ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.17, No.8, August 2006, pp.1753−1763
DOI: 10.1360/jos171753

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

挖掘多数据流的异步偶合模式的抗噪声算法*

陈安龙[1,2+], 唐常杰[1], 元昌安[1,3], 彭 京[1], 胡建军[1]

[1](四川大学 计算机学院,四川 成都 610065)

[2](电子科技大学 计算机科学与工程学院,四川 成都 610054)

[3](广西师范学院 信息技术系,广西 南宁 530001)

An Anti-Noise Algorithm for Mining Asynchronous Coincidence Pattern in Multi-Streams

CHEN An-Long[1,2+], TANG Chang-Jie[1], YUAN Chang-An[1,3], PENG Jing[1], HU Jian-Jun[1]

[1](College of Computer, Sichuan University, Chengdu 610065, China)

[2](College of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China)

[3](Department of Information Technology, Guangxi Teachers Education University, Nanning 530001, China)

+ Corresponding author: Phn: +86-28-85466105, E-mail: chenanlong@126.com, http://www.cs.scu.edu.cn/~tangchangjie

Chen AL, Tang CJ, Yuan CA, Peng J, Hu JJ. An anti-noise algorithm for mining asynchronous coincidence pattern in multi-streams. *Journal of Software*, 2006,17(8):1753−1763. http://www.jos.org.cn/1000-9825/17/1753.htm

Abstract: Mining asynchronous coincidence pattern is a difficult task in multi-data streams. The main contributions of this work included: (1) The filter technique of Haar Wavelet is investigated and applied to mining asynchronous coincidence pattern in multi-streams; (2) The Wavelet coefficient series are applied to the measurement of asynchronous coincidence between data streams. A series of theorems are proved to ensure the validity of measuring asynchronous coincidence; (3) The anti-noise increment algorithms are designed on loop sliding windows to mine asynchronous coincidence pattern and implemented with complexity $O(n^2)$; (4) The extensive experiments on real data are given to validate algorithms.

Key words: multi-data stream; asynchronous coincidence pattern; Haar wavelet; loop sliding window

摘 要: 挖掘多数据流的异步偶合模式是具有挑战性的工作.主要的研究工作包括:(1) 研究 Haar 小波滤波技术在挖掘流数据的异步偶合模式中的应用;(2) 引入小波系数序列来度量数据流的异步局域偶合度;证明了一系列定理,保证了度量方法的正确性;(3) 设计了环形滑动窗口和挖掘异步偶合模式的抗噪声增量算法,其时间复杂性小于 $O(n^2)$;(4) 使用真实数据进行模拟实验,验证了算法的有效性.

关键词: 多数据流;异步偶合模式;Haar 小波;环形滑动窗口

中图法分类号: TP311 文献标识码: A

由于网络通信技术和传感器技术的高速发展,产生了大量的流数据,从而促进了对流数据内在规律的研究.

* Supported by the National Natural Science Foundation of China under Grant No.60473071 (国家自然科学基金); the National Research Foundation for the Doctoral Program of Higher Education of China under Grant No.20020610007 (国家教育部博士点基金)

Received 2005-07-28; Accepted 2005-12-01

例如,电话通话记录、股票交易指数、气象监测数据、传感器监测数据以及网管日志数据等.相对于数据库的静态数据,流数据呈现快速变化、海量无限和实时连续出现等特点,引起了学术界的极大研究兴趣.文献[1]研究了挖掘流数据中临时频繁模式的方法,根据滑动窗口的频繁模式支持度的变化,使用线性回归的方法对频繁模式支持度进行预测;文献[2]用概率分布函数的变化来描述数据流的变化特征,并运用分布函数变化的度量方法来发现数据流的变化规律;文献[3]研究了如何从数据流中发现假阴性或假阳性的频繁模式集;文献[4]研究了滑动窗口中的数据流的统计问题,提出了一种联机数据结构,在滑动窗口上发现统计特征;文献[5]使用直方图的方法,研究了多数据流聚类的近似计算问题;文献[6]研究了静态数据库中的时间序列数据的近似查询问题;文献[7,8]研究了小波变换技术应用于数据流特征的近似描述,并利用小波变换对数据流进行聚类分析、近似搜索和压缩数据.这些研究工作主要集中在数据流的查询、统计分析和频繁模式的发现上,但对数据流之间的偶合性研究较少.文献[9]运用离散的傅里叶变换研究了多数据流之间的偶合性,但没有深入研究数据流的异步偶合特性,傅里叶变换存在难于表征具有局域突变特性的信号的局限性.小波技术正是解决该问题的有效数学工具,本文深入研究了 Haar 小波技术与流数据处理技术的融合,并应用于数据流之间的异步偶合模式的发现.

1 本文主要贡献

研究多数据流偶合模式的工作主要集中在同步偶合规律的发现上,某些呈现同步弱偶合的数据流平移一定的相差之后,将呈现异步强偶合特征.本文继承了同步偶合数据流的研究成果,融合 Haar 小波的滤波和压缩技术,主要研究工作包括:(1) 融合了 Haar 小波的滤波技术和流数据挖掘技术,研究了数据流的异步偶合模式的挖掘方法,定义了流数据间的局域异步偶合度的度量方法;(2) 探索了 Haar 小波技术处理数据流的理论依据,提出局域异步偶合模型,证明了流的局域中心距定理、异步偶合等价定理、距离等价定理、偶合度距离化定理和强偶合判定定理等一系列定理,揭示了 Haar 小波技术与数据流处理相融合的内在本质,为使用压缩后的小波系数信息计算偶合度提供了理论依据,避免了重构数据流信息的开销;(3) 设计了嵌套环形滑动窗口,解决有限内存空间和数据无限的矛盾,同时为增量式的小波变换提供数据结构;(4) 设计了挖掘异步偶合模式的抗噪声增量算法,提高了计算效率,并从局域异步偶合模型的相关定理导出了相关推论,保证了算法的合理性;(5) 采用真实数据对算法进行了模拟实验,验证了抗噪声增量算法的有效性以及小波系数度量数据流的异步偶合程度的合理性.

2 Haar 小波方法介绍

傅里叶变换在处理非平稳信号时存在局限性,小波方法正是解决该问题的数学工具.该方法主要包括尺度函数$\varphi(x)$和小波函数$\phi(x)$,本节引用文献[10]的小波定义和分解定理,导出矩阵表示和相关引理.

定义 **2.1**. Haar 尺度函数$\varphi(x)$定义为

$$\varphi(x)=\begin{cases}1, & x\in[0,1)\\ 0, & \text{otherwise}\end{cases}.$$

定义 **2.2**. Haar 小波函数$\phi(x)$定义为$\phi(x)=\varphi(2x)-\varphi(2x-1)$.

函数空间 $V_j=\{f(x)|f(x)=\Sigma(a_k\times 2^{j/2}\varphi(2^jx-k));k\in Z,a_k\in R\}$,$W_j=\{g(x)|g(x)=\Sigma(a_k\times 2^{j/2}\phi(2^jx-k));k\in Z,a_k\in R\}$;$V_{j+1}=V_j\oplus W_j$,即 W_j 是 V_j 在 V_{j+1} 的正交补.$f_n(x)=\Sigma(a_{n,k}\times 2^{n/2}\varphi(2^nx-k))\in V_n$,根据 Haar 分解定理,$f_n(x)=w_{n-1}+f_{n-1}(x)$,其中:$f_{n-1}(x)=\Sigma(a_{n-1,k}\times 2^{(n-1)/2}\varphi(2^{n-1}x-k))\in V_{n-1}$,$w_{n-1}(x)=\Sigma(b_{n-1,k}\times 2^{(n-1)/2}\phi(2^{n-1}x-k))\in W_{n-1}$;系数 $b_{n-1,k}=(a_{n,2k}-a_{n,2k+1})/2^{1/2}$,$a_{n-1,k}=(a_{n,2k}+a_{n,2k+1})/2^{1/2}$.按照此方法逐层分解,直到 $f_n(x)=w_{n-1}+w_{n-2}+\ldots+w_1+f_0(x)$且 $w_j\in W_j(1\le j\le n-1)$.Haar 分解定理给出了逐层分解数据信息的方法.借助于此分解方法,可以把具有 2^n 个数据点的序列看作是函数 $f_n(k/2^n)=a_{n,k}\times 2^{n/2}$ $(0\le k\le 2^n-1)$的 2^n 个系数值 $a_{n,k}$,利用 $b_{n-1,k}=(a_{n,2k}-a_{n,2k+1})/2^{1/2}$,$a_{n-1,k}=(a_{n,2k}+a_{n,2k+1})/2^{1/2}$,对数据序列的 2^n 个数据 $a_{n,0},a_{n,1},\ldots$进行层次分解,经过一次分解后,得到 2^{n-1} 个数据 $b_{n-1,0},b_{n-1,1},\ldots$和 2^{n-1} 个数据 $a_{n-1,0},a_{n-1,1},\ldots$此分解可以表示为矩阵关系,矩阵为 2^{n-1} 行和 2^n 列:

$$\begin{bmatrix} b_{n-1,0} \\ b_{n-1,1} \\ \vdots \\ b_{n-1,2^{n-1}-1} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \times \begin{bmatrix} a_{n,0} \\ a_{n,1} \\ \vdots \\ a_{n,2^n-1} \end{bmatrix},$$

$$\begin{bmatrix} a_{n-1,0} \\ a_{n-1,1} \\ \vdots \\ a_{n-1,2^{n-1}-1} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \times \begin{bmatrix} a_{n,0} \\ a_{n,1} \\ \vdots \\ a_{n,2^n-1} \end{bmatrix}.$$

第 2 次采用类似的方法对 2^{n-1} 个数据 $a_{n-1,0},a_{n-1,1},\dots$进行分解,可等价于对 $a_{n,0},a_{n,1},\dots,a_{n,m}(m=2^n-1)$进行如下变换:

$$b_{n-2,0}=\begin{bmatrix} \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 0 & \dots & 0 \end{bmatrix} \times [a_{n,0}\ a_{n,1}\ \dots\ a_{n,m}]^T,$$

$$b_{n-2,1}=\begin{bmatrix} 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 0 & \dots & 0 \end{bmatrix} \times [a_{n,0}\ a_{n,1}\ \dots\ a_{n,m}]^T\dots,$$

直到

$$b_{0,0}=\begin{bmatrix} \frac{1}{\sqrt{2^n}} & \dots & \frac{1}{\sqrt{2^n}} & -\frac{1}{\sqrt{2^n}} & \dots & -\frac{1}{\sqrt{2^n}} \end{bmatrix} \times [a_{n,0}\ a_{n,1}\ \dots\ a_{n,m}]^T,$$

$$a_{0,0}=\begin{bmatrix} \frac{1}{\sqrt{2^n}} & \dots & \frac{1}{\sqrt{2^n}} & \frac{1}{\sqrt{2^n}} & \dots & \frac{1}{\sqrt{2^n}} \end{bmatrix} \times [a_{n,0}\ a_{n,1}\ \dots\ a_{n,m}]^T.$$

由 Haar 小波的层次分解算法,上述矩阵的各行向量是相互正交的单位向量(参见文献[10]).为了方便后续问题的讨论,本文将 Haar 层次分解法表示为下面的矩阵形式(M 称为小波变换矩阵),并给出了两个引理.

$$\begin{bmatrix} b_{n-1,0} \\ b_{n-1,1} \\ \vdots \\ b_{n-1,2^{n-1}-1} \\ b_{n-2,0} \\ \vdots \\ b_{0,0} \\ a_{0,0} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\sqrt{2^n}} & \frac{1}{\sqrt{2^n}} & \dots & \frac{1}{\sqrt{2^n}} & \frac{1}{\sqrt{2^n}} & -\frac{1}{\sqrt{2^n}} & -\frac{1}{\sqrt{2^n}} & \dots & -\frac{1}{\sqrt{2^n}} & -\frac{1}{\sqrt{2^n}} \\ \frac{1}{\sqrt{2^n}} & \frac{1}{\sqrt{2^n}} & \dots & \frac{1}{\sqrt{2^n}} & \frac{1}{\sqrt{2^n}} & \frac{1}{\sqrt{2^n}} & \frac{1}{\sqrt{2^n}} & \dots & \frac{1}{\sqrt{2^n}} & \frac{1}{\sqrt{2^n}} \end{bmatrix} \times \begin{bmatrix} a_{n,0} \\ a_{n,1} \\ \vdots \\ a_{n,2^n-1} \end{bmatrix} = M \times \begin{bmatrix} a_{n,0} \\ a_{n,1} \\ \vdots \\ a_{n,2^n-1} \end{bmatrix}.$$

为了便于理解,假设有流数据{3,3,6,8},可简记为数组:X=[3,3,6,8].分别计算$(x_0+x_1)/2^{1/2}=(3+3)/2^{1/2}=3\times 2^{1/2}$;$(x_0-x_1)/2^{1/2}=(3-3)/2^{1/2}=0,\dots$,构造新序列[$3\times 2^{1/2}$,$7\times 2^{1/2}$],反复使用该方法,计算过程见表 1(其中,序列 X=[3,3,6,8]经过变换后的序列为 Y^T=[0,$-2^{1/2}$,−4,10]).

Table 1 Haar wavelet transform of stream data

表 **1** 流数据的 Haar 小波分解过程

$(x_{2i}+x_{2i+1})/2^{1/2}$	$(x_{2i}-x_{2i+1})/2^{1/2}$
[3,3,6,8]	
[$3\times 2^{1/2}$,$7\times 2^{1/2}$]	[0,$-2^{1/2}$]
[10]	[−4]

引理 **2.1**. 设 M 是小波变换矩阵,M^T 为 M 的转置矩阵,则 M^T 是 M 的逆矩阵.

证明:由 Haar 小波分解过程可知,M 的行向量是单位正交向量,则矩阵 M^T 的列向量也是单位正交向量.所以,矩阵 M 的任意行向量 X 与 M^T 的列向量 Y 的乘积,当 $Y=X^T$ 时,$X\times Y=1$;当 $Y\neq X^T$ 时,则有 $X\times Y=0$.所以,$M\times M^T=E$(E 为单位向量).可推得$(M\times M^T)^T=E\Rightarrow M^T\times M=E$,则 M^T 是 M 的逆矩阵. □

引理 **2.2**. 设 M 是小波变换矩阵,对于给定的向量 $X=[x_1,x_2,\ldots,x_m]$和 $Y=[y_1,y_2,\ldots,y_m]$($m=2^n$),如果 $Y^T=M\times X^T$,则向量的模满足$\|X\|^2=\langle X,X\rangle=\|Y\|^2=\langle Y,Y\rangle$.

证明:由向量的范数的定义可知:$\|Y\|^2=\langle Y,Y\rangle$,则有 $\|Y\|^2=\langle Y,Y\rangle=[y_1,y_2,\ldots,y_m]\times[y_1,y_2,\ldots,y_m]^T=(M\times X^T)^T\times(M\times X^T)=X\times M^T\times M\times X^T=X\times(M^T\times M)\times X^T=X\times E\times X^T=\langle X,X\rangle=\|X\|^2$. □

使用小波变换的主要目的是消除高频噪声和压缩处理数据.对于小波压缩后的数据处理,一般需用重构算法重构信息.本文研究在不重构数据的条件下,小波滤波技术在数据流的异步偶合特征挖掘中的应用.

3 数据流的局域异步偶合

3.1 数据流的异步偶合度

流数据具有无限性,不可能存储数据的完整信息;数据呈现具有不可再现性,只能对数据流进行一次性处理.因此,采用滑动窗口机制对数据流间的局部偶合关系进行研究.但是,许多数据流之间在一定的相位差(时延)后,才表现出明显的偶合关系,在此主要研究数据流之间的异步偶合关系.从概率统计的角度对数据流间的局域异步偶合度作出如下定义:

定义 **3.1**(异步偶合度). 设 $X_1=[x_{1,1},x_{1,2},\ldots,x_{1,w}]$和 $X_2=[x_{2,1},x_{2,2},\ldots,x_{2,w}]$是含有 w 个数据点的两个数据流序列.相距时延 t 的序列 X_1 和 X_2 的异步偶合度为

$$Corr(X_1,X_2,t)=\frac{\sum_{i=t+1}^{w}\left(x_{1,i-t}-\lambda(x_1)\right)\left(x_{2,i}-\lambda(x_2)\right)}{\sqrt{\sum_{i=t+1}^{w}\left(x_{1,i-t}-\lambda(x_1)\right)^2\sum_{i=t+1}^{w}\left(x_{2,i}-\lambda(x_2)\right)^2}},$$

其中, $\lambda(x_1)=\frac{1}{w-t}\sum_{i=t+1}^{w}x_{1,i-t}$, $\lambda(x_2)=\frac{1}{w-t}\sum_{i=t+1}^{w}x_{2,i}$.

定义 3.1 给出了分别具有 w 个数据点的两个序列,序列间的时差为 t,其偶合度的计算方法是:当 $t=0$ 时,则表示两个同步数据流之间的偶合程度,是异步偶合度的特例.

3.2 局域异步偶合的等价度量模型

为了更准确地度量数据流的异步偶合程度,采用 Haar 小波技术消除数据流中的高频噪声,同时达到压缩数据的目的.为了避免由小波系数重构数据信息的时间开销,在此研究小波系数度量数据流的异步偶合度的方法,进而引入下列定理,以保证度量方法的正确性.

定理 **3.1**(局域中心距定理). 设 $x_1,x_2,\ldots,x_m$ 是有 $m(m=2^n)$个数据点的序列,记 $X=[x_1,x_2,\ldots,x_m]$,M 为小波变换矩阵, $\lambda(x)=\frac{1}{m}\sum_{i=1}^{m}x_i$.如果 $Y^T=[y_1,y_2,\ldots,y_m]^T=M\times X^T$,则有 $\sigma_x^2=\sum_{i=1}^{m}(x_i-\lambda(x))^2=\sum_{i=1}^{m-1}y_i^2$.

证明:因为
$$\begin{aligned}\sigma_x^2&=\sum_{i=1}^{m}(x_i-\lambda(x))^2=\sum_{i=1}^{m}(x_i^2-2\lambda(x)\times x_i+(\lambda(x))^2)\\&=[x_1,x_2,\ldots,x_m]\times[x_1,x_2,\ldots,x_m]^T-m\times(\lambda(x))^2\\&=[x_1,x_2,\ldots,x_m]\times M^T\times M\times[x_1,x_2,\ldots,x_m]^T-m\times(\lambda(x))^2\\&=(M\times[x_1,x_2,\ldots,x_m]^T)^T\times M\times[x_1,x_2,\ldots,x_m]^T-m\times(\lambda(x))^2\\&=[y_1,y_2,\ldots,y_m]\times[y_1,y_2,\ldots,y_m]^T-m\times(\lambda(x))^2\\&=\sum_{i=1}^{m}y_i^2-m\times(\lambda(x))^2,\end{aligned}$$

由线性变换 $Y^T=[y_1,y_2,\ldots,y_m]^T=M\times X^T$ 可知: $y_m=\frac{1}{\sqrt{2^n}}\sum_{i=1}^{m}x_i=\frac{1}{\sqrt{m}}\sum_{i=1}^{m}x_i\Rightarrow y_m^2=m\times(\lambda(x))^2$,

则有 $\sigma_x^2=\sum_{i=1}^{m}(x_i-\lambda(x))^2=\sum_{i=1}^{m-1}y_i^2$ 成立. □

定理 **3.2**(异步偶合等价定理). 设 $X_1=[x_{1,1},x_{1,2},\ldots,x_{1,w}]$和 $X_2=[x_{2,1},x_{2,2},\ldots,x_{2,w}]$是两个流序列,$M$ 为 $2^n\times2^n$ 的小波变换矩阵($w=t+2^n$),对于时延 t,$X_{(1,t)}=[x_{1,1},x_{1,2},\ldots,x_{1,w-t}]$,$X_{(2,t)}=[x_{2,t+1},x_{2,t+2},\ldots,x_{2,w}]$, $Y_{(1,t)}^T=[y_{1,1},y_{1,2},\ldots,y_{1,w-t}]^T=M\times X_{(1,t)}^T$, 且 $Y_{(2,t)}^T=[y_{2,t+1},y_{2,t+2},\ldots,y_{2,w}]^T==M\times Y_{(2,t)}^T$,则序列 X_1 和 X_2 的异步偶合度为

$$Corr(X_1,X_2,t)=\frac{\sum_{i=t+1}^{w-1}(y_{1,i-t}\times y_{2,i})}{\sqrt{\sum_{i=t+1}^{w-1}y_{1,i-t}^2\times\sum_{i=t+1}^{w-1}y_{2,i}^2}}.$$

证明:首先记 $\lambda(x_1)=\frac{1}{w-t}\sum_{i=t+1}^{w}x_{1,i-t}$, $\lambda(x_2)=\frac{1}{w-t}\sum_{i=t+1}^{w}x_{2,i}$;

$\Delta_1=[x_{1,1}-\lambda(x_1),x_{1,2}-\lambda(x_1),\ldots,x_{1,w-t}-\lambda(x_1)]$, $\Delta_2=[x_{2,t+1}-\lambda(x_2),x_{2,t+2}-\lambda(x_2),\ldots,x_{2,w}-\lambda(x_2)]$.

因为 M 是 $2^n\times2^n$ 的小波变换矩阵,且 M^T 是逆矩阵,

$$\sum_{i=t+1}^{w}(x_{1,i-t}-\lambda(x_1))(x_{2,i}-\lambda(x_2))=\Delta_1\times\Delta_2^T=\Delta_1\times M^T\times M\times\Delta_2^T=(M\times\Delta_1^T)^T\times M\times\Delta_2^T$$

$$=(Y_{(1,t)}^T-[0,0,\ldots,0,\sqrt{2^n}\lambda(x_1)]^T)^T\times(Y_{(2,t)}^T-[0,0,\ldots,0,\sqrt{2^n}\lambda(x_2)]^T),$$

由小波变换 $Y=M\times X^T$ 可知: $y_{1,w-t}=\sqrt{2^n}\lambda(x_1)$, $y_{2,w}=\sqrt{2^n}\lambda(x_2)$,

所以, $\sum_{i=t+1}^{w}(x_{1,i-t}-\lambda(x_1))(x_{2,i}-\lambda(x_2))=\sum_{i=t+1}^{w-1}(y_{1,i-t}\times y_{2,i})$.

由定理 3.1 可知: $\sum_{i=t+1}^{w}(x_{1,i-t}-\lambda(x_1))^2=\sum_{i=t+1}^{w-1}y_{1,i-t}^2$, $\sum_{i=t+1}^{w}(x_{2,i}-\lambda(x_2))^2=\sum_{i=t+1}^{w-1}y_{2,i}^2$.

所以,$Corr(X_1,X_2,t)=\frac{\sum_{i=t+1}^{w}(x_{1,i-t}-\lambda(x_1))(x_{2,i}-\lambda(x_2))}{\sqrt{\sum_{i=t+1}^{w}(x_{1,i-t}-\lambda(x_1))^2\sum_{i=t+1}^{w}(x_{2,i}-\lambda(x_2))^2}}=\frac{\sum_{i=t+1}^{w-1}(y_{1,i-t}\times y_{2,i})}{\sqrt{\sum_{i=t+1}^{w-1}y_{1,i-t}^2\times\sum_{i=t+1}^{w-1}y_{2,i}^2}}$ 成立. □

定理 3.2 给出了两个数据流在相差 t 时刻后的异步局域偶合度的计算方法,将局部数据序列构成的向量,经小波变换得到系数序列,使用两个系数序列计算异步偶合度,但与系数 $y_{1,w-t}$ 和 $y_{2,w}$ 无关,揭示了分别使用原始数据序列和小波系数序列计算异步局域偶合度的等价性.如果将异步数据流进行平移 t 后,则与 $t=0$ 时的偶合度计算等价.为了方便讨论问题,本文后续部分将假定 $t=0$.

定理 **3.3**(局域距离等价定理). 设 $X_1=[x_{1,1},x_{1,2},\ldots,x_{1,m}]$和 $X_2=[x_{2,1},x_{2,2},\ldots,x_{2,m}]$是两个流序列($m=2^n$),$M$ 为 $2^n\times2^n$ 的小波变换矩阵, $Y_1^T=[y_{1,1},y_{1,2},\ldots,y_{1,m}]^T=M\times X_1^T$ 且 $Y_2^T=[y_{2,1},y_{2,2},\ldots,y_{2,m}]^T=M\times X_2^T$,$X_1$ 和 X_2 的欧氏距离 $d(X_1,X_2)=((x_{1,1}-x_{2,1})^2+(x_{1,2}-x_{2,2})^2+\ldots+(x_{1,m}-x_{2,m})^2)^{1/2}$,$Y_1$ 和 Y_2 的欧氏距离 $d(Y_1,Y_2)=((y_{1,1}-y_{2,1})^2+(y_{1,2}-y_{2,2})^2+\ldots+(y_{1,m}-y_{2,m})^2)^{1/2}$,则 $d(X_1,X_2)=d(Y_1,Y_2)$.

证明:因为 $Y_1^T=[y_{1,1},y_{1,2},\ldots,y_{1,m}]^T=M\times X_1^T$, $Y_2^T=[y_{2,1},y_{2,2},\ldots,y_{2,m}]^T=M\times X_2^T$,

则有 $X_1^T=M^T\times Y_1^T$, $X_2^T=M^T\times Y_2^T$, $X_1=Y_1\times M$, $X_2=Y_2\times M$.

又因为 $d^2(X_1,X_2)=(x_{1,1}-x_{2,1})^2+(x_{1,2}-x_{2,2})^2+\ldots+(x_{1,m}-x_{2,m})^2$, $M\times M^T=E$(E 为单位矩阵),

所以

$$\begin{aligned}d^2(X_1,X_2)&=(X_1-X_2)\times(X_1-X_2)^T\\&=(Y_1\times M-Y_2\times M)\times(M^T\times Y_1^T-M^T\times Y_2^T)\\&=(Y_1-Y_2)\times(Y_1-Y_2)^T\\&=(y_{1,1}-y_{2,1})^2+(y_{1,2}-y_{2,2})^2+\ldots+(y_{1,m}-y_{2,m})^2,\end{aligned}$$

则有 $d(X_1,X_2)=d(Y_1,Y_2)$成立. □

3.3 环形嵌套滑动窗口模式

在数据流应用环境中,滑动窗口是对无限性数据流进行局部处理的重要方法之一.使用环形嵌套滑动窗口模式挖掘数据流的局部异步偶合关系,为便于处理,又不失一般性,本文假设主滑动窗口内含有连续数据点的数量为 2^{p+q},即窗口的宽度为 2^{p+q},并将每个滑动窗口划分为 2^p 个宽度为 2^q 的子窗口,主滑动窗口每次滑动步长为 2^q.如果两个数据流在相差时间 t 后呈现偶合关系,则称 t 为偶合时延.两个数据流的偶合时延 t 取 2^q 的整数倍.为了便于处理,主滑动窗口首尾相连而构成逻辑环路.当已经充满 2^{p+q} 个数据的主滑动窗口向前滑动 k 个数据点时,主窗口中的数据变为$[x_{m+1},\ldots,x_{m+k},x_{k+1},\ldots,x_m]$的形式.如图 1 所示.

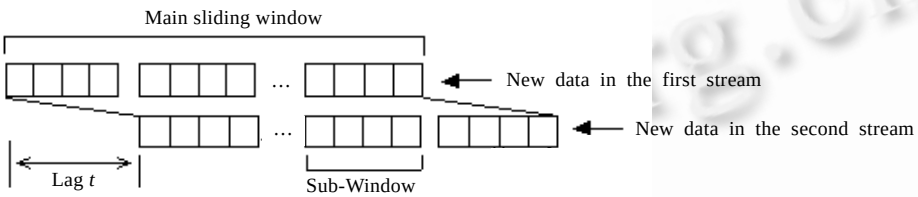

Fig.1 Double loop sliding windows

图 1 双环形滑动窗口

算法 **1**(小波系数的滑动算法). SubwinDWT(WinData[],Wsize,Coef[]).

输入:滑动窗口的数据 WinData[],子窗口的宽度 Wsize;

输出:小波系数 Coef[].

```
(1) {   int j=Wsize/2,k=0,i;
(2)     while (j>=1)
(3)     {   for (i=0;i<j;i++)
(4)           {Coef[k]=(WinData[2*i]–WinData[2*i+1])/2^1/2;
(5)              WinData[i]=(WinData[2*i]+DinData[2*i+1])/2^1/2,k=k+1;}
(6)           j=j/2,Coef[k]=0;}   //因为最后一个小波系数与计算偶合度无关
(7)   Return Coef[];}
```

算法 1 是计算滑动窗口数据的 Haar 小波变换的系数,语句(4)主要用于计算$(x_{2i}–x_{2i+1})/2^{1/2}$;而语句(5)用于计算$(x_{2i}+x_{2i+1})/2^{1/2}$.为了充分利用内存空间,将$(x_{2i}+x_{2i+1})/2^{1/2}$ 的结果暂存于 *WinData*[]的前 $j/2$ 元素中.例如,假如 *WinData*[]中的值为[3,3,6,8],当语句(4)~语句(5)循环 j=2 次后,*Coef*[]中的前 2 个值分别为$[0,–2^{1/2}]$,*WinData*[]中的值分别为$[3\times2^{1/2},3\times2^{1/2},6,8]$;当退出语句(2)~语句(6)的循环语句时,*Coef*[]中的值分别为$[0,–2^{1/2},–4,10]$.空间开销与子窗口宽度呈线性关系,近似为子窗口宽度的 3 倍.

3.4 小波系数的增量算法

经过一定时间之后,如果两个子序列的数据部分更新,在此将讨论利用原有的小波系数,计算更新后的新序列的小波系数,以达到减少计算时间的目的.为了方便讨论问题,假定两个系列的时延 t=0,给出如下相关推论:

推论 **3.1**. 设 $X_1=[x_{1,1},x_{1,2},\ldots,x_{1,m}]$和 $X_2=[x_{2,1},x_{2,2},\ldots,x_{2,m}]$是两个流序列($m=2^n$),新序列 $X_1^0=[x_{1,p+1},\ldots,x_{1,m},x_{1,1},\ldots,x_{1,p}]$和 $X_2^0=[x_{2,P+1},\ldots,x_{2,m},x_{2,1},\ldots,x_{2,p}]$,则 $Corr(X_1,X_2,0)=Corr(X_1^0,X_2^0,0)$.

证明:较为容易证明,此处从略. □

推论 **3.2**. 设 $Y=[y_1,y_2,\ldots,y_m]$($m=2^n$)是流序列 $X=[x_1,x_2,\ldots,x_m]$的小波系数,$Y^0=[y_1^0,y_2^0,\ldots,y_m^0]$是新序列 $X^0=[x_{m+1},\ldots,x_{m+p},x_{p+1},\ldots,x_m]$的小波系数,$M$ 为 $2^n\times2^n$ 的小波变换矩阵,则 $Y^0=Y+M\times[x_{m+1}–x_1,\ldots,x_{m+p}–x_p,0,\ldots,0]^T$.

证明:因为较为容易证明,此处从略. □

算法 **2**(小波系数的增量算法). IncrDWT(*WinData*[],*Coef*[],*k*,*First*,*Wsize*).

输入:主窗口的原 *WinData*[],原小波系数 *OldCoef*[],有 *k* 个数据更新,*First* 记录存储窗口数据更新的起始位置,窗口的宽度 *Wsize*;

输出:小波系数 *NewCoef*[],存储更新后的窗口数据的数组 *WinData*[]. //注意,*WinData*[]采用逻辑环形结构

(1) {将 *IncrData*[]中的数据清零;

(2) 将 *WinData*[]中从 *First* 开始的 *k* 个数据复制到 *IncrData*[]中从 *First* 开始的 *k* 个位置;

(3) 更新 *WinData*[]中从 *First* 开始的 *k* 个数据;

(4) *IncrData*[*First*…*First*+*k*−1]=*IncrData*[*First*…*First*+*k*−1]−*WinData*[*First*…*First*+*k*−1];

(5) *SubwinDWT*(*IncrData*[],*Wsize*,*IncrCoef*[]); //语句(4)、语句(5)体现了推论 3.1

(6) *Coef*[]=*Coef*[]+*IncrCoef*[]; //该处体现了推论 3.2

(7) Return *Coef*[];

(8) }

在推论 3.2 的理论指导下,算法 2 描述了在原有序列的小波系数的基础上,计算滑动距离为 *k* 的新序列的小波系数的增量算法.

推论 **3.3**. 设 $X_1=[x_{1,p+1},\ldots,x_{1,m},x_{1,m+1},\ldots,x_{1,m+p}]$和 $X_2=[x_{2,p+1},\ldots,x_{2,m},x_{2,m+1},\ldots,x_{2,m+p}]$是两个流序列($m=2^n$),$Y_1=[y_{1,1},y_{1,2},\ldots,y_{1,m}]$和 $Y_2=[y_{2,1},y_{2,2},\ldots,y_{2,m}]$分别是序列 $X_1^0=[x_{1,m+1},\ldots,x_{1,m+p},x_{1,p+1},\ldots,x_{1,m}]$和 $X_2^0=[x_{2,m+1},\ldots,x_{2,m+p},x_{2,p+1},\ldots,x_{2,m}]$的小波系数,则 X_1 和 X_2 的同步偶合度可由 Y_1 和 Y_2 计算,即为

$$Corr(X_1,X_2,0)=\frac{\sum_{i=1}^{m-1}(y_{1,i}\times y_{2,i})}{\sqrt{\sum_{i=1}^{m-1}y_{1,i}^2\times\sum_{i=1}^{m-1}y_{2,i}^2}}.$$

证明:由推论 3.1 和定理 3.2 容易证明,此处从略. □

4 异步偶合度的抗噪声算法

为了简化小波系数的滤波阈值α的确定,在此对小波系数序列进行单位规范化处理,使规范化的序列数据取值在−1 到 1 之间,且平方和等于 1,α取接近于 0 的值.为方便讨论,我们给出如下定义和定理.

定义 **4.1**(扩展小波变换). 设序列$[y_1,y_2,\ldots,y_m]$是向量 $X=[x_1,x_2,\ldots,x_m]$的小波变换系数,EDWT(X)=$Y^0=[y_1,y_2,\ldots,y_{m-1}]\times(y_1^2+y_2^2+\ldots+y_{m-1}^2)^{-1/2}$, 则称 EDWT($X$)是对 X 的扩展小波变换.

定理 **4.1**(偶合度距离化定理). 设 $m=2^n$,$X_1=[x_{1,1},x_{1,2},\ldots,x_{1,m}]$和 $X_2=[x_{2,1},x_{2,2},\ldots,x_{2,m}]$是两个流序列,经 EDWT($X_1$)= Y_1^0 和 EDWT(X_2)= Y_2^0 变换后,$d(Y_1^0,Y_2^0)$表示 Y_1^0 和 Y_2^0 的欧氏距离,则有

$$Corr(X_1,X_2,0)=1-d^2(Y_1^0,Y_2^0)/2.$$

证明:显然有$\|Y_1^0\|=\|[y_{1,1},y_{1,2},\ldots,y_{1,m-1}]\times(y_{1,1}^2+y_{1,2}^2+\ldots+y_{1,m-1}^2)^{-1/2}\|=1$,$\|Y_1^0\|$表示向量 Y_1^0 的模;

$\|Y_2^0\|=\|[y_{2,1},y_{2,2},\ldots,y_{2,m-1}]\times(y_{2,1}^2+y_{2,2}^2+\ldots+y_{2,m-1}^2)^{-1/2}\|=1$,$\|Y_2^0\|$表示向量 Y_2^0 的模;

由定理 3.2 可知:$Corr(X_1,X_2,0)=\langle Y_1^0,Y_2^0\rangle$,$\langle Y_1^0,Y_2^0\rangle$为向量的内积,

$$d^2(Y_1^0,Y_2^0)=\|Y_1^0-Y_2^0\|^2=\|Y_1^0\|^2-2\times\langle Y_1^0,Y_2^0\rangle+\|Y_2^0\|^2=2-2\times\langle Y_1^0,Y_2^0\rangle=2-2Corr(X_1,X_2,0),$$

所以,$Corr(X_1,X_2,0)=1-d^2(Y_1^0,Y_2^0)/2$ 成立. □

定义 **4.2**. 设 X_1 和 X_2 是在局域窗口上的有序流序列($\varepsilon\geq0$),如果 $Corr(X_1,X_2,t)\geq1-\varepsilon^2/2$,则称序列 X_1 和 X_2 为ε-局域正偶合;如果 $Corr(X_1,X_2,t)\leq-(1-\varepsilon^2/2)$,则称序列 X_1 和 X_2 为ε-局域负偶合;否则为ε-局域弱偶合.

由上述定理 4.1 和定义 4.1 可知,算法 3 的主要思想为:对小波系数进行单位规范化,在给定阈值α下进行消除噪声处理后,计算两个规范系数序列的欧氏距离.

算法 **3**(局域系数距离算法). LocalDistance($coef_1$[],$coef_2$[],α) //按照定义 4.1 和定理 4.1 计算局域系数距离

输入:两个数据流局域小波变换系数 $coef_1[],coef_2[]$;小波变换滤波阈值α;

输出:扩展小波变换系数距离 *Distance*.

(1) { double $CoefSum_1$=0,$CoefSum_2$=0; //分别用于存储两个流的窗口小波系数的平方和

(2) double *Distance*; //用于存储扩展小波系数的欧氏距离

(3) for (i=0;i<*Wsize*−1;i++) //计算窗口小波系数的平方和

(4) {$CoefSum_1=CoefSum_1+coef_1[i]\times coef_1[i]$,$CoefSum_2=CoefSum_2+coef_2[i]\times coef_2[i]$;}

(5) for (i=0;i<*Wsize*−1;i++) //计算窗口对滤波后的偶合度

(6) {使用窗口小波系数的平方和 $CoefSum_1$,$CoefSum_2$ 对小波系数单位化;

(7) if 单位化的小波系数<α then 将相应的小波系数置 0;}

(8) 计算扩展小波系数的欧氏距离 *Distance*;

(9) Return *Distance*;

(10) }

定理 **4.2**(局域偶合判定定理). 设 X_1 和 X_2 是两数据流在局域窗口上的有序序列,如果给定$\varepsilon\geq 0$,经 $EDWT(X_1)=Y_1^0$ 和 $EDWT(X_2)=Y_2^0$ 变换后,则有:

(1) $Corr(X_1,X_2,0)\geq 1-\varepsilon^2/2\Leftrightarrow d(Y_1^0,Y_2^0)\leq\varepsilon$;

(2) $Corr(X_1,X_2,0)\leq -(1-\varepsilon^2/2)\Leftrightarrow d(-Y_1^0,Y_2^0)\leq\varepsilon$ 或 $d(Y_1^0,-Y_2^0)\leq\varepsilon$.

证明:(1) 由定理 4.1 可知:$Corr(X_1,X_2,0)=1-d^2(Y_1^0,Y_2^0)/2$,则 $Corr(X_1,X_2,0)\geq 1-\varepsilon^2/2\Leftrightarrow d(Y_1^0,Y_2^0)\leq\varepsilon$;

(2) 由定理 3.2 和定理 4.1 可知:$Corr(X_1,X_2,0)=\langle Y_1^0,Y_2^0\rangle$,$\langle Y_1^0,Y_2^0\rangle$为向量的内积,

由小波变换的原理可知:$EDWT(-X_1)=-Y_1^0\Leftrightarrow -Corr(X_1,X_2,0)=Corr(-X_1,X_2,0)=\langle -Y_1^0,Y_2^0\rangle$,

由定理 4.1 知:$Corr(-X_1,X_2,0)=1-d^2(-Y_1^0,Y_2^0)/2$,

$Corr(X_1,X_2,0)\leq -(1-\varepsilon^2/2)\Leftrightarrow Corr(-X_1,X_2,0)\geq 1-\varepsilon^2/2\Leftrightarrow 1-d^2(-Y_1^0,Y_2^0)/2\geq 1-\varepsilon^2/2\Leftrightarrow d(-Y_1^0,Y_2^0)\leq\varepsilon$;

同理可证 $d(Y_1^0,-Y_2^0)\leq\varepsilon$. □

下述算法 4 主要使用扩展小波系数距离计算两个数据流的局域异步偶合度的方法.其主要特点是:(1) 使用了两个逻辑上为环形的数组构成两个环形窗口;(2) 如果是强偶合关系,则两个环形的窗口同时滑动更新,如程序的语句(9)~语句(12);(3) 如果是弱偶合关系,则只有辅助窗口滑动更新,如语句(14)、语句(15);(4) 如果数据流偶合时延达到指定的最大值,则交换两个窗口,其程序语句为(17)、(18).假设窗口数为 n,子窗口宽度为 *SubWsize*,滑动窗口滑动步长为 *SubWsize*;时间复杂度为 $O(n*SubWsize)$.

算法 **4**(异步滑动算法). AsynCoin(*Wsize*,*SubWsize*,α,ε,max*lag*).

输入:主窗口的宽度 *Wsize*,子窗口的宽度 *SubWsize*,滤波阈值α,距离阈值ε,max*lag* 指定最大时延;

输出:局域偶合度 *LocalCoin* 和时延 *lag*.

(1) 第 1 个数据流的数据充满主窗口 $Windata_1[]$; //$Windata_1[]$数组为逻辑循环结构

(2) 第 2 个数据流的数据充满主窗口 $Windata_2[]$; //$Windata_2[]$数组为逻辑循环结构

(3) *SubwinDWT*($WinData_1[]$,*Wsize*,$Coef_1[]$); //在局域窗口上对第 1 个流小波变换

(4) *SubwinDWT*($WinData_2[]$,*Wsize*,$Coef_2[]$); //在局域窗口上对第 2 个流小波变换

(5) 初始化更新数据窗口的起始位置 $first_1$=0 和 $first_2$=0,并初始化偶合时延 *lag*=0;

(6) for All data in Streams

(7) { D=*LocalDistance*($coef_1[],coef_2[],\alpha$); //计算局域扩展小波系数距离

(8) if $D<\varepsilon$ //根据定理 4.2 扩展小波系数距离大于阈值,两个流的窗口同时滑动 k.

(9) {计算并记录某时刻的局域偶合度 $LocalCoin=1-0.5\times D^2$ 和时延; //根据定理 4.1 计算局域偶合度.

(10) $first_1=(first_1+k)$ mod *Wsize*; $first_2=(first_2+k)$ mod *Wsize*; //窗口滑动步长 k,并计算更新的起始点.

(11) *IncrDWT*($WinData_1[]$,$Coef_1[]$,*SubWsize*,$first_1$,*Wsize*); //应用推论 3.1 和推论 3.2 增量式计算小波

系数

(12) $IncrDWT(WinData_2[],Coef_2[],SubWsize,first_2,Wsize);\}$ //同语句(11)

(13) else if $lag<=\max lag$ //两个数据流偶合时延小于给定的最大时延

(14) $\{first_2=(first_2+k) \bmod Wsize, lag=lag+1;$ //第 1 个窗口不滑动,滑动第 2 个窗口

(15) $IncrDWT(WinData_2[],Coef_2[],SubWsize,first_2,Wsize);\}$ //同语句(11)

(16) else //在第 2 个窗口长时间与第 1 个窗口的弱偶合

(17) {将第 1 个窗口与第 2 个窗口重新对齐,将两个窗口互换;

(18) 滑动第 2 个窗口并进行增量小波变换;}

(19) }

5 实验及性能分析

该算法的模拟实验采用深圳证券交易所和上海证券交易所的股票日交易数据.实验平台和主要参数如下:(1) CPU 为 PIII600;(2) 内存为 256M;(3) 操作系统使用 Windows2000;(4) 使用 1990 年~2005 年 2 月的 2 157 支股票的日数据,记录数为 250 万条左右;(5) 数据存储方式为 SQL SERVER2000 数据库.主要进行了如下实验:

5.1 不同滤波阈值α的偶合度比较

我们研究了在给定不同滤波阈值α下的局域偶合度,将小波系数序列进行去噪声处理的偶合度与使用原始数据计算的偶合度进行比较.实验假定主窗口大小为 512,子窗口大小为 64,滤波阈值α分别取 0.025,0.05,0.075,0.1 时,对 200 多对数据流进行研究.图 2 给出了任意一对数据流在不同的滑动窗口中,当α不同时的偶合度.从图 2 可以初步看出:当α较小时,由小波方法滤波处理后,使用小波系数计算偶合度与原始数据计算的偶合度较为接近,甚至相等;随着α的增加,与原始数据计算的偶合度的偏差呈增大的趋势.

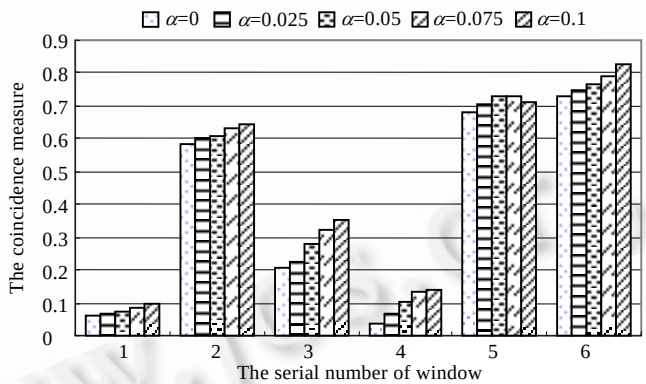

Fig.2 The coincidence measure versus α

图 2 不同α时,偶合度之间的比较

5.2 运行时间效率

我们对算法的运行效率主要进行了如下两个方面的实验:

(1) 当主滑动窗口选定 512 时,子窗口大小(即滑动)分别取 64,128,256;在滤波阈值α=0.05 的条件下,数据序列的平均长度为 3 200 左右,对 1 078 对数据序列分组模拟计算偶合度.从图 3 可以看出:随着数据流序列数的增加而时间开销逐步增加;时间开销随着子窗口的大小增加而减少.

(2) 实验比较了直接使用经小波算法去噪声后,系数计算偶合度与重构数据信息后计算偶合度占用的时间.当主滑动窗口选取 512 时,子窗口大小取 128.由图 4 可知,用小波系数重构数据流信息计算偶合度的时间大约是小波系数直接计算偶合度的 1.5 倍.

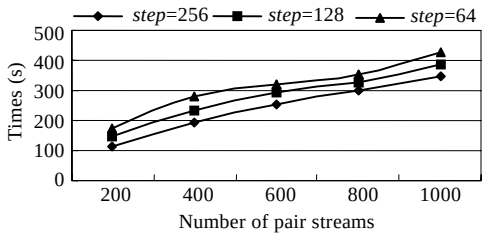

Fig.3 Time in different sliding step

图 3 不同滑动步长时,计算时间的比较

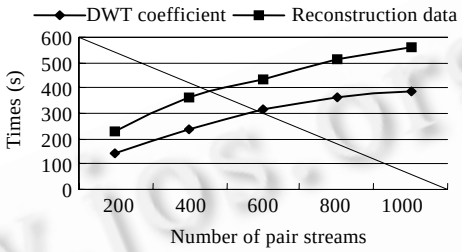

Fig.4 Times versus DWT coefficient and reconstruction data

图 4 小波系数和重构的数据的比较

5.3 平均偶合时延t与滑动步长

平均偶合时延是由数据流在局域窗口发生强偶合时的时延之和除以强偶合的次数而得到的.实验的主滑动窗口取 256,滤波阈值α=0.025,最小偶合度为 0.8,数据流为 1 078 对.滑动步长取 4,8,16,32,64 和 128,分别进行实验.结果表明,滑动步长的变化将影响计算异步平均偶合时延的准确性.图 5 说明:平均偶合时延与实际平均偶合时延的平均距离随着滑动步长的增大而有所增加;当滑动步长较小时,偶合时延的平均值更接近真实的平均偶合时延.

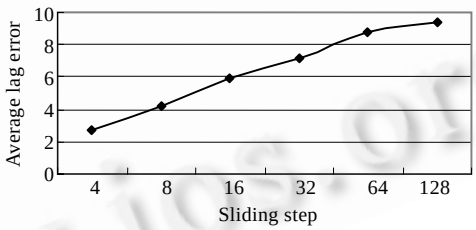

Fig.5 Sliding step versus average lag

图 5 滑动步长对平均时延的影响

6 总 结

本文研究了融合 Haar 小波技术、挖掘多数据流的异步偶合模式的方法;证明了一系列定理和相关推论,从理论上保证了使用小波系数计算数据流的异步偶合度的正确性,为从小波方法压缩的流数据中发现偶合模式提供了理论依据.本文还设计了基于环形滑动窗口的增量式的抗噪声算法.实验表明,该算法不但能消除数据流中的高频噪声,而且能够直接从小波技术压缩的流数据中高效地挖掘数据流的局域偶合关系.

References:

[1] Teng WG, Chen MS, Yu PS. A regression-based temporal pattern Ming scheme for data streams. In: Freytag JC, Lockemann PC, eds. Proc. of the 29th Int'l Conf. on Very Large Data Bases (VLDB 2003). Berlin: Morgan Kaufmann Publishers, 2003. 93−104.

[2] Ben-David S, Gehrke J, Kifer D. Detecting change in data streams. In: Nascimento MA, Kossmann D, eds. Proc. of the 30th Int'l Conf. on Very Large Data Bases (VLDB 2004). Toronto: Morgan Kaufmann Publishers, 2004. 180−191.

[3] Yu JX, Chong ZH, Lu HJ, Zhou AY. False positive or false negative: Mining frequent Itemsets from high speed transactional data streams. In: Nascimento MA, Kossmann D, eds. Proc. of the 30th Int'l Conf. on Very Large Data Bases (VLDB 2004). Toronto: Morgan Kaufmann Publishers, 2004. 204−215.

[4] Datar M, Gionis A, Indyk P, Motwani R. Maintaining stream statistics over sliding windows. In: Eppstein D, ed. Proc. of the 13th Annual ACM-SIAM Symp. on Discrete Algorithms. San Francisco: ACM Press, 2002. 635−644.

[5] Gehrke J, Korn F, Srivastava D. On computing correlated aggregates over continual data streams. Walid GA, ed. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2001. 13−24.

[6] Rafiei D, Mendelzon A. Similarity-Based queries for time series data. In: Peckham J, ed. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. Tucson: ACM Press, 1997. 13−25.

[7] Dula S, Kim C, Shim K. XWAVE: Optimal and approximate extended wavelets for streaming data. In: Nascimento MA, Kossmann D, eds. Proc. of the 30th Int'l Conf. on Very Large Data Bases (VLDB 2004). Toronto: Morgan Kaufmann Publishers, 2004. 288−299.

[8] Gilbert AC, Kotidis Y, Muthukrishnan S, Strauss MJ. Surfing wavelets on streams: One-Pass summaries for approximate aggregate queries. In: Apers PMG, Atzeni P, eds. Proc. of the 27th Int'l Conf. on Very Large Data Bases (VLDB 2001). Roma: Morgan Kaufmann Publishers, 2001. 79−88.

[9] Zhu YY, Shasha D. StatStream: Statistical monitoring of thousands of data streams in real time. In: Bressan S, Chaudhri AB, eds. Proc. of the 28th Int'l Conf. on Very Large Data Bases (VLDB 2002). Hong Kong: Springer-Verlag, 2002. 358−369.

[10] Burrus CS, Gopinath RA, Guo HT. Introduction to Wavelets and Wavelet Transform. Englewood Cliffs: Prentice Hall, 1998. 1−145.

陈安龙(1971－),男,四川仪陇人,博士,主要研究领域为数据挖掘.

彭京(1973－),男,博士,主要研究领域为数据挖掘.

唐常杰(1946－),男,教授,博士生导师,CCF高级会员,主要研究领域为数据挖掘.

胡建军(1970－),男,博士,主要研究领域为数据挖掘.

元昌安(1964－),男,博士,教授,主要研究领域为数据库,空间数据挖掘.