ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.2, February 2007, pp.177−184
DOI: 10.1360/jos180177

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

基于小波和偶合特征的多数据流压缩算法*

陈安龙[1,2+], 唐常杰[1], 元昌安[1,3], 朱明放[1], 段 磊[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)

A Compression Algorithm for Multi-Streams Based on Wavelets and Coincidence

CHEN An-Long[1,2+], TANG Chang-Jie[1], YUAN Chang-An[1,3], ZHU Ming-Fang[1], DUAN Lei[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, Zhu MF, Duan L. A compression algorithm for multi-streams based on wavelets and coincidence. *Journal of Software*, 2007,18(2):177−184. http://www.jos.org.cn/1000-9825/18/177.htm

Abstract: Methods based on Haar wavelets and coincidence characteristics are proposed to compress multi-streams. The main contributions include: (1) Energy conservation law of Haar wavelets transform is proved to compress data streams. (2) The relation between the coincidence measure and trend of streams is revealed as along with the invariability under parallel shift and the equivalence law over coincidence measure to approximately express data-streams by the wavelet coefficient of the characteristic stream and its energy. (3) Multi-Scales energy decomposition model is proposed to improve the compression precision. (4) The multi-scales compression algorithm and the energy conservation reconstruction algorithm are designed. (5) Extended experiments show that the compression ratio of the new methods is 2~4 times as the traditional method.

Key words: data stream; Haar wavelet; coincidence characteristic; data compression; hierarchy decompose

摘 要: 提出了基于 Haar 小波技术和偶合特征的多数据流压缩方法.主要研究成果包括:(1) 证明了 Haar 小波变换服从能量守恒规律,并用于压缩数据流;(2) 揭示了数据流的偶合度与变化趋势的相关性、偶合度的平移不变性及等价规律,采用特征流序列的小波系数和流能量近似表示流的趋势,达到压缩的目的;(3) 提出了多尺度能量分解模型,提高了表示精度;(4) 设计了多尺度能量分解压缩算法以及多尺度重构算法;(5) 在真实数据集上的实验表明,新方法的压缩比是传统小波方法的 2~4 倍.

关键词: 数据流;Haar 小波;偶合特征;数据压缩;层次分解

* Supported by the National Natural Science Foundation of China under Grant Nos.60473071, 10476006 (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 (国家教育部博士点专项基金)

Received 2006-03-08; Accepted 2006-06-30

中图法分类号: TP391 文献标识码: A

网络技术的迅速发展催生了大量的流数据,例如通话记录、股票交易数据、交通监测数据、气象监测数据、传感器数据等.相对于传统数据库的静态数据,数据流呈现出快速变化、海量无限、实时连续等特点,传统的存储技术难以满足快速变化的海量流数据的存储需求.为了掌握数据流的历史信息,研究描述数据流信息的方法具有重要意义.文献[1,2]研究了用小波变换压缩数据流的方法,仅存储少量的小波系数来表征数据流,而此方法仅针对单个数据流压缩,没有研究对多个数据流压缩的方法.某些数据流之间可能存在不同程度的偶合,研究如何利用数据流之间的偶合特征,对多数据流进行压缩具有重要意义,这正是本文的研究内容与文献[1,2]的重要区别之一;文献[3]运用离散的傅立叶变换技术研究了多数据流之间的偶合关系,但没有研究数据流的历史信息的存储方法;文献[4]研究了用小波技术消除数据流的噪声,在不重构数据流原貌的条件下,运用小波系数发现多数据流之间的偶合关系,但该方法没有研究如何运用数据流的偶合特征有效地管理流数据;文献[5]在不存储数据流的历史信息的情况下,发现了多数据流的偶合关系以及探测数据流的潜在跳变规律,但没有考虑对历史描述信息的存储;文献[6]研究了发现异步偶合特征的方法;文献[7]研究了数据流的实时在线查询操作,但没有考虑对历史信息的检索.在实际应用中,可能需要掌握某些历史阶段的信息,我们在文献[4]中已经研究了数据流的偶合关系,为了弥补已有研究的不足,本文提出一种新的数据流压缩方法.

1 本文的主要贡献

本文融合 Haar 小波技术[1,8]于数据流的压缩处理,利用多数据流之间的偶合特征,探索了多数据流压缩方法,做了如下研究工作:(1) 证明了对数据流序列的小波变换服从能量守恒规律,并运用于多数据流压缩处理;(2) 揭示了数据流之间的偶合度和变化趋势的相关性.在多数据流环境中,当两个数据流的偶合达到一定程度时,两个数据流变化趋势具有极大的相似性.使用代表变化规律的特征流序列的小波系数近似表示与之强偶合的数据流的变化趋势,此时,仅存储数据流序列的能量,以达到压缩数据流的目的.用数据流的能量和特征流序列的小波系数可重构数据流.传统方法是对每个数据流分别压缩,而新方法是将多个偶合数据流作为整体压缩,以提高压缩率;(3) 给出了按尺度分解小波能量的算法.借鉴小波分解的层次特征,将数据流能量在不同尺度下分解,按尺度分别存储,设计了多尺度的压缩算法,提高了压缩精度.设计了基于能量守恒的多尺度重构算法,使用不同尺度下的小波系数与相应的能量进行合成并重构数据流.融合能量分解和偶合相似性原理的多数据流的压缩方法是本文的重要创新点;(4) 用真实数据进行了实验,验证了算法的有效性.

2 多数据流的压缩存储

2.1 基本思想方法

在多数据流环境中,数据流之间存在不同程度的偶合关系.数据流的偶合程度反映了数据流变化趋势的相似度.我们将研究利用数据流之间的偶合特征对多数据流进行压缩的新方法.为便于讨论,先给出形式化概念:

定义 **2.1**(局域偶合度). 设 $X_1=[x_{1,1},x_{1,2},\ldots,x_{1,n}]$和 $X_2=[x_{2,1},x_{2,2},\ldots,x_{2,n}]$分别为数据流 S_1 和 S_2 中的具有 n-数据点的子序列,则子序列 X_1 和 X_2 的局域偶合度为 $Corr(X_1,X_2)=L_{12}\Big/\sqrt{L_1\times L_2}$, 其中, $L_1=\sum_{i=1}^{n}(x_{1,i}-\overline{X}_1)^2$; $L_2=\sum_{i=1}^{n}(x_{2,i}-\overline{X}_2)^2$; $L_{12}=\sum_{i=1}^{n}(x_{1,i}-\overline{X}_1)(x_{2,i}-\overline{X}_2)$; $\overline{X}_1=\frac{1}{n}\sum_{i=1}^{n}x_{1,i}$; $\overline{X}_2=\frac{1}{n}\sum_{i=1}^{n}x_{2,i}$.

定义 **2.2**. 设 X_1 和 X_2 是两个数据流序列,对于给定的正数$\varepsilon\le 1$,如果 $Corr(X_1,X_2)\ge\varepsilon$,则称序列 X_1 和 X_2 为ε-正偶合;如果 $Corr(X_1,X_2)\le-\varepsilon$,则称序列 X_1 和 X_2 为ε-负偶合;否则,称序列 X_1 和 X_2 为ε-弱偶合.

定义 **2.3**(特征流序列). 设 S 是由多个数据流构成的序列集合,如果存在 $X_0\in S$,对于$\forall X_k\in S$ 以及给定的正数 $\varepsilon\le 1$,满足$|Corr(X_0,X_k)|\ge\varepsilon$,则称 X_0 为集合 S 的特征流序列,简称特征流;X_k 为偶合流序列.

定义 **2.4**(流序列能量). 设 $X_k=[x_{k,1},x_{k,2},\ldots,x_{k,n}]$为数据流序列构成的向量,$E(X_k)=\|X_k\|^2$,则称 $E(X_k)$为流序列 X_k 的能量,简称流能量.

定理 **2.1**(小波变换能量守恒). 设 $X_k=[x_{k,1},x_{k,2},\ldots,x_{k,n}]$为数据流序列构成的向量,$Y_k=[y_{k,1},y_{k,2},\ldots,y_{k,n}]$是由小波变换[1,8]得到的系数构成的向量,则 $E(X_k)=E(Y_k)$.换言之,流序列在小波变换下服从能量守恒定律(证明略).

在上述概念及性质的基础上,为了近似表示数据流,本文取数据流的两个主要特征:(1) 数据流的变化规律;(2) 数据流内含的流序列能量.数据流变化规律使用与之强偶合的特征流序列近似表示,仅存储特征流序列的小波系数和数据流能量,以达到压缩数据流的目的.具体算法与数据流之间是ε-正偶合还是ε-负偶合,以及数据流中的数据点的正负有关,下面将分不同情况加以讨论.

2.2 ε-正偶合的流序列

在多数据流环境中,如果数据流之间是ε-正偶合,则说明数据流的变化规律呈现一定程度的相似性,可以用一个流的规律近似表征另一个数据流.根据流能量的定义,任何数据流的能量非负数.为了简化问题的讨论,假定两个呈现ε-正偶合的数据流,根据流数据序列的数值特征分别进行讨论:

2.2.1 非负数流序列

如果数据流中数据点均为非负数,称这样的数据流为非负数流序列.假定表现正偶合的两个数据流均为非负数流序列,其中一个为特征流序列.使用特征流序列的变化模式表示与之偶合的流序列的变化模式.根据定理 2.1,小波系数所表征的能量和原数据流蕴涵的能量相等;文献[8]研究表明,流序列的大部分能量分布在少数的小波系数中.因此,对非负ε-正偶合流序列集压缩的主要思想是:(1) 对特征流序列进行 Haar 小波变换,将小于给定阈值的小波系数置 0;(2) 计算偶合流和特征流序列的能量比;(3) 存储特征流序列的非 0 小波系数和能量比;(4) 在重构偶合流序列时,用能量比乘以特征流的小波系数,然后用重构算法得到近似偶合流序列.

2.2.2 带有负数的流序列

在实际应用中,许多流数据可能为负数,使用前面介绍的非负数流的压缩思想,在重构时不能近似再现原数据流的信息.如图 1 中数据流 S_0 和 S_1,假定 S_0 是特征流序列,S_1 经前面的压缩思想处理后,使用非负数的数据流重构得到序列 S_2,从图 1 可以看出,S_1 和 S_2 是关于时间轴 t 对称的.其主要原因是:特征流序列在时间轴 t 的上方波动变化,数据流 S_1 的能量为正数,按照上述重构方法得到时间轴 t 上方的 S_2.为了解决该问题,下面使用坐标平移的方法,使每个数据流序列为正数;然后使用非负数的流序列的压缩和重构思想,再进行平移则可得到流序列的近似序列.为了说明平移不影响数据序列之间的偶合度,给出如下定理.

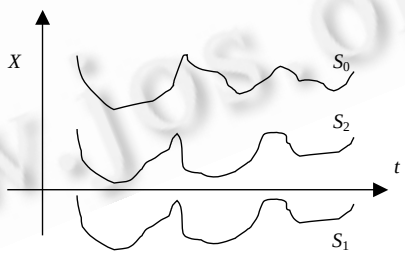

Fig.1 An example of stream with negative

图 1 数据流中含有负数的例子

定理 **2.2**(偶合度的平移不变性). 设 $X_1=[x_{1,1},x_{1,2},\ldots,x_{1,n}]$和 $X_2=[x_{2,1},x_{2,2},\ldots,x_{2,n}]$为流序列向量,分别平移 p 和 q 后得到 $X_1'=[x_{1,1}-p,x_{1,2}-p,\ldots,x_{1,n}-p]$和 $X_2'=[x_{2,1}-q,x_{2,2}-q,\ldots,x_{2,n}-q]$,则 $Corr(X_1,X_2)=Corr(X_1',X_2')$.

证明:因为 $\overline{X}_1=\frac{1}{n}\sum_{i=1}^{n}x_{1,i}$, $\overline{X}_2=\frac{1}{n}\sum_{i=1}^{n}x_{2,i}$,则 $\overline{X}_1'=\frac{1}{n}\sum_{i=1}^{n}(x_{1,i}-p)=\overline{X}_1-p$, $\overline{X}_2'=\frac{1}{n}\sum_{i=1}^{n}(x_{2,i}-q)=\overline{X}_2-q$.

又由 $L_1=\sum_{i=1}^{n}(x_{1,i}-\overline{X}_1)^2$, $L_2=\sum_{i=1}^{n}(x_{2,i}-\overline{X}_2)^2$, $L_{12}=\sum_{i=1}^{n}(x_{1,i}-\overline{X}_1)(x_{2,i}-\overline{X}_2)$,

可知 $L_1'=\sum_{i=1}^{n}((x_{1,i}-p)-\overline{X}_1')^2=L_1$, $L_2'=\sum_{i=1}^{n}((x_{2,i}-q)-\overline{X}_2')^2=L_2$, $L_{12}'=\sum_{i=1}^{n}((x_{1,i}-p)-\overline{X}_1')((x_{2,i}-q)-\overline{X}_2')=L_{12}$,

所以, $Corr(X_1,X_2)=L_{12}\Big/\sqrt{L_1\times L_2}=L_{12}'\Big/\sqrt{L_1'\times L_2'}=Corr(X_1',X_2')$ 成立. □

定理 2.2 揭示了平移对数据流序列之间的偶合度不产生影响的规律,利用该规律可以将含有负数的数据流序列进行平移转化为非负数的流序列,然后对新生成的流序列进行压缩处理.在重构偶合流序列时,先使用能量比乘以虚拟流序列的小波系数,然后用重构算法得到近似的流序列,并根据平移参数进行反向平移还原.

2.3 ε-负偶合的流序列

前面讨论了ε-正偶合流序列的存储和重构思想,但是在实际的应用环境中,某些数据流间表现为ε-负偶合,则两个数据流具有相反的变化规律.定理 2.3 揭示了ε-负偶合流的压缩规律.

定理 **2.3**(偶合度的等价规律). 设 $X_1=[x_{1,1},x_{1,2},\ldots,x_{1,n}]$和 $X_2=[x_{2,1},x_{2,2},\ldots,x_{2,n}]$为流序列构成的向量,如果 $0\le\varepsilon\le1$,则有:(1) $Corr(X_1,X_2)\ge\varepsilon$;(2) $Corr(-X_1,X_2)\le-\varepsilon$;(3) $Corr(X_1,-X_2)\le-\varepsilon$;(4) $Corr(-X_1,-X_2)\ge\varepsilon$相互等价.

证明:因为 $\overline{X}_1=\frac{1}{n}\sum_{i=1}^{n}x_{1,i}$, $\overline{X}_2=\frac{1}{n}\sum_{i=1}^{n}x_{2,i}$,则 $\overline{X}_1'=-\overline{X}_1=\frac{1}{n}\sum_{i=1}^{n}(-x_{1,i})$, $\overline{X}_2'=-\overline{X}_2=\frac{1}{n}\sum_{i=1}^{n}(-x_{2,i})$.

又由 $L_1'=\sum_{i=1}^{n}((-x_{1,i})-\overline{X}_1')^2=L_1$, $L_2'=\sum_{i=1}^{n}((-x_{2,i})-\overline{X}_2')^2=L_2$, $L_{12}'=\sum_{i=1}^{n}((-x_{1,i})-\overline{X}_1')(x_{2,i}-\overline{X}_2)=-L_{12}$,

所以,

$$Corr(-X_1,X_2)=L_{12}'\Big/\sqrt{L_1'\times L_2}=-L_{12}\Big/\sqrt{L_1\times L_2}=-Corr(X_1,X_2) \quad ①$$

又因为 $L_{12}''=\sum_{i=1}^{n}((-x_{1,i})-\overline{X}_1')(x_{2,i}-\overline{X}_2)=\sum_{i=1}^{n}(x_{1,i}-\overline{X}_1)((-x_{2,i})-\overline{X}_2')=-L_{12}$,

所以,

$$Corr(X_1,-X_2)=L_{12}''\Big/\sqrt{L_1\times L_2'}=-L_{12}\Big/\sqrt{L_1\times L_2}=-Corr(X_1,X_2) \quad ②$$

又因为 $L_{12}'''=\sum_{i=1}^{n}((-x_{1,i})-\overline{X}_1')((-x_{2,i})-\overline{X}_2')=\sum_{i=1}^{n}((-x_{1,i})-(-\overline{X}_1))((-x_{2,i})-(-\overline{X}_2))=L_{12}$,

所以,

$$Corr(-X_1,-X_2)=L_{12}'''\Big/\sqrt{L_1'\times L_2'}=L_{12}\Big/\sqrt{L_1\times L_2}=Corr(X_1,X_2) \quad ③$$

综合式①~式③可知,式(1)~式(4)相互等价. □

定理 2.3 揭示了流按时间轴对称处理后的偶合规律,如果将两个呈现ε-负偶合的流序列中的一个关于时间轴对称,那么它将与另外一个流序列之间呈现ε-正偶合.在流序列压缩中:(1) 如果某数据流与特征流呈现ε-负偶合,则将该数据流的数据取相反数;(2) 如果新的流序列含有负数,则采用含负数的流序列的压缩思想;(3) 如果新的流序列的数据为非负数,则采用非负数的流序列的压缩思想;(4) 在重构偶合流序列时,先使用能量比乘以特征流的小波系数,使用重构算法重构,并将各数据取相反数.

2.4 偶合的流序列压缩算法

综合上述分析的数据流偶合的特殊情况,下面先给出对这些特殊情况进行预处理的算法.

算法 **1**. 流的预处理:MultiStreamProcess( ).

输入:特征流序列 X_0,偶合流序列的集合$\{X_1,X_2,\ldots,X_m\}$.

输出:处理后流序列的集合.

(1) {if $\exists x\in X_0$ and $x<0$ then $X_0=X_0+b_0$ //特征流序列中含有负数,平移特征流序列

(2) for each $X_i\in\{X_1,X_2,\ldots,X_m\}$ do

(3) {记录 $Corr(X_i,X_0)$的值;

(4) if $Corr(X_i,X_0)\le-\varepsilon$ then $X_i=-X_i$; //与特征流序列呈ε-负偶合

(5) if $\exists x\in X_i$ and $x<0$ then $X_i=X_i+b_i$; //如果存在负数据点,则流序列平移 b_i 使其为正序列

(6) 记录平移量 b_i 的值}

(7) Return 经平移和对称处理后的流序列}

算法 1 是对与特征流之间呈现负偶合或者含有负数据点的数据流进行预处理,语句(1)是对含有负数据点的特征流平移;循环语句(3)~语句(6)对每个流进行预处理,重复执行 m 次,则该算法的时间复杂度为 $O(m)$.下面给出多偶合流的能量守恒压缩算法,仅存储特征数据流的小波系数、偶合数据流的能量、偶合度以及平移参数等信息.

算法 **2**. 多偶合流的能量守恒压缩算法:MSNCStorage( ).

输入:特征流序列 X_0,偶合流序列的集合$\{X_1,X_2,\ldots, X_m\}$.

输出:特征流的小波系数、流序列的 ID 号、平移参数 b_i 以及能量比 R_i.

(1) {Call for MultiStreamProcess( ); //对流序列进行预处理

(2) 对 X_0 进行小波变换,并存储大于滤波阈值的系数和平移量 b_0;

(3) 计算特征流序列 X_0 的总能量 $E(X_0)$;

(4) for each $X_i\in\{X_1,X_2,\ldots,X_m\}$ do

(5) {计算 X_i 的总能量 $E(X_i)$,以及 $E(X_i)$与 $E(X_0)$的能量比 R_i;

(6) 存储偶合流的 ID 号和能量比 R_i}}

算法 2 是运用数据流的偶合相似性原理和小波变换压缩多数据流的算法.语句(4)、语句(5)是计算偶合流与特征流序列的能量比,重复执行 m 次.当特征流序列数据点的个数 N 远大于 m 时,则时间复杂度为 $O(N)$.

算法 **3**. 偶合流重构算法:RebuildStream( ).

输入:特征流 X_0 的小波系数、偶合流的 ID、$Corr(X_i,X_0)$的值、能量比 R_i,平移参数 b_i.

输出:流序列的重构信息.

(1) {用能量比的平方根分别乘以 X_0 的小波系数得到序列 $dwt[N]$;

(2) 对序列 $dwt[N]$重构生成 X_i;

(3) if $b_i\neq 0$ then $X_i=X_i-b_i$;

(4) if $Corr(X_i,X_0)<0$ then $X_i=-X_i$;

(5) 输出流序列 X_i}

算法 3 给出了经算法 2 压缩后重构数据流的算法.时间开销主要在语句(2),时间复杂度为 $O(N)$.

定义 **2.5**(小波系数通过率). 设 $X=[x_0,x_1,\ldots,x_{N-1}]$是由 $N=2^n$ 个数据点构成的流序列,经 Haar 小波变换并经给定阈值滤波后,非零系数个数与 N 之比称为小波系数通过率(overpass ratio).

定义 **2.6**(压缩比). 数据经给定压缩算法处理,压缩前与压缩后的数据量之比为μ,称μ为算法的压缩比.

定理 **2.4**. 设$\{X_0,X_1,\ldots,X_l\}$是 l+1 个长度为 2^n 的流序列,X_0 是满足$|Corr(X_i,X_0)|\geq\varepsilon$的特征序列,其中 $i\in\{1,2,\ldots,l\}$.X_0 经小波滤波后的系数通过率为α,算法 2 的压缩比μ满足: $\max\left\{\frac{l+1}{\alpha+4},\frac{2^n}{\alpha+4}\right\}\leq\mu\leq\frac{l+1}{\alpha}$ (证明略).

3 能量分解压缩算法

小波变换是将数据序列分解为不同尺度下系数的序列.小波系数代表了数据流的变化特征,同时蕴涵了不同尺度下的数据流能量分布.下面给出数据流序列在不同尺度下的能量度量及多尺度能量分解压缩算法.

定义 **3.1**(流序列的尺度能量). 设 $Y=[y_0,y_1,\ldots,y_{N-1}]$是 $X=[x_0,x_1,\ldots,x_{N-1}]$的小波系数构成的向量,则 X 在尺度 $\tau=2^k(0\leq k\leq n-1)$下的能量:$E(X|\tau=2^k)=\sum_{i=0}^{D-1}y_{i+B}^2$,其中,$D=2^{n-k}$,$B=\sum_{i=0}^{k}2^{n-i-1}$.

算法 **4**. 能量分解压缩算法:MSCStorage( ).

输入:特征流序列 X_0,流的集合$\{X_1,X_2,\ldots,X_m\}$.

输出:特征流的小波系数、流序列的 ID 号、平移参数 b_i 以及各尺度下的能量比.

(1) {Call for MultiStreamProcess( ); //对流序列进行预处理

(2) 对 X_0 进行小波变换,并存储大于滤波阈值的系数和平移参数 b_0;

(3) 计算特征流 X_0 在不同尺度下的能量 $E(X_0|\tau)$;

(4) for each $X_i \in \{X_1,X_2,\ldots,X_m\}$ do

(5) {计算不同尺度下的能量 $E(X_i|\tau)$,以及 $E(X_i|\tau)$与 $E(X_0|\tau)$的能量比 $R_i(\tau)$;

(6) 存储偶合流的 ID 号、平移参数 b_i 以及各尺度下的能量比}}

算法 4 给出将数据流的能量分解为不同尺度下的能量,除存储特征流序列的小波系数以外,还需存储偶合流与特征流在相同尺度下的能量比.与算法 2 相比,虽然压缩比有所降低,但更能准确地描述偶合流的能量分布特征.算法 4 的时间复杂度为 $O(N)$.下面将讨论用算法 4 压缩流序列的重构算法.

算法 **5**. 多尺度重构算法:MultScaleRebuild( ).

输入:特征流 X_0 的小波系数、偶合流的 ID、能量比、平移参数 b.

输出:流序列的重构信息.

(1) {for each $R_i(\tau)$ do 用 $R_i(\tau)$的平方根分别乘以相应尺度的小波系数,得到序列 $dwt[N]$;

(2) 对序列 $dwt[N]$重构生成 X_i;

(3) if $b_i \neq 0$ then $X_i=X_i-b_i$;

(4) if 存储的与特征流 X_0 偶合度$\leq -\varepsilon$ then $X_i=-X_i$;

(5) 输出流序列 X_i}

算法 5 给出了从经算法 4 压缩后的信息重构数据流的原貌的算法.如果偶合度为负,则进行对称化得到原有流序列.时间开销主要集中在小波变换的语句(2)上,则时间复杂度为 $O(N)$.

定理 **3.1**. 设$\{X_0,X_1,\ldots,X_l\}$是 l+1 个长度为 2^n 的流序列,X_0 是满足$|Corr(X_i,X_0)|\geq\varepsilon$的特征序列,且 $i\in\{1,2,\ldots,l\}$,X_0 经小波滤波后系数的通过率为α,算法 4 的压缩比μ满足: $\max\left\{\frac{l+1}{\alpha+n+3},\frac{2^n}{\alpha+n+3}\right\}\leq\mu\leq\frac{l+1}{\alpha}$ (证明略).

4 实验结果及分析

实验数据使用深圳证券交易所和上海证券交易所的股票交易价格数据.实验测试平台和主要参数如下:(1) CPU 为 PIII600 和内存为 256M 的计算机;(2) 操作系统为 Windows 2000;(3) 数据量大约为 50 万条记录;(4) 存储方式为微软 SQL SERVER2000.我们进行了如下几方面的实验.

4.1 比较不同算法的压缩比

在实验中选用了 200 对股票数据,窗口宽度为 512,偶合度阈值分别取 0.75,0.8,0.85,0.9.在给定不同偶合度阈值ε的情况下,研究了偶合度与算法的平均压缩比的关系,使用以下 3 种算法分别在等宽窗口上进行了实验:(1) 使用文献[1,2]中的传统小波算法,分别对每个数据流单独进行压缩处理,简记为 TCStorage 算法;(2) 使用多偶合数据流能量守恒的压缩算法(简称 MSNCStorage 算法)对多个偶合数据流进行压缩;(3) 使用多尺度的能量分解压缩算法(简称 MSCStorage 算法)对多个偶合数据流进行压缩处理.实验研究了不同压缩算法的压缩比,图 2 表明:在不同的偶合度阈值下,MSNCStorage 算法的压缩比最大;TCStorage 算法最低;MSCStorage 算法介于之间.其主要原因是,MSNCStorage 算法使用特征流的小波系数和偶合流的总能量表示多个流的特征;MSCStorage 算法使用特征流的小波系数和偶合流在不同尺度下的能量表示多个流;而 TCStorage 算法对每个数据流分别用其小波系数表示.实验表明:MSNCStorage 算法和 MSCStorage 算法的压缩比随着偶合阈值的升高而降低,主要原因是偶合度阈值升高,与特征流呈现强偶合的数据流的数量减少,则需要更多的特征流表示流信息;TCStorage 算法的压缩比与偶合度无关,主要是传统的小波压缩算法存储每个数据流的小波系数.

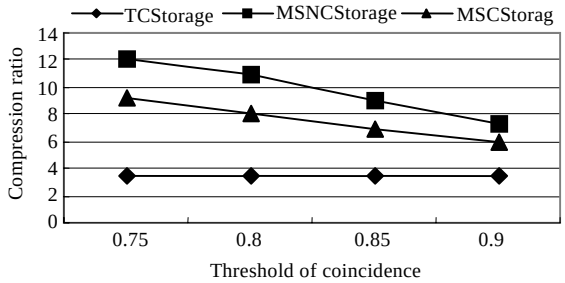

Fig.2 The compression ratio of algorithms

图 2 比较算法的压缩比

4.2 比较不同算法的相对误差

偶合度反映出两个数据流的变化规律的相似程度,两个数据流的偶合度越高,变化规律越相似.尽管小波变换是无损变换,但将系数小于给定的阈值进行归零化处理,则重构的序列与原序列会存在一定的误差;并且 MSNCStorage 算法和 MScaleStorage 算法为有损压缩.在此,我们研究了偶合度阈值ε与误差之间的关系.同样选用了 200 对流数据,窗口宽度为 512,偶合度阈值分别取 0.75,0.8,0.85,0.9,分别用 3 种算法压缩后重构生成新序列 g 与原数据流 f 的相对误差为$\|f-g\|/\|f\|$,且$\|f\|$表示 L^2 范数.研究了 3 种压缩算法的相对误差.如图 3 所示,MSNCStorage 算法的相对误差大于 MSCStorage 算法,且随着偶合阈值的升高而降低.其主要原因包括:(1) MSNCStorage 算法是用特征流的小波系数和偶合流的总能量来描述偶合流的特征,而 MSCStorage 算法不同于 MSNCStorage 算法,是用不同尺度的能量来描述数据流;(2) 偶合度阈值越高,特征流与偶合数据流的相似程度越高,则特征流更能准确表示与之偶合的数据流.因为传统的小波压缩算法是独立处理每个数据流,所以传统小波算法的相对误差与偶合度无关.从总体来看,传统的小波方法的相对误差较小,MSCStorage 算法的相对误差略大一些,而 MSNCStorage 算法的相对误差最大.

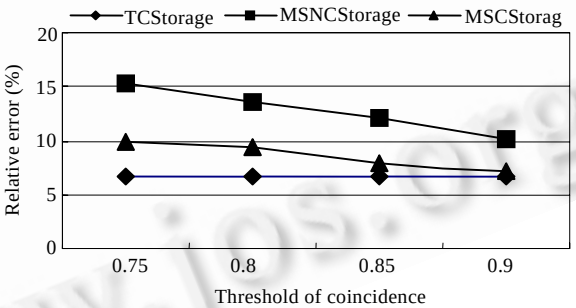

Fig.3 The relative error of algorithms

图 3 比较算法的相对误差

5 结束语

本文揭示了小波变换服从能量守恒规律,给出了小波能量多层次性的分解算法,提出了基于数据流偶合特征和小波能量分解的多数据流压缩的新方法.分析了不同的强偶合流序列的特点,设计了能量守恒的多数据流的压缩方法及多尺度下的能量分解压缩算法.实验验证了新算法的效率是传统的小波压缩方法的 2~4 倍.

References:

[1] 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. Toronto: Morgan Kaufmann Publishers, 2004. 288−299.

[2] Gilbert AC, Kotidis Y, Muthukrishnan S, Strauss MJ. One-Pass wavelet decompositions of data streams. IEEE Trans. on Knowledge

and Data Engineering, 2003,15(3):541−554.

[3] 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. New York: Springer-Verlag, 2003. 358−369.

[4] Chen AL, Tang CJ, Yuan CA, Peng J, Hu JJ. Mining correlations between multi-streams based on haar wavelet. In: Sui GL, Vianu V, eds. Advances in Computer Science: ASIAN 2005 the 10th Asian Computing Science Conf. Kunming: Springer-Verlag, 2005. 270−271.

[5] Papadimitriou S, Sun J, Faloutsos C. Streaming pattern discovery in multiple time-series. In: Böhm K, Jensen CS, Haas LM, Kersten ML, Larson PÅ, Ooi BC, eds. Proc. of the VLDB 2005 the 31st Int'l Conf. on Very Large Data Bases. ACM Press, 2005. 697−708. http://www.vldb2005.org/program/full_program.php.

[6] Sakurai Y, Papadimitriou S, Faloutsos C. BRAID: Stream mining through group lag correlations. In: Özcan F, ed. Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2005. 180−201. http://cimic.rutgers.edu/sigmodpods05

[7] Ma LS, Viglas SD, Li M, Li Q. Stream operators for querying data streams. In: Fan W, Wu Z, Yang J, eds. Proc. of the 6th Int'l Conf. on Web-Age Information Management. Berlin, Heidelberg: Springer-Verlag, 2005. 404−415.

[8] Burrus CS, Gopinath RA, Guo HT. Introduction to Wavelets and Wavelet Transform. Beijing: China Machine Press, 2005. 1−145.

陈安龙(1971－),男,四川仪陇人,博士,讲师,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.

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.

朱明放(1970－),男,博士生,副教授,主要研究领域为数据挖掘.

段磊(1980－),男,博士生,主要研究领域为数据挖掘.