82, s_{19}: 0, s_{20}: 86, s_{21}: -87, s_{22}: 0, s_{23}: -78, s_{24}: 70\}$

$Q' = \{q_l: 81, 0, q_r: -79\}$

由分段的直线倾斜角度数据创建表 1.

表 1 时间序列 S和 Q按本文算法分类

-8	-7	0	7	8
s_1 s_4 s_6 s_{11} s_{13}	s_{23} s_8	s_3 s_7 s_{10}	s_2 s_{24}	s_5 s_9 s_{12} s_{14} s_{17}
s_{16} s_{18} s_{21}		s_{15} s_{19} s_{22}		s_{20} s_{26} s_{28}
	q_3	q_2		q_1

根据允许误差范围, 调整 S' 在 Q' 中的列分布: q_1: (76 ~ 86), q_2: (-5 ~ 5), q_3: (-74 ~ -84), 创建表 2

表 2 时间序列 S和 Q按允许误差重新分类

q_1: (76 ~ 86) L	q_2: (-5 ~ 5)	q_3: (-74 ~ -84) R
s_2 s_5 s_9 s_{12} s_{14} s_{20} s_{26}	s_3 s_7 s_{10} s_{15} s_{19} s_{22}	s_1 s_4 s_6 s_{11} s_{16} s_{18} s_{23}

扫描表 2, 将 L列中每个 s_i 子段的 i序号后加上Q的分段数减去 1, 如结果序号能在 R列中出现, 则保留该项记录 (如: s_2的序号加 3减去 1为 s_4, 保留 s_2、s_4), 这一项操作可剪去大量不相似的记录, 保留的记录中检索余下子段是否在对应的连续列中出现 (本例表 2中 q_2列, 继续检查 s_3是否在 q_2列中), 在 S 中快速查找到与 Q 相似的子序列, 最后 S 与 Q 相似的时间子序列为 (图 5中标记) $\{s_2\ s_3\ s_4\}$、$\{s_9\ s_{10}\ s_{11}\}$、$\{s_{14}\ s_{15}\ s_{16}\}$ 3段子序列 (如图 6椭圆标出所示).

3 结论

时间序列相似检索挖掘时, 分别建立时间序列变化关键点的集合, 以关键点为边界分段, 线性拟合各分段. 以函数的斜率为相似度量标准, 建立倾斜角度分布表, 采用快速比较检索算法, 剪去非相似子序列, 能够实现时间序列中相似子序列的查找.

[参考文献]

[1] Das G, LinK Mannila H, Renganathan G, *et al*. Rule Discovery From Time Series, Fourth Annual Conference on Knowledge Discovery and Data Mining[C]. AAAI Press, 1998. 16-22.

[2] 李斌, 谭立湘. 面向数据挖掘的时间序列符号化方法研究[J]. 电路与系统学报, 2000, 4(5): 9-14.

[3] 李爱国, 覃征. 在线分割时间序列数据[J]. 软件学报, 2004, 15(11): 1671-1679.

[4] 郑诠, 朱明, 王俊普, 等. 相似时间序列的快速检索算法[J]. 小型微型计算机系统, 2004, 25(5): 785-789.

[5] 王达, 荣冈. 时间序列的模式距离[J]. 浙江大学学报(工学版), 2004, 38(7): 795-798.

[责任编辑: 刘健]