计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
49(12): 2645-2655, 2012

一种新的时间序列延迟相关性分析算法——三点预测探查法

林子雨[1] 江 弋[1] 赖永炫[2] 林 琛[1]
[1](厦门大学计算机科学系 福建厦门 361005)
[2](厦门大学软件学院 福建厦门 361005)
(ziyulin@xmu.edu.cn)

A New Algorithm on Lagged Correlation Analysis Between Time Series: TPFP

Lin Ziyu[1], Jiang Yi[1], Lai Yongxuan[2], and Lin Chen[1]
[1](*Department of Computer Science, Xiamen University, Xiamen, Fujian* 361005)
[2](*School of Software, Xiamen University, Xiamen, Fujian* 361005)

Abstract Lagged correlation analysis plays an important role in data mining based on time series, which can be used extensively in real life such as weather forcast, stock market analysis, network analysis, moving object tracking, sensor monitoring, and so on. Lagged correlation analysis is to find out the time series which are correlated with lags. Here three phenomena in lagged correlation analysis between time series, namely, continuous distribution of lags, lag mutation and mutation amplitude distribution feature, are found based on extensive experiments. It is proved that existing research work can achieve desirable performance when the lag is small, but there is sometimes large error when the lag is large. Furthermore, available methods can not deal with the occasion of lag mutation, which means that the lag changes suddenly at certain point on the lag correlation curve and is much different from before. This brings many difficulties for those existing methods. Based on the above three phenomena, three points forecast-based probing (TPFP) is proposed here to overcome the disadvantages of the existing methods, which is able to achieve small error when the lag is large, and also it can perform well on the occasion of lag mutation. Extensive experiments show that TPFP can achieve better performance than available methods.

Key words time series; lagged correlation; correlation analysis; lag mutation; three points forecast-based probing (TPFP)

摘 要 延迟相关性分析是时间序列数据挖掘的重要研究内容,它可以在很多领域得到应用,比如股票市场分析、天气预报、网络分析、移动对象跟踪和传感器监控等;通过实验发现和验证了时间序列延迟相关性分析中存在的3个现象,即连续分布性、延迟突变和突变幅度分布特性;证明了已有研究或者在延迟位置较大时具有较大的误差,或者无法解决延迟突变问题;根据3个实验现象,提出了三点预测探查法(three points forecast-based probing, TPFP),它可以克服已有算法的缺陷,在延迟位置较大时也可以具有较小的误差,并且可以有效处理大部分延迟突变情形.大量实验证明,三点预测探查法可以比已有方法取得更好的性能.

关键词 时间序列;延迟相关;相关性分析;延迟突变;三点预测探查法

中图法分类号 TP311

收稿日期:2010-11-09;**修回日期**:2011-06-24

基金项目:国家自然科学基金项目(61001013,61001143,61102136);福建省自然科学基金项目(2011J05156,2011J05158);厦门大学基础创新科研基金项目(中央高校基本科研业务费专项资金项目)(2011121049)

时间序列的延迟相关性是时间序列数据挖掘[1]领域的一个重要研究问题. 简单地说,对于两个时间序列 $\boldsymbol{X}=\{x_i \mid i=0,\cdots,n-1\}$ 和 $\boldsymbol{Y}=\{y_j \mid j=0,\cdots,n-1\}$,二者的延迟相关是指 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的相关性系数最大值并不发生在 $i=0$ 的位置,而是 $i=l(l\neq 0)$ 的位置,l 就是延迟的大小. 在实际应用中,有很多情形涉及到延迟相关问题:

1) 股市分析. 在股票市场上存在着多只股票,一种股票(尤其是权重股)的价格走势常常会影响到其他股票的行情,但是,这种相关性往往不会立即表现出来,可能会存在一个延迟.

2) 气候预测. 大气环流使得发生在一个地方的气候情况会在不远的将来给另一个地方带来影响.

目前已有一些研究时间序列延迟相关性的成果[2-4],但是,本文通过大量实验发现,这些方法或者在延迟位置较大时具有较大的误差,或者无法解决延迟突变问题. 所谓延迟突变,是指对于一些相邻的时间点,最大延迟相关点位置分布范围会发生突变. 除了延迟突变,本文还发现另外两个重要现象:首先,对于很多相邻的时间点,最大延迟相关点总是连续集中地分布在一个较小的范围内;其次,对于相邻的时间点,绝大多数最大延迟相关点位置突变的幅度位于区间$[0,m/2]$之内,其中 m 表示允许的最大延迟.

基于以上发现,本文提出了三点预测探查法(three points forecast-based probing, TPFP)方法,它可以克服已有算法的缺陷,在延迟位置较大时也可以具有较小的误差,并且在大部分延迟突变情形下也可以取得较好的性能;大量的实验证明,本文的方法可以比已有方法取得更好的性能.

1 问题描述

定义 1. 时间序列延迟相关[2]. 对于两个时间序列 $\boldsymbol{X}=\{x_1,\cdots,x_n\}$ 和 $\boldsymbol{Y}=\{y_1,\cdots,y_n\}$,二者在延迟 l 处($\boldsymbol{Y}$ 延迟于 $\boldsymbol{X}$)的相关系数 $R(l)$ 计算公式如式(1)所示:

$$R(l)=\frac{\sum_{t=l+1}^{n}(x_t-\bar{x})(y_{t-l}-\bar{y})}{\sqrt{\sum_{t=l+1}^{n}(x_t-\bar{x})^2}\sqrt{\sum_{t=1}^{n-l}(y_t-\bar{y})^2}},$$
$$\bar{x}=\frac{1}{n-l}\sum_{t=l+1}^{n}x_t,\bar{y}=\frac{1}{n-l}\sum_{t=1}^{n-l}y_t, \quad (1)$$

其中,l 的最大取值为 $n/2$. 当 l 的值从 0 变化到 $n/2$ 时就可以得到多个 $R(l)$ 值,令 $R(l)$ 最大值所对应的延迟为 $l_{\max}$,如果 $R(l_{\max})>\sigma$(一个人为定义的阈值),那么就说时间序列 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 具有延迟相关性,$\boldsymbol{Y}$ 延迟于 $\boldsymbol{X}$ 的量为 $l_{\max}$. 以 l 为横坐标、$R(l)$ 为纵坐标的曲线被称为"延迟相关曲线",该曲线上 $R(l)$ 最大值所在的点被称为"最大延迟相关点".

这里以一个实例来解释延迟相关问题. 如图 1(a)所示,有两个时间序列 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$,长度均为 300,图 1 中横坐标为时间点 t,纵坐标为时间序列在时间点 t 处的值 V;图 1(b)给出了它们的延迟相关曲线,横坐标为不同的延迟 Lag,纵坐标为在当前 Lag 值下的相关系数. 对于特定的 Lag 值,利用式(1)计算 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的相关系数时,实际上参与计算的只是 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 二者的等长部分. 比如,当延迟 $Lag=0$ 时,$\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的等长部分是长度为 300 的整段时间序列,即用 $\{x_1,x_2,x_3,\cdots,x_{300}\}$ 和 $\{y_1,y_2,y_3,\cdots,y_{300}\}$ 计算相关系数;而当 $Lag=100$ 时,$\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的等长部分是长度为 200 的子序列 $\{x_{101},x_{102},x_{103},\cdots,x_{300}\}$ 和 $\{y_1,y_2,y_3,\cdots,y_{200}\}$. 由此可以看出,随着 Lag 值的增加,$\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的等长部分的长度会逐渐减小,但是,当参与计算的等长部分的长度太小时,得到的相关系数就不具有实际应用价值. 因此,目前的研究[2,4]大都把 Lag 的最大值限定为时间序列 $\boldsymbol{X}$ 长度的一半(即 $n/2$). 从图 1(b)中可以看出,当 $Lag=14$ 时,$\boldsymbol{X}$ 和 $\boldsymbol{Y}$

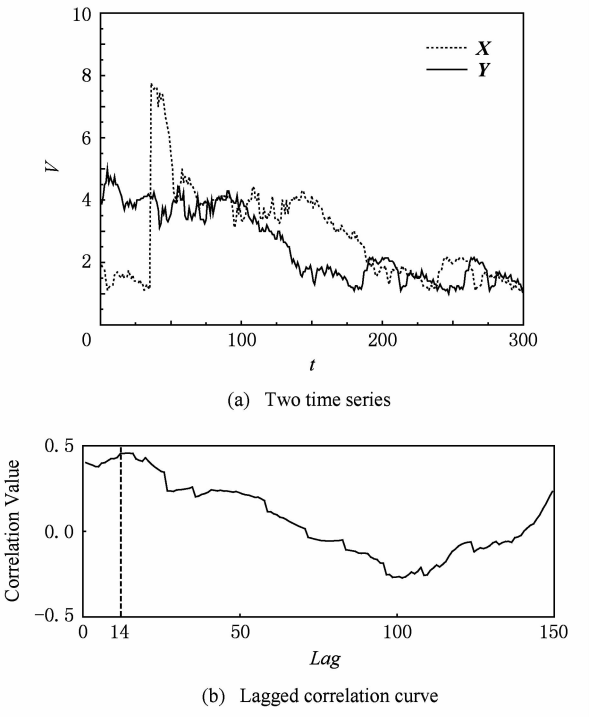


Fig. 1 Two time series and their lagged correlation curve.

图 1 2 个时间序列及其延迟相关曲线

的相关系数取得最大值 0.458,该 Lag 值所对应的延迟相关曲线上面的点就是最大延迟相关点.如果让 $\sigma=0.4$,就可以认为时间序列 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 具有延迟相关性,$\boldsymbol{Y}$ 延迟于 $\boldsymbol{X}$ 的量为 14.

图 1 中的时间序列延迟相关曲线是通过一种简单的"蛮力"算法得到的,即让 Lag 取值从 0 依次变化到 150,分别计算每个 Lag 值所对应的相关系数,然后求出相关系数最大值.如果时间序列长度为 n,那么,这种算法求解延迟相关问题时就要进行 $O(n)$ 次相关系数计算(采用式(1)计算相关系数).但是,已有研究[2]表明,没有必要对每个 Lag 值计算相关系数,可以采用"几何渐进探查"的方式对少数 Lag 值计算相关系数,然后,采用数学插值方法得到其他 Lag 值所对应的相关系数值,从而最终求得相关系数最大值,这时,只需要进行 $O(\log n)$ 次相关系数计算.从式(1)可以看出,每次相关系数计算都需要大量时间,因此,减少相关系数的计算次数,可以大大减少延迟相关问题的整体求解时间.基于以上论述,下面简要描述本文所要解决的问题.

问题描述:对于长度为 n 的时间序列 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$,通过 $O(\log n)$ 次相关系数计算和数学插值方法得到二者的延迟相关曲线,从而求出最大相关系数所对应的 Lag 值,即找到最大延迟相关点的位置.

2 BRAID 方法及其不足

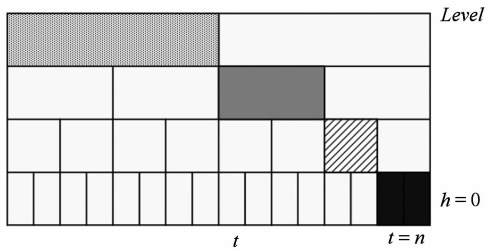
本节将首先简单介绍已有的时间序列延迟相关分析方法——BRAID,然后指出其不足之处.

2.1 BRAID 方法介绍

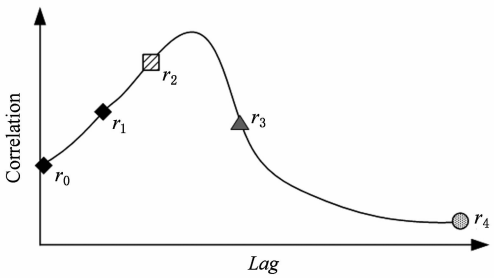
Sakurai 等人[2]最早出了 BRAID 方法,它采用了 2 种策略加速求解过程:几何渐进探查和近似平滑序列.

1) 几何渐进探查.在求两段时间序列的延迟相关曲线时,它并没有去计算所有的点,而是采用"几何渐进"的方式对延迟相关曲线上的少数点进行探查,即只对少数 Lag 值计算相关系数,然后,采用样条(spline)插值方法得到延迟相关曲线上的其他点,从而最终求得曲线上相关系数最大值所在的点,即最大延迟相关点.采用"几何渐进"方式探查时,Lag 的取值为 0 和 2^i,即 $0,2^0,2^1,2^2,2^3,\cdots,2^{\lfloor\log m\rfloor}$,其中,$m$ 为允许的最大延迟量.比如,在图 2 中,r_0 对应的 Lag 值为 0,r_1 对应的 Lag 值为 1,r_2 为 2,r_3 为 4,r_4 为 8.

2) 近似平滑序列.在每次针对具体 Lag 值计算相关系数时,并不采用原来的时间序列,而是采用近似的、平滑后的时间序列.具体而言,就是在原来的时间序列上放置一个大小为 p 的滑动窗口,每次移动一步(相邻两步之间的两个窗口不能重叠),并计算窗口内子序列的平均值,这些平均值就构成一个新的时间序列,称为"移动均值"[5],它是原来时间序列的近似表示.使用大小为 p 的滑动窗口得到的均值序列称为"p-阶移动均值".显然,p 值越大均值序列长度越小.比如,对于长度为 n 的时间序列,它的 2-阶移动均值序列长度为 $n/2$.在 BRAID 方法中,作者把窗口大小 p 设置为 2^h,并把不同窗口大小时的"p-阶移动均值"组织成一个层次结构(如图 2 所示),在 $h=0$ 层,窗口大小为 1,该层对应的是 1-阶移动均值,依次类推.在计算不同 Lag 值所对应的相关系数值时,Lag 越大计算相关系数所用近似序列的层次越高.比如,在图 2 中,延迟相关曲线上有 5 个点:r_0,r_1,r_2,r_3,r_4,其中 r_0 和 r_1 在计算时采用 1-阶移动均值,r_2 采用 2-阶移动均值,r_3 采用 4-阶移动均值,r_4 采用 8-阶移动均值.显然,移动均值的阶数越高,采用式(1)计算相关系数所涉及的计算量就越小.BRAID 方法同时采用了增量更新的方法,可以快速对各阶移动均值进行动态更新,从而可以在数据流环境中使用.



(a) The hierachy of moving average



(b) The lagged correlation curve

Fig. 2 BRAID method.

图 2 BRAID 方法示意图

2.2 BRAID 方法的不足

BRAID 采用几何渐进方式进行探查,每次探查时都是从延迟为 0 的点开始探查,即 $Lag=0,1,2,$

4,8,…. 对于这种探查方式,当最大延迟相关点具有比较小的 Lag 值时,BRAID 具有很高准确性. 但是,当最大延迟相关点具有较大的 Lag 值时,该方法准确性不高. 这是因为 BRAID 在 Lag 值较小的范围内设置的探查点比较密集,随着 Lag 值的增大探查点越来越稀疏. 这样,该方法对延迟相关曲线上面那些 Lag 值较小的点的探查比较密集,能够准确地反映出这部分曲线的本来面貌. 但是,对于延迟相关曲线上面那些 Lag 值较大的点的探查非常稀疏,用这些稀疏的探查点来插值还原这部分曲线的本来面貌,准确度就会大大降低. 为了说明这个问题,下面给出两个实例.

1) 最大延迟相关点在 Lag 值较小处的情形. 例如,图 1 中的两个时间序列的延迟相关点就出现在 Lag 值较小($Lag=14$)的地方,对于这种情形,采用 BRAID 方法就可以取得比较好的效果. 如图 3 所示,粗糙的曲线(naive)表示采用蛮力算法计算得到的延迟相关曲线;Probing Point 表示的 8 个点是 BRAID 方法设置的 8 个探查点,这些探查点在 Lag 值较小处分布比较密集,在 Lag 值较大处分布很稀疏;比较平滑的曲线(BRAID)表示根据 8 个探查点采用样条插值方法得到的近似延迟相关曲线. 从图 3 可以看出,采用蛮力算法和 BRAID 方法得到的最大延迟相关点是相同的,都在 $Lag=14$ 这个位置. 因此,在这种情形下,BRAID 方法是准确的.

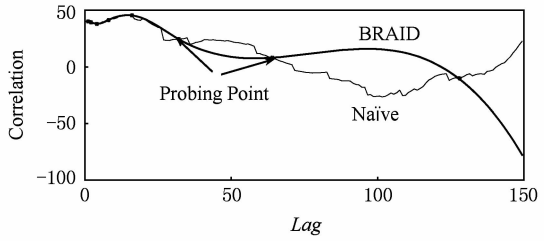


Fig. 3 When the max lagged correlation location is small.

图 3 最大延迟相关点较小的情形

2) 最大延迟相关点在 Lag 值较大处的情形. 如图 4 所示,采用蛮力算法得到的最大延迟相关点在 $Lag=140$ 处,而采用 BRAID 方法得到的最大延迟相关点则在 $Lag=15$ 处,因此,BRAID 方法计算结果的误差很大.

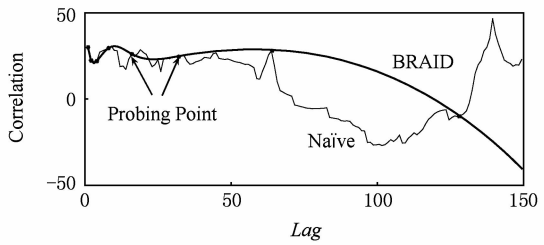


Fig. 4 When the max lagged correlation location is large.

图 4 最大延迟相关点较大的情形

3 TPFP 方法

针对 BRAID 方法的不足,本文提出了三点预测探查法. 本部分内容在简述 TPFP 方法之后,将介绍本文发现的 3 个实验现象,然后,详细阐述基于这些现象而提出的 TPFP 方法.

3.1 TPFP 方法描述

与 BRAID 方法不同,TPFP 方法在延迟相关曲线上设置探查点时,不是从 Lag 值为 0 的地方开始探查,而是在最可能出现最大延迟相关点的地方设置第 1 个探查点,然后,以第 1 个探查点为基础,在该探查点左右两侧分别以"几何渐进"的方式设置其他探查点. 通过这种方式得到的探查结果,可以较准确地利用插值方法还原真实的延迟相关曲线.

由于 TPFP 方法要在最可能出现最大延迟相关点的地方开始探查,因此,TPFP 方法若要取得较好的性能,算法需要事先预测最大延迟相关点最可能出现的位置. 本文后面的内容将具体阐述如何解决这个问题,但是,在给出具体解决方案之前,这里将先介绍 3 个实验现象,本文的提出 TPFP 方法正是以这 3 个现象为基础的.

3.2 实验现象

1) 实验现象 1. 对于很多相邻的时间点,最大延迟相关点总是连续集中地分布在一个较小的范围内.

通过大量基于股票市场真实数据的实验分析,本文发现了最大延迟相关点具有平稳分布的特性. 图 5 显示了两个数据流时间序列(长度为 300)的最大延迟相关点的位置变化曲线. 图 5 中的横坐标 t 表示时间点,从 0 变化到 900;纵坐标 $MaxLag$ 表示在某个时间点 t 上时间序列 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的最大延迟相关点的 Lag 值. 从图 5 可以观察到存在很多用横向椭圆标注的区域(图 5 中只标出了一部分作为示例),这些区域被称为"稳定分布区域". 在这些区域中,最大延迟相关点位置具有连续稳定的区间分布,没有发生大的变化. 比如,当 t 为 401,402,403,404,405,406,407,408,409,410 时,$MaxLag$ 分别为 87,83,83,83,83,83,87,83,83,87,数值分布非常稳定. 图 6 是采用另一组时间序列统计得到的延迟突变幅度 f 分布饼状图,从中可以看出,突变幅度 f 为 0 的情形占 32%,幅度 f 在(0,5]之间的情形占 52%,二者之和占 84%. 由此可见,最大延迟相关点的位置在大多数情形下具有连续分布的特性.

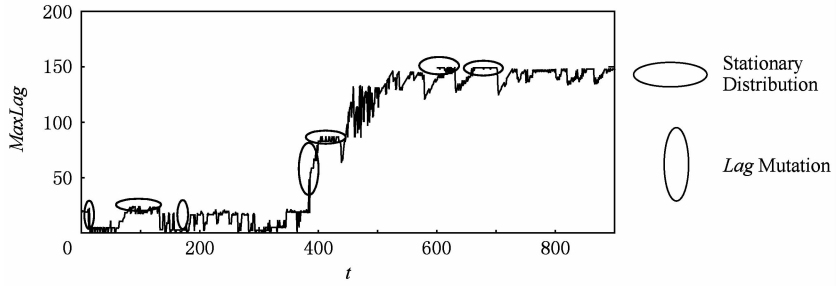


Fig. 5 The stationary distribution of max lagged correlation location and lag mutation.

图 5 最大延迟相关点位置的平稳分布和延迟突变

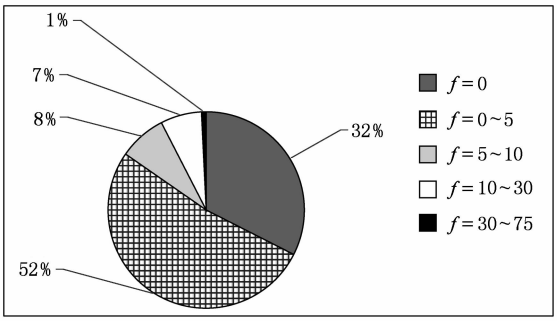


Fig. 6 Range distribution of lag mutation amplitude.

图 6 延迟突变幅度区间分布

针对这些稳定分布区域,由于最大延迟相关点具有连续集中小范围分布的特性,就可以使用当前已知的最大延迟相关点位置来预测下一个时间点的最大延迟相关点位置,显然,这种方法具有很高的准确度.

2) 实验现象 2. 对于一些相邻的时间点,最大延迟相关点位置分布范围会发生突变.

本文通过大量实验也发现,在一些时间点,最大延迟相关点的位置分布特性会发生突然变化,本文把这种情况称为"延迟突变". 在图 5 中存在一些用竖向椭圆标注的区域,这些区域被称为"延迟突变区域". 比如,在 $t=381$ 处,$MaxLag$ 突然从 19 快速变化到 87. 在 $t=697$ 处,$MaxLag$ 突然从 147 变化到 125. 从图 6 也可以看出,发生突变的情形占 48%,而发生较大突变的情形则占 16%.

对于延迟突变区域,如果依然使用当前时间点的最大延迟相关点位置作为预测依据,就会产生较大的误差.

3) 实验现象 3. 对于相邻的时间点,绝大多数最大延迟相关点位置突变的幅度位于区间 $[0,m/2]$ 之内.

这里的 m 表示允许的最大延迟. 这个现象在本文已经开展的大量实验中都存在,在图 5 中就可以很容易观察到这一现象. 比如,在 $t=381$ 处,$MaxLag$ 突然从 19 快速变化到 87,变化幅度是 68,它属于区间 $[0,75]$,这一变化幅度在图 5 中已经是最大变化幅度. 当然,在实验过程中,也存在极少量位置突变幅度超出 $[0,m/2]$ 的情形,由于它出现的概率很小,在实际应用中可以作忽略处理. 图 7 是针对图 5 中的数据计算得到突变幅度后绘制的延迟突变幅度分布散点图,从中可以看出,大部分突变幅度都较小,只有少数几个位置的突变幅度大于 30,而且所有突变幅度都在 $[0,75]$ 区间内,也就是小于允许最大延迟($m=150$)的一半.

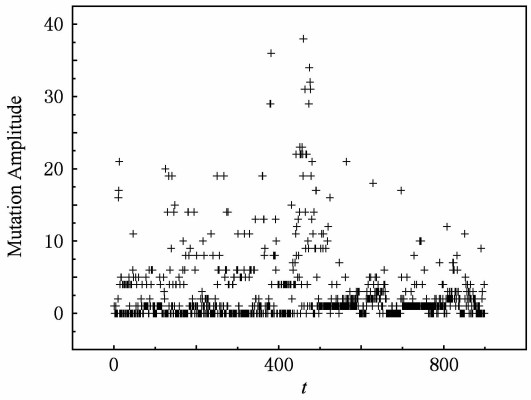


Fig. 7 Scatter diagram of lag mutation amplitude.

图 7 延迟突变幅度分布散点图

这一实验现象将为本文的最大延迟相关点位置预测提供重要依据.

3.3 TPFP 具体算法

如算法 1 所示,在 TPFP 方法中,首先需要确定 3 个预测探查点,具体方法如下:

1) 把上一次最大延迟相关点位置 $preMaxLag$ 作为第 1 个预测探查点 $probe\_one$;

2) 把第 2 个预测探查点设置为 $preMaxLag+m/4$;

3) 把第 3 个预测探查点设置为 $preMaxLag+m/2$.

第 1)步的依据是"实验现象 1",即对于很多相

邻的时间点,最大延迟相关点总是连续集中地分布在一个较小的范围内.因此,当前时间点的最大延迟相关点位置很有可能就在上一个时间点的最大延迟相关点位置附近.

第3)步的依据是"实验现象3",即对于相邻的时间点,绝大多数最大延迟相关点位置突变的幅度在$[0,m/2]$.因此,如果产生突变,新的最大延迟相关点应该在$[preMaxLag, preMaxLag+m/2]$范围内.

第2)步的依据是三次样条函数拟合曲线精确通过各采样点,在采样点处拟合误差为零,两采样点间拟合误差最大值出现在中点附近,呈现中间大、两端小的趋势.因此,在$[preMaxLag, preMaxLag+m/2]$的中点处增加一个探查点,可以在很大程度上减少插值误差.

在确定这3个预测探查点后就可以把它们作为输入参数,利用$MaxProbing()$函数求得下一步进行几何渐进探查的起始点.得到起始探查点以后就可以从该点出发,向左右两边同时进行几何渐进探查,如果碰到边缘就结束探查,最终就可以求得当前时间点的最大延迟相关点位置$curMaxLag$.

这里需要说明的是,在算法1中,假设$preMaxLag<m/2$,这样可以简化程序书写,易于理解,对于$preMaxLag>m/2$的情形,处理方法是类似的.

算法1. TPFP算法.

输入:① 时间序列$\boldsymbol{X}$和$\boldsymbol{Y}$;/* 当前时间点的两段时间序列 */

②$preMaxLag$;/* 上一个时间点的最大延迟相关点位置 */

输出:① $curMaxLag$;/* 当前时间点的最大延迟相关点位置 */

```
begin
    probe_one←preMaxLag; /* 把上一个时间点的最大延迟相关点位置作为第1个预测探查点 */
    probe_two←preMaxLag+m/4; /* 确定第2个预测探查点 */
    probe_three←preMaxLag+m/2; /* 确定第3个预测探查点 */
    probe_first←MaxProbing(probe_one, probe_two, probe_three); /* 确定几何渐进探查的起始点 */
    curMaxLag←LagProbing(probe_first, X, Y); /* 从起始探查点开始采用几何渐进方式向两边探查得到结果 */
  return curMaxLag;
end
```

3.4 起始探查点的确定

三角法:如图8所示,由3个预测探查点P_1,P_2和P_3构成一个三角形,令边P_1,P_2的延长线和边P_2,P_3构成的夹角为θ,当边P_2,P_3在边P_1,P_2的延长线的下方时,$\theta<0$,反之,$\theta\geqslant 0$.由此,可用式(2)确定起始探查点$probe_first$:

$$probe\_first=\begin{cases}P_1, \theta<0, P_1>P_2\\ P_2, \theta<0, P_1<P_2\\ P_3, \theta\geqslant 0\end{cases}. \tag{2}$$

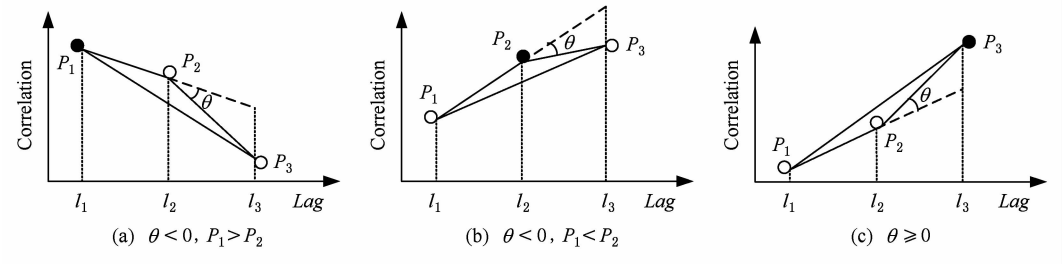


Fig. 8 Determining the first probing location in TPFP.

图8 TPFP方法中起始探查点的确定

3.5 算法分析

计算最大延迟相关点时,若采用蛮力算法,则需要进行$O(n)$次相关系数的计算,其中,n表示时间序列的长度;若采用BRAID方法,则只需要进行$O(\log n)$次相关系数计算,大大减少了计算量;而采用TPFP方法,同样只需要计算$O(\log n)$次相关系数,并没有增加计算量.在TPFP方法中采用三点预测探查法,虽然看起来比BRAID方法增加了3次相关系数的计算,但是,实际上并非如此,因为,对于这3个预测探查点,由于已经对它们进行了相关系数计算,那么,这3个点就可以直接作为后面的插值采样点,也就是说,在确定初始探查点以后进行的几何渐进探查中,只需要再探查$(\log n)-3$个点,所以,最终的相关系数计算总次数仍然是$O(\log n)$.

此外,与 BRAID 方法相比,TPFP 方法多了采用三角法确定初始探查点的计算量.但是,三角法所涉及的计算量非常小,给出 3 个点以后,根据式(2)可以快速计算得到初始探查点,这其中所涉及的计算量要远远小于采用式(1)计算相关系数所涉及的计算量,因此,三角法所涉及的计算量几乎可以忽略不计.

因此,本文可以得出结论:TPFP 方法和 BRAID 方法具有几乎同样的时间复杂度.

4 实验设计与结果

本节将介绍实验设计和实验结果,目的在于说明采用 TPFP 方法可以比 BRAID 等方法取得更好的性能.

4.1 实验设计

实验环境:AMD Athlon II X2 2.70 GHz CPU;2 GB 内存,Visual C++ 6.0 编程;WINDOWS XP 操作系统.本文选取了美国股市 1980～2003 年的所有股票交易记录,获得了 1 228 764 条交易记录,每条记录包含交易日期、开盘价、收盘价和成交量等信息.本文选取某只股票的"收盘价"信息构成一个时间序列,由此可以得到多只股票的多个时间序列.在实验中,每次选择两个时间序列 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$,在它们上面放置两个滑动窗口并同步移动,对于滑动窗口内的子序列,分别采用 BRAID 方法和 TPFP 方法进行最大延迟相关点的计算,并比较二者的性能.在以下实验中,如果没有特殊说明,TPFP 方法都采用三角法确定初始探查点.

4.2 实验 1:针对具体某段时间序列的算法准确性

本实验的目的在于比较 BRAID 方法和 TPFP 方法在确定最大延迟相关点方面的准确性.本文选取了 6 对时间序列 DS_1,DS_2,DS_3,DS_4,DS_5,DS_6 进行实验,时间序列的长度统一取为 3 000,然后分别采用 Naïve,BRAID 和 TPFP 方法计算每对时间序列之间的最大延迟相关点的位置.从表 1 可以看出,当采用 Naïve 方法时,6 对时间序列的最大延迟相关点位置分别为 168,1 256,684,193,1 104,1 359,这种方法计算得到的位置是准确的结果,从中可以得知每对时间序列的最大延迟相关点的实际位置.对于一些最大延迟相关点位置较小的时间序列对,比如 DS_1,DS_3,DS_4,BRAID 和 TPFP 方法都具有比较好的性能,二者准确性也很接近.但是,对于一些最大延迟相关点位置较大的时间序列对,比如 DS_2,DS_5,BRAID 和 TPFP 方法的准确性差别很大,比如,对于 DS_2 和 DS_5 而言,BRAID 的误差率是 70.70% 和 47.30%,误差非常大,而这时 TPFP 方法的误差率只有 1.59% 和 1.18%,比 BRAID 方法小很多.从上述结果可以看出,在最大延迟相关点位置较大的情况下,BRAID 方法在多数情况下(比如 DS_2 和 DS_5)的准确性会变得很差,只有在少数情况下(比如 DS_6),具有较好的准确性;而 TPFP 则始终表现出了稳定的、较好的性能.

Table 1 Algorithm Accuracy Comparison

表 1 算法准确性比较

Datasets	*Lag*			Estimation Error/%	
	Naïve	BRAID	TPFP	BRAID	TPFP
DS_1	168	169	169	0.60	0.60
DS_2	1 256	368	1 236	70.70	1.59
DS_3	684	665	689	2.78	0.73
DS_4	193	196	195	1.55	1.04
DS_5	1 104	582	1 117	47.28	1.18
DS_6	1 359	1 322	1 369	2.72	0.74

4.3 实验 2:滑动窗口在时间序列上移动时的算法准确性

本实验在于说明滑动窗口在时间序列上移动时,BRAID 和 TPFP 算法的准确性的动态变化情况.这里让滑动窗口的大小取为 3 000,并在两个时间序列 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 上同步滑动 500 次,当滑动窗口位于时间序列 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的某个位置上时,可以得到两段长度分别为 3 000 的子序列,分别用 Naïve,BRAID 和 TPFP 方法计算这两个子序列的最大延迟相关点.其中,Naïve 方法计算得到的结果是准确的,BRAID,TPFP 方法计算得到的结果需要和 Naïve 方法计算得到的结果进行比较,计算误差值,即误差=(|BRAID 计算结果－Naïve 计算结果|)/Naïve 计算结果×100%.

从图 9 可以看出,BRAID 方法在大多数情形下具有良好的性能.图 10 显示了 BRAID 方法误差 e 的区间分布情况,从图 10 可以看出,误差 $e<1\%$ 的情形占 58%,误差 $e<2\%$ 的情形占 80%,但是,仍有 15.2% 的情形下误差 e 高达 5% 以上,这些情形通常是由于延迟突变或者最大延迟相关点位置较大而造成的,而 BRAID 方法又无法有效处理这些情形.

从图 11 可以看出,TPFP 方法的误差 e 绝大部分都小于 2%,说明该算法在大多情形下具有良好的稳定的性能,只有极少数情况下误差 $e>2\%$.经过

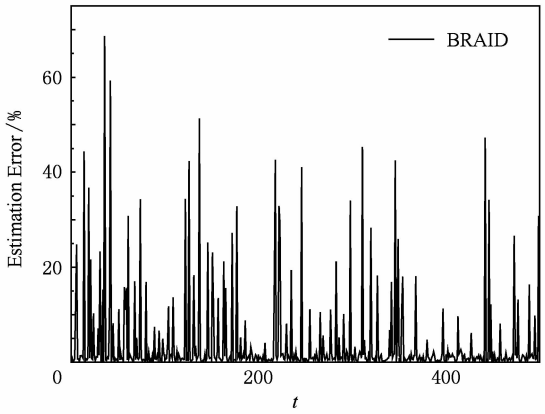


Fig. 9 Estimation error of BRAID when moving 500 times.

图 9 滑动 500 次时 BRAID 误差变化

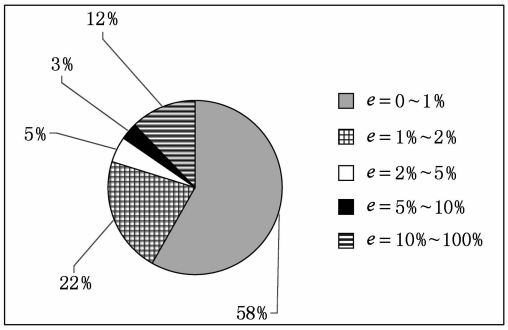


Fig. 10 Estimation error distribution of BRAID.

图 10 BRAID 误差区间分布

统计(如图 12 所示),在 TPFP 方法得到的结果中,误差 $e<1\%$ 的情形占 65%,误差 $e>1\%$,$<2\%$ 的情形占 27%,二者之和高达 92%,误差 $e>10\%$ 的情形只占 3.2%. 因此,可以看出,TPFP 方法对于绝大部分情形都具有较高的准确性,对于极少数误差较大的情形,这是由于 TPFP 方法没有正确探测到“延迟突变”造成的.

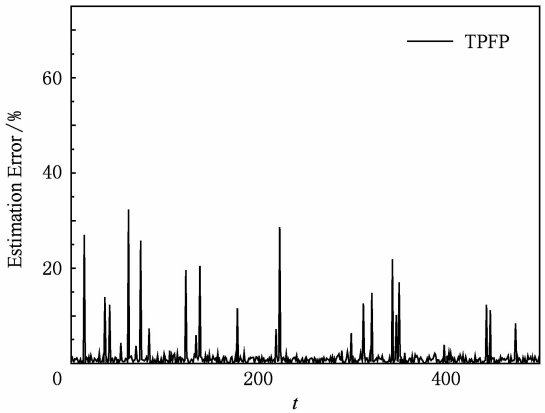


Fig. 11 Estimation error of TPFP when moving 500 times.

图 11 滑动 500 次时 TPFP 误差变化

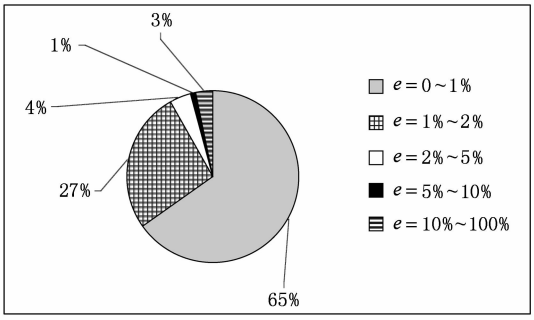


Fig. 12 Estimation error distribution of TPFP.

图 12 TPFP 误差区间分布

图 13 显示了当 BRAID 误差较大时,BRAID 和 TPFP 算法性能的比较情况. 从上面滑动 500 次计算得到的结果当中,共发现 61 次 BRAID 结果误差大于 10%,经过统计发现,这些情形或者是由于延迟位置较大,或者是由于存在延迟突变. 针对这 61 次结果,本文找出与之相对应的 TPFP 结果,并把二者用叠加柱状图的形式在图 13 中展现出来. 从图 13 可以发现,当采用 TPFP 方法时,误差大于 10% 的个数只有 15 个,误差大于 5% 的个数也只有 18 个,而误差小于 2% 的个数更是达到了 41 个. 也就是说,对于同样的时间序列,在采用 BRAID 方法误差大于 10% 的 61 个结果中,TPFP 方法可以有效处理其中 67.2% 的情形,让误差减小到 2% 以内. 对于剩余的 32.8% 的情形,虽然 TPFP 方法无法把误差减小到 2% 以内,但是,误差大小也都要比 BRAID 方法的误差小.

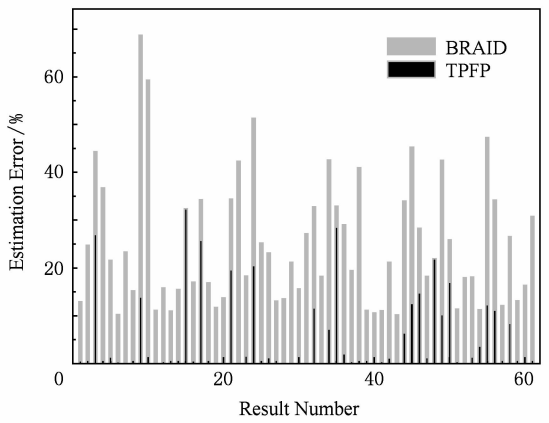


Fig. 13 Performance comparison when estimation error of BRAID is large.

图 13 BRAID 误差较大时的性能比较

类似地,图 14 显示了当 TPFP 误差较大时 BRAID 和 TPFP 算法性能的比较情况. 从中可以看出,在 500 次结算结果中,只有 16 个结果的误差大于 10%. 而对于这些情形,采用 BRAID 方法时,

结果同样具有很大的误差,而且绝大多数误差比TPFP大,只有在图14中的第3个和第4个结果中,BRAID方法误差比TPFP方法小,即BRAID方法误差分别为9%和30.8%,而TPFP方法误差分别为12.3%和32.36%.

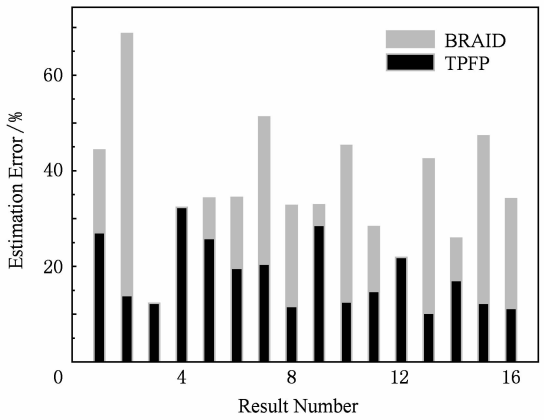


Fig. 14 Performance comparison when estimation error of TPFP is large.

图14 TPFP误差较大时的性能比较

IA(interesting area)方法[4]虽然采用了"兴趣区域"的概念,可以处理延迟较大的情形,但是,却无法有效处理延迟突变问题.为了验证这一点,本文设计了相关实验.本文对BRAID误差大于10%的61个结果进行分析,从中挑选出15个属于不存在延迟突变但是延迟较大的情形,并另外挑选出15个属于延迟突变的情形,然后,针对这些情形分别开展实验测试BRAID,TPFP和IA算法的性能.如图15所示,当不存在延迟突变但是延迟较大时,TPFP和IA算法都取得较好的准确性,性能比较接近.如图16所示,当存在延迟突变时,BRAID方法和IA方法的误差都比较大,只有在第2个和第5个结果中,IA的准确性接近或略好于TPFP.在其他情形下,TPFP的准确性都要优于IA和BRAID.这说明,IA无法有效处理延迟突变的情形,而TPFP则具有这个能力.

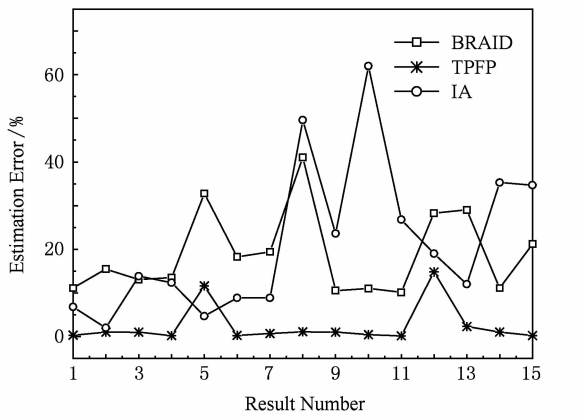


Fig. 16 Accuracy comparison between TPFP and IA on the occasion of lag mutation.

图16 当存在延迟突变时TPFP和IA的准确性比较

4.4 实验3:三角法和插值法的性能比较

本实验的目的在于说明,在确定初始探查点时,采用简单的三角法可以取得和相对复杂的插值法几乎同样好的性能.为了进行区分,这里把采用三角法的TPFP方法仍旧称为"TPFP",把采用插值法的TPFP方法称为"TPFP-INTER".通过实验,本文发现采用TPFP-INTER时的误差分布和TPFP基本类似,不再详细展示,这里只给出对于那些BRAID方法误差大于10%的情形,BRAID,TPFP和TPFP-INTER的性能比较.从图17可以看出,对于BRAID方法误差较大的情形,TPFP和TPFP-INTER方法

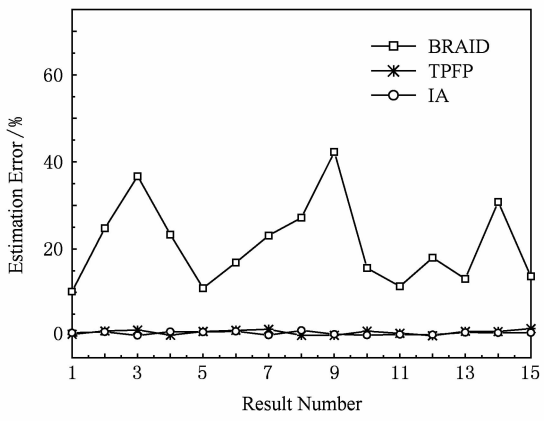


Fig. 15 Accuracy comparison between TPFP and IA on the occasion of large lag.

图15 当延迟较大时TPFP和IA的准确性比较

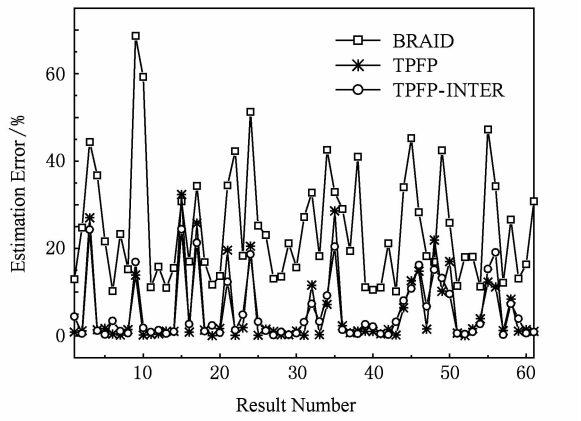


Fig. 17 Accuracy comparison between TPFP and TPFP-INTER.

图17 TPFP和TPFP-INTER的准确性比较

都能够进行有效地处理,从处理效果而言,二者并没有明显差别.经过统计,在 61 个 BRAID 误差较大的情形中,在其中的 33 个地方,TPFP-INTER 的处理效果好于 TPFP,即把误差降低到更小的值,但是,在其他的 28 个地方,采用三角法的 TPFP 方法性能要好于 TPFP-INTER.因此,可以得出,二者在准确性方面没有明显差别.但是,在计算时间方面(如图 18 所示),TPFP 明显好于 TPFP-INTER,随着时间序列上面滑动窗口的长度的增加,TPFP 的时间优势更加明显.

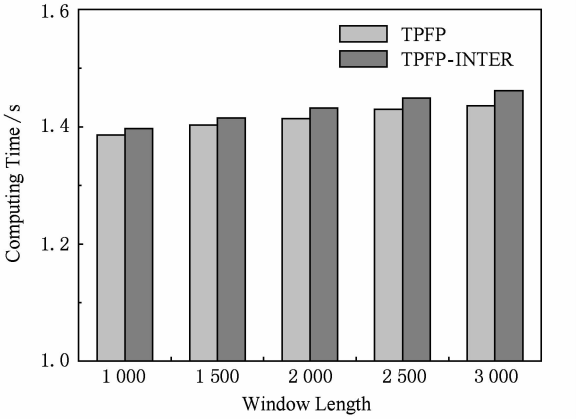


Fig. 18 Computing time comparison between TPFP and TPFP-INTER.

图 18 TPFP 和 TPFP-INTER 的运行时间比较

5 相关工作

针对时间序列的数据挖掘[1,6-8]是一个热门的研究领域.比如,文献[1]研究了基于分段时间弯曲距离的时间序列挖掘,提出了一种新的序列分段方法和距离度量方法,在比较时间序列相似性方面取得了较好的性能.时间序列数据挖掘的一个重要研究内容就是时间序列之间的相关性[9-13].文献[14]提出了一种网格结构对成千上万个时间序列进行检测,找出高度相关(相关系数大于常量 R)的时间序列对,算法的时间复杂度为 $O(d)$,其中 d 表示数据的维数.但是,对于不同的 R 值,查询都要重新扫描整个时间序列数据库.为此,文献[12]采用了一种基于树的索引机制,可以避免重复扫描时间序列数据库,并且支持即席相关性查询和变长查询.文献[9]研究了空间时间序列的相关性问题,作者利用空间相邻时间序列之间的空间自相关性来减少计算代价.文献[15]分析了多个气候时间序列的线性相关问题,并且使用聚类来构造气候索引.文献[16]研究了随着时间演化的时间序列之间的局部相关性问题,作者提出的时间序列相关性度量方法具有很好的鲁棒性和高效性,会随着时间演化准确地反映时间序列之间变化的关系.文献[3]针对数据仓库中的大量时间序列之间的相关性问题,提出了基于 DFT 和图分区的相关性系数计算方法,大大减少了算法运行时间,降低了 CPU 和 I/O 开销.

时间序列之间的延迟相关性也是一个重要的研究方向.文献[2]提出了适用于数据流环境的 BRAID 方法,可以在 $O(\log l)$ 时间复杂度内发现延迟相关的最大值,即相关性系数最大时对应的延迟,其中,l 是最大的可能延迟.BRAID 不会对所有延迟值下的相关系数进行计算,而是采用几何递增渐进的方式,只选择少数延迟值进行相关性系数计算,然后采用插值方法得到剩余延迟值对应的相关性系数,最后比较得到相关性系数最大时对应的延迟值.BRAID 这种思路在其他研究中得到了继承,比如,文献[3]在计算延迟相关最大值时也采用了这种几何渐进探查思想,二者的主要区别在于 BRAID 是在时间领域内计算相关性,而文献[3]则是在频率领域内计算相关性.但是,由于 BRAID 总是从 $l=0$ 的地方开始探查,因此,误差会随着 l 的增加而增大.因此,文献[4]提出了"兴趣区域"的概念,并对区域的关键位置(波峰和边界)进行跟踪,从每个局部最大波峰开始探查.这样,即使 l 很大时也可以得到比较准确的结果.该方法在一定程度上弥补了 BRAID 的不足,但是仍有无法克服的缺陷,即它无法处理"延迟突变"问题.而本文提出的方法可以很好解决这一问题.

6 结束语

时间序列延迟相关分析是时间序列研究领域的一个重要问题.目前已有的研究都存在一定的缺陷,比如,类似 BRAID 的方法每次都从延迟为 0 的地方开始探查,使得这类方法在最大延迟相关点位置较小时可以取得较好的性能,但是,在最大延迟相关点位置较大时往往误差很大.另一些类似 IA 的方法可以根据"兴趣区域"设置探查点,在最大延迟相关点位置较大时也可以获得较小的误差,但是,这类方法无法解决延迟突变的问题.

本文通过大量实验发现和验证了延迟相关分析中存在的 3 个实验现象,即连续分布性、延迟突变和突变幅度分布特性.利用这 3 个实验现象,本文提出了 TPFP 方法,它即可以处理最大延迟相关点位置

较大的情形,也可以有效处理延迟突变问题.本文还用大量实验证明了 TPFP 方法比其他方法具有更好的性能.

在未来的研究工作当中,将研究探查点预测错误纠正方法.因为在当前的 TPFP 方法中,当发生延迟突变时,仍有少部分初始探查点预测结果会导致较大的延迟位置计算误差,因此,需要研究相关的误差反馈和探查点位置纠正方法.

参 考 文 献

[1] Xiao Hui, Hu Yunfa. Data mining based on segmented time warping distance in time series database [J]. Journal of Computer Research and Development, 2005, 42(1): 72-78 (in Chinese)
(肖辉,胡运发. 基于分段时间弯曲距离的时间序列挖掘[J]. 计算机研究与发展, 2005, 42(1): 72-78)

[2] Sakurai Y, Papadimitriou S, Faloutsos C. BRAID: Stream mining through group lag correlations [C] //Proc of ACM SIGMOD'05. New York: ACM, 2005: 599-610

[3] Mueen A, Nath S, Liu Jie. Fast approximate correlation for massive time-series data [C] //Proc of ACM SIGMOD'10. New York: ACM, 2010: 171-182

[4] Wu Di, Ke Yiping, Yu J, et al. Detecting leaders from correlated time series [G] //LNCS 5981: Proc of DASFAA'10. Berlin: Springer, 2010: 352-367

[5] Lin Ziyu, Yang Dongqing, Wang Tengjiao. Similarity search of time series with moving average based indexing [J]. Journal of Software, 2008, 19(9): 2349-2361 (in Chinese)
(林子雨,杨冬青,王腾蛟. 用基于移动均值的索引实现时间序列相似查询[J]. 软件学报, 2008, 19(9): 2349-2361)

[6] Lee A, Wu H, Lee T, et al. Mining closed patterns in multi-sequence time-series databases [J]. Data Knowledge Engineering, 2009, 68(10): 1071-1090

[7] Shieh J, Keogh E. SAX: Disk-aware mining and indexing of massive time series datasets [J]. Data Mining and Knowledge Discovery, 2009, 19(1): 24-57

[8] Wu H, Lee A. Mining closed flexible patterns in time-series databases [J]. Expert System Application, 2010, 37(3): 2098-2107

[9] Zhang Pusheng, Huang Yan, Shekhar S, et al. Correlation analysis of spatial time series datasets: A filter-and-refine approach [C] //LNCS 2637: Proc of PAKDD'03. Berlin: Springer, 2003: 532-544

[10] Idé T, Papadimitriou S, Vlachos M. Computing correlation anomaly scores using stochastic nearest neighbors [C] //Proc of ICDM'07. Piscataway, NJ: IEEE, 2007: 523-528

[11] Vlachos M, Wu K, Chen S, et al. Fast burst correlation of financial data [G] //LNCS 3721: Proc of PKDD'05. Berlin: Springer, 2005: 368-379

[12] Nguyen P, Shiri N. Fast correlation analysis on time series datasets [C] //Proc of CIKM'08. New York: ACM, 2008: 787-796

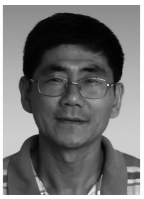
[13] Dorr D, Denton A. Establishing relationships among patterns in stock market data [J]. Data Knowledge Engineering, 2009, 68(3): 318-337

[14] Zhu Yunyue, Shasha D. Statstream: Statistical monitoring of thousands of data streams in real time [C] //Proc of VLDB'02. San Fransisco: Morgan Kaufmann, 2002: 358-369

[15] Steinbach M, Tan Pangning, Kumar V, et al. Discovery of climate indices using clustering [C] //Proc of KDD'03. New York: ACM, 2003: 446-455

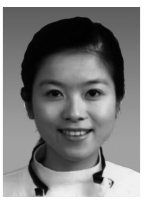
[16] Papadimitriou S, Sun J, Yu P. Local correlation tracking in time series [C] //Proc of ICDM'06. Piscataway, NJ: IEEE, 2006: 456-465

Lin Ziyu, born in 1978. Received his PhD degree in computer software and theory from Peking University in 2009. Now he is an assistant professor and master supervisor at Xiamen University. Member of China Computer Federation. His main research interests include data warehouse, OLAP, data mining, etc.



Jiang Yi, born in 1960. Associate professor and master's supervisor at Xiamen University. His main research interests include data warehouse, OLAP, data mining, etc (jiangyi@xmu.edu.cn).

Lai Yongxuan, born in 1981. Received his PhD degree in applied technology of computer science from Renmin University of China in 2009. Now he is an assistant professor and master's supervisor at Xiamen University. His main research interests include database, data management on sensor network, opportunistic network, etc (laiyx@xmu.edu.cn).



Lin Chen, born in 1982. Received her PhD degree in computer software and theory from Fudan University in 2010. Now she is an assistant professor and master's supervisor at Xiamen University. Her main research interests include retrieving, mining and organizing unstructured and semistructured data (chenlin@xmu.edu.cn).