维普资讯 http://www.cqvip.com

基于时间序列的相似子模式发现算法

张 军[1,2],马志民[1]
(1.东南大学 计算机系,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:基于时间序列的数据挖掘时,一般需要对时间序列离散化,再聚类成不同的子模式。已有的方法常忽略时间序列本身的位置和整体特征,并且计算量大。针对其不足,文中提出一种检索时间序列分段关键点的算法,以关键点为边界分段,使用形态距离测度和快速剪除的算法,高效简便地检索出相似子模式。

关键词:时间序列;数据挖掘;相似子模式;形态距离
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1005-3751(2006)01-0140-03

An Algorithm of Finding Similar Subpattern Based on Time Series

ZHANG Jun[1,2], MA Zhi-min[1]
(1. Department of Computer Science & Engineering, Southeast University, Nanjing 210096, China;
2. Department of Information Technology, Jiangsu Maritime Institute, Nanjing 211170, China)

Abstract: Based on the data mining of the time series, generally discretizes time series, then clusters different sub-pattern. Some methods constantly ignore time series of position and dynamic attribute furthermore calculation large. For its shortage, presents searching time series key point algorithm, with the key point for boundary, using the appearance distance method, measures each cent a subsequence, mining latency information from the new angle of view.

Key words: time series; data mining; similar subsequence; pattern distance

1 时间序列简介

时间序列是一种在金融、气象等许多领域都普遍存在的重要数据。时间序列相似性模式发现是时间序列数据挖掘中的一个重要方面。但连续数值形式表示的结构模式不便于挖掘过程的描述和计算,为此,一般需将连续数值形式表示的时间序列转换成离散的、相对抽象的符号序列。

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。

文中提出一种检索关键点的算法,给定时间间距阈值与幅值比阈值,通过分段函数拟合、最大似然函数和最小二乘法算法校验分段线性拟合程度,确定数据挖掘应用中关键的变化点。然后以检索出的点为边界划分成各个子序列,结合时间特征的形态相似性测度计算,抽取出相似模式。

2 检索关键点的算法

时间序列 $S=(s_1,s_2,\cdots,s_i,\cdots,s_n)$,其中 $s_i=(x_i,$

收稿日期:2005-04-12
作者简介:张 军(1973—),男,江苏海安人,讲师,硕士研究生,主要从事数据挖掘方面的研究。

维普资讯 http://www.cqvip.com

$y_i)$,x_i 为时间刻度值,y_i 为 x_i 时刻的幅度值,i 为均匀采样时刻点。假设 X 能用分段函数模型[3] 拟合,即:

$$X=\begin{cases} f_1(t,w_1)+e_1(t), & 1\leqslant t<\alpha_1 \\ \quad M \\ f_k(t,w_k)+e_k(t), & \alpha_{k-1}\leqslant t<\alpha_k=N \end{cases}$$

式中向量 $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_k)$,是 X 的时间序列关键分段点的集合;$e_1(t),e_2(t),\cdots,e_k(t)$ 是第 i 段的误差项,$e_i(t)$ 均值为零的高斯白色噪声函数;$f_i(t,w_i)$ 为时间序列第 i 段的拟合多项式函数,$1\leqslant i\leqslant k$,w_i 是系数向量,$f_i(t,w_i)\in M$,M 可以为多种形式,如线性多项式、DFT、DWT 等模型,通常形式为多项式函数 $f(t)=w_0+w_1t+w_2t^2+\cdots+w_pt^p$,$p$ 为阶次,文中取 $p=1$。

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,如果检索关键点无限多,G 值近似为零,就失去挖掘的意义。最优方法是给定 G 值情况下,希望分段总数最少。

为此,提出如下方法:给定幅值比例阈值 $\{\delta_1(0\leqslant\delta_1\leqslant1);\delta_2(1<\delta_2\leqslant\infty);$ 且 $\delta_1=1/\delta_2\}$,时间间距阈值 T,先扫描时间序列数据库,以时间先后为序,依次求出后面时刻的幅值与起点 t_0 时刻的幅值 A_0 比例值(且记录比例值和此段平均值),同时求得时间间距 $t-t_0$。如满足比例值 $\leqslant\delta_1$ 或比例值 $>\delta_2$ 条件时,则记录此时刻 t_i、幅值 A_i;如一直不满足条件,但时间间距超过 T,同样记录此时刻 t_i 和幅值 A_i(且记录比例值和此段平均值),调整 A_0 为 t_i 时刻的 A_i 值,t_0 为 t_i,记录每一起点(关键点)的时刻 t_i 和幅值 A_i。以 t_i 为边界将连续时间序列分成若干子序列,可根据拟合度参数需要,多次人工交互调整 δ_1,δ_2,T 的阈值。

算法 1 如下:

输入:时间序列 X;给定时间间距阈值 T;幅度比例阈值 δ_1,δ_2;采样结束时刻 t_e。

输出:关键点集合 cp。

算法:

```
FST:在每个采样点时刻;do
    if A/A0>δ2 OR A/A0≤δ1
    then go JILU;
    else if t-t0≥T
    then go JILU;
    go CON
JILU:A0=A:t0=t(且记录 t-t0 段幅值平均值,t 时刻的比例值)
        cp=cp∪t;
CON:if t≤te goto FST
    end if
  end if
end do
```

以 cp 集合中每一点为分界点,分时间序列为各子序列,对其作一元线性回归拟合。拟合程度指标 r 计算如下:

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

ε 为随机误差。

用最小二乘法求 a,b 的最大似然估计:

1) 作似然函数。

因为 $y_i\sim N(a+bx_i,\sigma^2)(i=1,2,\cdots,n)$,所以 y_i 的分布密度为:

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

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) 求平方和函数。

令 $Q(a,b)=\sum_{i=1}^{n}(y_i-a-bx_i)^2$,求其最小点。

显然,要使 L 最大,只要使 $Q(a,b)$ 最小,用最小二乘法计算 a,b。具体方法如下[4]:

先由二元函数的极值性质建立正规方程组:

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

整理得:

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

式中 $\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n}$,$\bar{y}=\frac{\sum_{i=1}^{n}y_i}{n}$,即各子序列中每个采样点幅度的平均值。

将 $\hat{a},\hat{b}$ 代入得到经验回归方程:

$$y=\hat{a}+\hat{b}x$$

现验证回归直线 y 与实际数据的拟合程度,提出一拟合程度指标:

$$r=\frac{L_{xy}}{\sqrt{L_{xx}}\cdot\sqrt{L_{yy}}}$$

其中:

$$L_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$L_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$L_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2$$

如:$r=0$,则 $L_{xy}=0$,因而 $\hat{b}=0$,x,y 无线性关系,需重新调整 δ,T,检索关键点;$0<|r|<1$,则 x,y 存在一定线性关系,可人为指定 r 值,根据需要重新调整 δ,T,再重

维普资讯 http://www.cqvip.com

新检索关键点;设置合适的 δ,T,检索的关键点通常也是感兴趣的点。$|r|=1$,x,y 完全线性相关。基于线性回归技术和最小均方差意义下的时间序列分割,可实现低维数高精度的特征空间重构。

3 模式相似性计算

由上述算法得到时间序列关键分割点的集合 cp,每两个关键点之间构成一组形态各异的有限长分段。在查询相似模式子序列时,常采用计算序列在多维相空间中点距离,如欧氏距离公式。

实际的采样时间和记录时间有一定的差距,可以把采样获得的(Value,Time) 表示为(Value,Time + ζ,ζ 为噪音),基于点距离的计算,结果对时间噪音十分敏感。

分段序列最基本的计算单元是分段而不是一维数据点,采用基于点距离方法显然不能完全描述分段之间的形态相似性。为此提出一个新的度量,称为线性分段子序列形态相似性测度。根据算法 1,提出算法 2 检索相似子模式,以各自分段点比例值之差作为形态相似距离的测度。

设有子序列 Q,在时间序列 S(S 的长度远大于 Q 的长度) 中查询与之相似的子序列集。为此提出前缀比较算法2,将 Q 用算法1及 S 分段时相同的阈值分割,可根据需要结合时间跨度 T,每段的平均值等特征值来进一步抽取相似模式。

算法 2 如下:

输入:时间序列 S,关键点集合 cp;子序列 Q,相似度阈值 ε。

输出:与 Q 相似的子序列分段时刻点集合 mp。

算法:

(1) 用算法 1 将 Q 分成 n 个子段($q_1,q_2,\cdots,q_i,\cdots,q_n$);

(2) 计算出 n 段的比例值总和 K;

(3) 在 S 中检索与 q_1 相似(s_i,q_i 各自比例值的差 ε 为标准) 的子段时刻 t 的集合 mp;

(4) 依次以 t 为起点,计算各 n 个分段比例值之和,在 mp 中舍去与 K 值不同的 t;

(5) 继续检索与 q_2 相似的子段(q_1 为前缀),在 mp 中舍去不相似的 t;

(6) 继续检索 Q 的余下的子序列,在 mp 中舍去不相似的 t;

(7) 最后在 mp 中,如有时间跨度 T 和幅度均值 A' 的相似要求,继续舍去不符合相似 T,A' 条件的 t,得到与 Q 相似的子模式集合。以 q_1 为前缀搜索,比例值总和 K 为剪除条件,这是快速有效的剪除非相似 t 时刻点方法,为继续检索减少了很多的计算量。

4 实验与评估

使用 Wiener 过程来产生实验数据。Wiener 过程是一类描述实际物理世界的随机现象的有效逼近,可以用来模拟股票市场的价格波动等。同样可以利用其过程产生模拟的时间序列,序列根据下面的递推式获得 RAWK 数据库[5]:

$$x_w(k+1)=x_w(k)+\varepsilon,1\leqslant k\leqslant 6399$$

$$x_w(1)=\varepsilon_0$$

ε_0 满足在[0,1]上平均分布,而 ε 满足[-0.05,0.05]上的平均分布。

实际可对上式得到的序列进行归一化处理,得下式:

$$x(k)=\frac{x(k)-\min\limits_{1\leqslant i\leqslant 6400}\{x(i)\}}{\max\limits_{1\leqslant i\leqslant 6400}\{x(i)\}-\max\limits_{1\leqslant i\leqslant 6400}\{x(i)\}},1\leqslant k\leqslant 6400$$

如:取时间序列数据长度 $n=200000$,子序列的长度 $m=2000$,使用 Wiener 过程随机产生 100 个长度为 n 的时间序列,在每一个时间序列中随机抽取一长度为 m 的子序列。做相似模式查询实验,实验过程如图 1 所示。

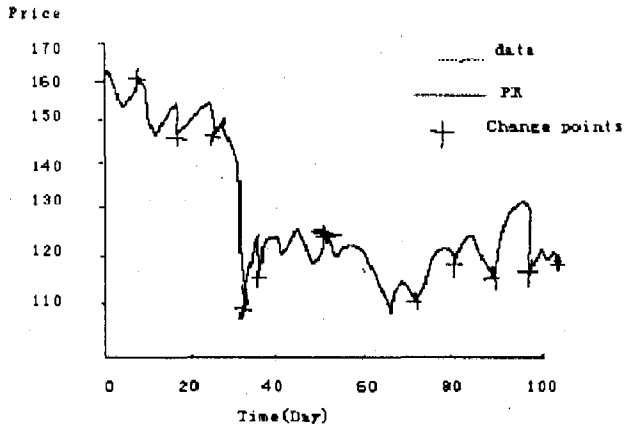

图 1 文中算法对时间序列作关键点分段的结果

采用文中的检索关键点算法,M 依次取线性模式和二阶模式($p=1$,$p=2$),p 依次取 1 或 2,反复实验,得到如图 1 的变化点,从图中(M 取线性模式,p 取 2) 可以看出,关键点往往也是极值点(兴趣点),共计检索出 24 个点,误差值 $G(x)=45$;用 SW 滑动窗口算法作为对比,如图 2 所示,原始时间序列数据分为两个阶段的关键点,SW 没有发现,共计检索出 35 个点,误差值 $G(x)=523$。

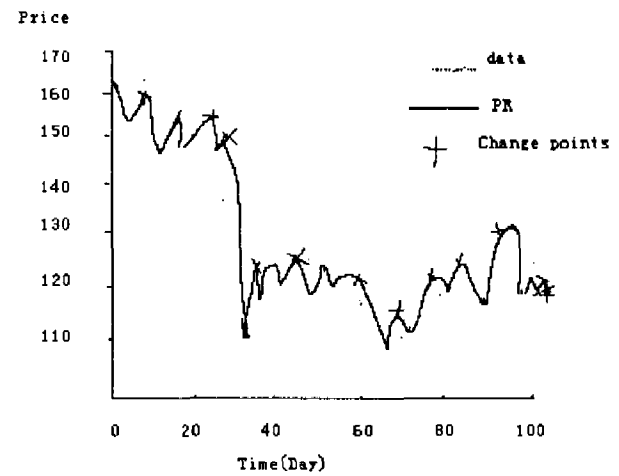

图 2 SW 滑动窗口算法对时间序列分段的结果

(下转第 146 页)

维普资讯 http://www.cqvip.com

协议翻译和协议中继来实现，随着 IPv4 地址的耗竭这种服务将会逐渐增多。

以上从前提到用例，给出了文中所提方案的整体描述。它利用SNMP管理模式解决了对各种IPv6过渡技术工具的统一管理问题，针对过渡技术本身的复杂性提供了基于策略的灵活的管理方式，使得对 IPv6 过渡网络的管理可以与传统 IPv4 网络的管理协调起来并存在于一个统一的管理系统之中，各种过渡技术本身的管理问题也得到了解决。某高校采用基于该方案的管理系统后所得到的一项隧道代理数据结果如图 3 所示。

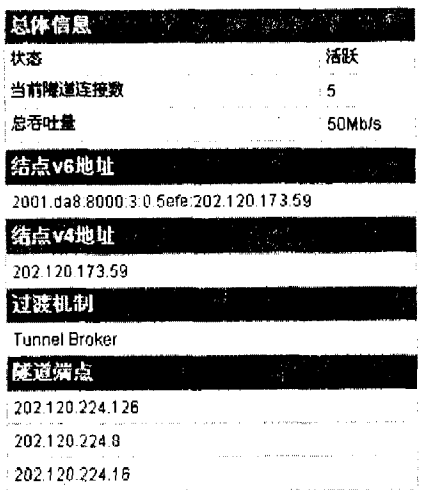

图 3 过渡方案的实施结果示例

5 总 结

对复杂的过渡技术进行统一形式的信息建模，将通用的管理模式与灵活的策略化管理相结合，采用层次化的框架来融合整个管理体系，这是一种适合 IPv6 过渡网络特点的管理方案。该方案为复杂网络的管理提供了参考，解决了过渡网络的管理问题。

参考文献：

[1] Schild C, Strauf T. Initial IPv4 to IPv6 transition cookbook for end site networks/universities[EB/OL]. http://www.6net.org/publications/deliverables/D2.3.2.pdf, 2003-02.

[2] RFC 1157. The Simple Network Management Protocol[S]. IETF, 1990.

[3] Rajan R. A Policy Framework for Integrated and Differentiated Services in the Internet[J]. IEEE Network, 1999, 13(5): 36-41.

[4] RFC 2748. The COPS (Common Open Policy Service) Protocol[S]. IETF, 2000.

[5] RFC 3512. Configuring Networks and Devices with SNMP [S]. IETF, 2003.

[6] Baudot A, Egeland G, Hahn C, et al. Interaction of transition mechanisms [EB/OL]. http://www.atm.tut.fi/list-archive/ietf-announce/msg10435.html, 2004-11.

(上接第 142 页)

在每一个时间序列中随机抽取一长度为 m 的子序列，通过形态距离公式，计算与各分段的形态距离，找出最相似的子序列。通过文中形态相似性算法，抽取一段子序列分析，如图 3 所示，如采用欧氏距离计算，差距值很大，

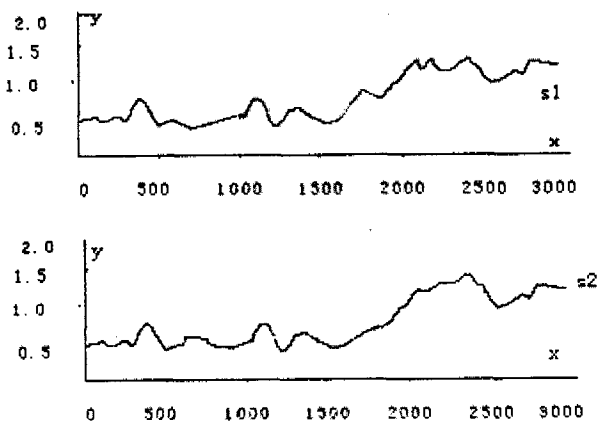

图 3 形态距离计算方法下两相似序列

s_1, s_2 不是两相似序列。但用文中形态距离计算方法，s_1, s_2 是相似模式，从图 3 中可以直观看出，相互间有 80% 的相似性，可以划分为一类。当然在不同的分段数目下，实验数据间的模式距离值也不一样，但此时它们的模式很接近，相当于对应不同的“分辨率”，时间序列模式的相似程度会不一样，此特性为数据的挖掘应用提供新角度信息，更有利于潜在信息的知识发现。

5 总 结

时间序列挖掘研究和应用中的主要任务就是时间序列变化关键点(兴趣点)的检索。文中提出在给定拟合总误差条件下，建立产生关键点的集合且使分段总数最少。用拟合程度指标交互修正分段计算时的阈值。在此基础上，提出形态相似的测度计算方法，在检索过程中，应用比例值总和相似条件的剪除计算，快速有效地抽取相似子序列的分段点。

参考文献：

[1] Das G, Lin K-I, Marmila H, et al. Rule Discovery From Time Series[A]. In Proc. of the 4th Int. Conf. on Knowledge Discovery and Data Mining[C]. [s.l.]: 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. 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] 郭斯羽. 动态数据中的挖掘研究[D]. 杭州：浙江大学，2002. 14-17.