

文章编号:1001-9081(2005)06-1369-04

一种基于小波大纲的数据流在线预测方法

郭吉平

(佳木斯大学 公共计算机教研部, 黑龙江 佳木斯 154007)

(guojiping6@163.com)

摘 要:描述了一种基于时间序列数据流大纲的预测框架,提出了构建具有有效降噪效果的小波大纲的方法,可根据背景噪声而分层自适应设置去噪(保留)阈值。并且在这种小波大纲的基础上实现了多尺度概要的分析和预测方法,能够分析动态变化的高频数据流的趋势、拐点、周期、方差的变化,用来为时间序列数据流提供实时的注解。在实际电力负荷数据上的仿真实验证明这种方法可以提供快速的精确的近似预测。

关键词:数据流;预测;小波;大纲;自适应阈值

中图分类号: TP311.12 **文献标识码:** A

Online prediction method for data streams based on wavelet synopses

GUO Ji-ping

(Commonality Teaching Department of Computer, Jiamusi University, Jiamusi Helongjiang 154007, China;)

Abstract: Several studies in recent years were demonstrated that wavelets can be efficiently used to compress large quantities of data down to compact wavelet synopses and provided fast and fairly accurate approximate answers to queries. In this paper author presented a wavelet synopses and prediction framework to analyze dynamic high-frequency data streams. A novel construction method for wavelet synopses provided with efficient De-noise ability was proposed. Its varied threshold schema for every decomposition level could adapt itself to the change of background noise. Based on this wavelet synopsis, a multi-scale prediction and analysis method for summarization was used to separate out the trend, turning points, cyclical fluctuations and autocorrelational effects etc. This framework was used to provide annotation for time series data streams at real-time. Experimental results with real power load datasets demonstrate that our approach achieves improved velocity and accuracy to approximate prediction queries when compared to existing techniques.

Key words: data streams; forecasting; wavelet; synopses; adaptive thresholding

0 引言

数据流应用广泛引用于网络、工控、气象以及交通等各个领域。数据流就是在这些应用中连续采样所得到的数据序列。这些新应用的特点是:查询时将数据作为序列而不是集合;借助新数据的插入来更新查询模式;由于流的无限性,无法物化整个数据流,需要一遍扫描算法(one-pass algorithm);在很多实际统计类应用中,例如决策支持系统、查询优化等,查询的解答大部分是定性的,用户并不需要获得确切值,牺牲部分准确性来换取速度是可取的^[1]。因此,设计一遍扫描算法,实时地给出查询的近似结果就成为数据流模型下数据处理的目标。此类算法的关键在于设计一个远小于数据集规模的结构,从而可以在内存中完成对数据的处理。相对于数据流的规模而言,这种名为大纲数据结构(synopsis data structure)的规模至多应该是次线性的。即如果流的长度为 N ,则大纲数据结构大小不超过 $O(\text{polylog}(N))$ ^[2],并且处理流上每一组数据的时间不超过 $O(\text{polylog}(N))$ 。

小波分析是一种时域-频域分析法,它在时域和频域上同时具有良好的局部化性质,并且能根据信号频率高低自动调节采样的疏密,能对不同的频率成分采用逐渐精细的采样步长,从而可以聚集到信号的任意细节,尤其是对奇异信号很

敏感,能很好的处理微弱或突变的信号。其目标是将一个信号的信息转化成小波系数,从而能够方便地加以处理、储存、传递、分析或被用于重建原始信号。这些优点决定了小波分析可以有效地应用于预测问题的研究。

为了实现快速数据流统计信息的预测,我们提出了一种可变阈值的小波大纲算法,基于无偏似然估计原理为不同层次选择具有自适应性的阈值,能够适应数据流的变化;在使用可变小波系数阈值产生的大纲上采用不同的分析方法,通过对不同层次的预测信息的合成,得到最终预测信息。

1 相关工作

多尺度预测技术^[3]的基本思想是应用小波变换将信号的不同频带成分分解到不同的级别上,对不同频率性质的分解系数采用不同的预测方法,通过对不同层次的预测值的合成得到最终预测值。文献[4]提出了非筛选式小波系数的(冗余的)小波变换方法,实现对易受到噪声干扰的数据流值的预测,保证了预测精度。文献[5]为了对变化的时间序列形成统计注解,应用离散小波变换和多尺度分析方法生成了一种能够表示趋势、拐点、方差变化等统计信息的概要。文献[6]提出 MUSCLES 模型对高相关性的数据流作基于线性回

收稿日期:2005-01-06;修订日期:2005-03-14

基金项目:江苏省高技术项目(BG2004034);江苏省2004年度研究生创新计划项目(xm04-36)

作者简介:郭吉平(1962-),女,副教授,黑龙江佳木斯人,主要研究方向:操作系统、数据处理。

归的预测,算法需要在很长的序列上才能完成,需要更低计算复杂性的增量分析技术。已有的应用小波分解实现预测的方法都是保持全部小波变换产生的系数,这样固然能够得到较高的预测精度,但对于高速数据流产生的海量数据,在线算法的有限时空资源无法操纵全部的小波系数,因此保留了部分小波系数的大纲结构是提高数据流在线近似预测速度的关键。

关于小波系数保留问题,常用的有硬阈值和软阈值技术^[7]。硬阈值是将所有绝对值小于一个阈值 T_j 的小波系数都设置为零;软阈值是使用值 $\tilde{\omega}$ 替换每个小波系数:

$$\tilde{\omega}_{j,k} = \begin{cases} \text{sgn}(\omega_{j,k}) (|\omega_{j,k}| - T_j), & \text{if } |\omega_{j,k}| \geq T_j \\ 0, & \text{otherwise} \end{cases}$$

软阈值的选择主要有以下几种:基于 stein 无偏似然估计原理的自适应阈值选择^[8],最优预测变量阈值选择,采用极大极小原理阈值选择等。文献[9]提出概率小波阈值,根据重构时的重要程度为每个小波系数分配一个保留概率,可以为近似查询解答提供很高的误差保证。然而这种基于概率的技术可能会因为不好的翻硬币序列而导致质量差的方案,从而降低查询的精度。

2 问题定义

2.1 数据流和查询模式

数据流 S 是只能一次读取的数据项的有序序列 \dots, x_i, \dots , 数据项 x_i 以逐项连续的方式输入,用 $S(i)$ 表示数据流 S 在时刻 i 的观测值 x_i ,在最简单的情况下,假设 x_i 的值域离散且有序,数据流 S 可视为将 x_i 值映射到非负的整数序列上的一维函数。例如,电力负荷数据流 S 中,如果 x_i 表示不同负荷值, x_i 的值域为可能的负荷值集合,那么整数序列可能表示特定平均负荷的时段。

定义 x_{i-n}, \dots, x_i 为滑动窗口数据流模式,当一个新的数据项 x_i 到达时,查询只对最近 n 项滑动窗口中的项目 $x_i(t-n \leq i < t, t$ 表示当前时刻)感兴趣,窗口之外的项视为过期。

定义数据流上的统计类预测查询如下:

```
SELECT STA(S.attr)
FROM S
SAMPLERATE  $\Delta t$ 
WINDOW |W|, NOW +  $k * \Delta t$ 
```

式中 STA 表示统计函数,可能是相关性、方差、拐点、趋势、异常模式(附带误差范围)等; $attr$ 表示属性值; Δt 表示预测采样间隔; $|W|$ 表示滑动窗口长度; $NOW + k * \Delta t$ 表示数据流 S 的未来 $k * \Delta t$ ($k \in Z$) 时刻, k 表示采样间隔的倍数。本文仅涉及滑动窗口模式下单数据流的近似预测问题,由数据流处理引擎生成并维护流大纲,提供满足用户的精度要求的近似预测查询解答。

2.2 小波大纲

离散小波变换(DWT)将信号表示成能够同时在时域和频域定位的平稳量(As)和细节量(Ds),DWT是一个离散卷积的过程,可用如下公式表示:

$$w * x_i = \sum_{j=-\infty}^{\infty} w_j x_{i-j} \quad (1)$$

其中 x_i 是原始数据, w 是相对于小波基的低通或高通滤波,实际应用中通过金字塔算法^[10]实现 DWT。小波种类很多,最常见且最简单的是哈尔小波(Haar wavelet)。DWT能够适应数据流的非静态性和时变性,借助原始信号的小波分解

得到良好的时频分辨率,可以发现数据中有意义的全局或局部模式。在每个分解层关键的频率成分,可以分析最近数据流的行为,同时也可以预测未来的行为。

2.3 stein 无偏似然估计阈值

为了利用小波变换产生的大纲实现预测,而不是使用原始观测的流值向量直接预测,首先需要知道在每一尺度使用多少和哪些小波系数。

Donoho^[7]提出了源于统计学中 Stein 的无偏似然估计原理的阈值选取算法,方法是对一个给定的阈值 t ,得到它的似然估计,再将非似然 t 最小化,就得到了所选的阈值。这是一种软件阈值估计器,其优点是能够自适应背景噪声的强度变化,保留合适的小波系数。

3 基于软阈值小波大纲的预测方法

我们提出的基于数据流大纲的预测框架的基本思想是,根据小波分析和 Stein 的无偏似然估计软阈值理论,使用 DWT 分解和软阈值将数据流 S 分解成由不同频率的成分组成的大纲。然后再大纲上采用不同的分析方法通过对不同层次的预测信息的合成得到最终预测信息。

A_n 为 N 层平稳成分,是时间序列数据流 S 的主要成分,捕捉 S 的整体变化趋势。为预测平稳成分 A'_n ,适合使用自回归(AR)模式预测影响因素和 A_n 的关系。 D_1, D_2, \dots, D_n 是随机序列,表示细节成分,为获得 D_1, D_2, \dots, D_n 之间的关系和影响因素,应用 Fourier 能量谱分析理论,抽取季节项(cycle)。

将时间序列分解为细节成分和平稳成分的用途是对序列的动态注释。例如对股价走势图的注解,或在电力系统潮流分析中对电力负荷曲线的注解。小波分析可以识别方差变化,同时也可识别趋势和季节成分。一旦识别出来,就可以为整个时间序列或离散的序列点添加预先计算并存储的短语集,能够自动地处理有关时变、趋势和季节变化等的注解。

3.1 算法描述

算法 1 小波大纲的分层自适应阈值生成算法

input: DWT 的层数 k , 第 k 层的 N 个小波系数

output: 第 k 层阈值 T

对于第 k 层的小波系数

1) 将第 k 层 N 个小波系数的平方按升序排列,得到向量: $W = [w_1, w_2, \dots, w_N]$, 其中 $w_1 \leq w_2 \leq \dots \leq w_N$;

2) 由此计算似然估计向量 $R = [r_1, r_2, \dots, r_N]$,

$r_{\min} = +\infty$, i 值从 1 变到 N , 执行下列操作:

$$r_i = (n - 2i + (n - i)w_i + \sum_{j=1}^i w_j) / N;$$

if $r_i < r_{\min}$ then $r_{\min} = r_i$; $\min l = i$;

3) 以 R 向量中最小值 r_{\min} 作为似然估计值,由 r_{\min} 的下标变量 $\min l$ 求出对应的 $w_{\min l}$;

4) $T = \sqrt{w_{\min l}}$ 的噪声强度(噪声强度 $\delta = \frac{1}{0.674} \sum_{i=1}^{N-1} |D_i^k|$)。

定义原始序列的主要特征,包括拐点、趋势、周期、方差的变化等为概要。本文使用 DWT 抽取原始序列的概要,并在其基础上生成相应的图形和注释,描述时间序列数据流的变化。根据不同尺度上的主要特征成分,可以合并出总体预测信息。

算法 2 基于小波大纲的统计信息(注解)提取算法

input: 滑窗长度 $|W|$, 采样间隔 Δt , 采样的个数 M

output: 平稳成分的趋势,季节项(周期项),拐点,方差

1) 计算离散小波变换需要的层数 $L = \lceil \log_2 M \rceil$;

2) 执行 L 层的 DWT 变换,调用算法 1,判别小波系数与阈值关系,获得小波大纲 D_i , $i = 1, 2, \dots, L$ 和 A_L ;

3) 提取统计信息:

- a) 对 A_L 执行线性回归计算趋势;
- b) 抽取季节项(cycle) 在每个 D_i 上执行 Fourier 能量谱分析,选取最大能量的 D_i 为 D_s ;

$$D_i(k) = \frac{1}{N} \sum_{t=0}^{N-1} D_i(t) e^{-j2\pi kt} \quad k = 0, 1, \dots, N-1;$$

- c) 抽取每个 D_i 的极值点作为拐点(turning points);
- d) 通过用 C_k 的 NCCS 索引查找每个方差(variance);

$$\tilde{p}_k = \frac{\sum_{i=L_j-1}^k \tilde{w}_{j,i}^2}{\sum_{i=L_j-1}^{N-1} \tilde{w}_{j,i}^2}, k = L_j - 1, \dots, N-2, \tilde{w}_{j,i} \text{ 为第 } j \text{ 层}$$

DWT 的小波系数。

算法3 平稳成分 A_L 自回归(AR) 预测算法

input: n 个平稳成分 A_L
output: 下一个平稳成分 A'

- 1) 相对于当前时间,第 k 步预测模式如下:

$$A_{i,t+k} = \xi_{i,1} a_{i,t-1+k} + \xi_{i,1} a_{i,t-2+k} + \dots + \xi_{i,1} a_{i,t-l+k} + \mu_i, \\ i = 1, 2, \dots, n(n \in Z)$$

- 2) 最后平稳成分的估计值如下:

$$A' = A'_1 + A'_2 + \dots + A'_i + \dots + A'_n, i = 1, 2, \dots, n$$

在获得了序列的主要成分之后,可以分别抽取趋势项和周期项成分,进而组合生成整体趋势。周期项基于不同的振幅和周期进行对称地扩展;趋势项在第 L 层小波大纲中可以以多项式函数线性扩展。这种基于小波大纲的统计信息预测算法不是预测未来序列的精确值,如果保存了全部的小波系数(A_L 和每一个 D_i) 可以得到更加精确的预测。许多复杂的序列特征需要在建模之前进一步研究或者应用带有合适条件的扩展。

算法4 预测算法

input: 预测的步长 K
output: K 步预测的聚集和 $prediction_K$

- 1) 调用算法2生成概要;
- 2) 对于第 K 步预测,将季节(周期)成分 D_s 对称地扩展到右侧的第 K 个点处形成 $D_s, prediction$;
- 3) 呈线性地将趋势成分扩展到右侧的第 K 个点处得到 $A_K, prediction$;
- 4) 用2)、3)步的结果获得 K 步预测的聚集和 $prediction_K = D_s, prediction + A_K, prediction$

3.2 复杂度分析

算法1中: N 表示第 k 层的小波小波系数,第1)步假设应用快速排序法,时间复杂度 $O(kM \log_2 N)$,递归用到的栈最大深度为 $O(\log_2 N)$,保存 N 个小波系数 $O(N)$,空间复杂度为 $O(N)$;第2)步求似然估计向量时间 $O(N(N-1)/2)$,保存 N 个似然估计空间 $O(N)$;第4)步噪声强度函数时间 $O(N)$,空间 $O(1)$;故算法1运行期间最大时间复杂度为 $O(N^2)$,空间复杂度为 $O(N)$,而在内存中需驻留的仅为 $O(1)$ 空间。

算法2中:第2)步假设滑窗的中采样点为 M 个,应用 Haar 小波,DWT 变换需要 $O(\log_2 M)$ 时间和空间,因为第 k 层的小波系数的个数 $N = 2^{\log_2 M - k}$,DWT 共有 $\log_2 M$ 层,得到的小波大纲的时间为 $O(\log_2 M (2^{\log_2 M - k})^2)$,空间复杂度 $O(B)$,其中 $B \ll \log_2 M$ 。第a)步时间 $O(n)$,第b)步时间 $O(N)$,第c)步 $O(N-1-(L_j-1))$,故算法1运行期间最大间需求为 $O(M^2 \log_2 M)$,空间复杂度为 $O(\log_2 M)$,而需要在内存中驻留的仅需 $O(B)$ 空间。

4 实验分析

为检验算法的性能,我们参照文献[5]的实验方法,采用

南京某地区电力系统真实负荷数据集进行了测试。下面本文提出的方法用 ADWT 表示。实验计算机的配置是 2.66GHz Pentium/256M/80G,在 windows 2000 server 的环境下应用 VC 实现了主要的算法,部分算法引用了 matlab 工具箱函数。真实负荷数据集的采样间隔为 5min,日负荷数据为 288 点,点数密集的好处在于充分提供了负荷特征信息。实验使用了 2002/3/15 的 24 小时的负荷数据,如图 1 所示。

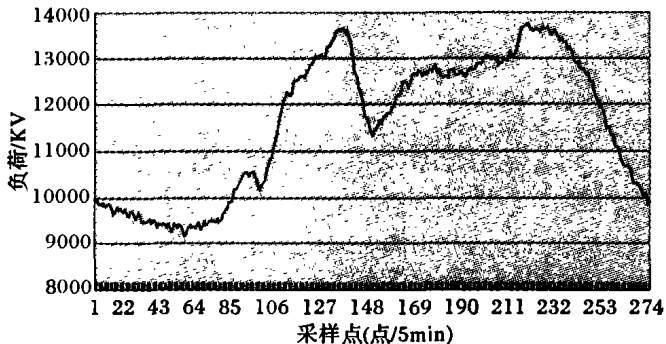


图1 2002/3/15 实际电力负荷曲线

应用算法2产生的序列特征信息(概要)如表1,趋势信息表示负荷的长期移动趋势是上升的,然而两个重要的拐点在 $t = 122$ 和 $t = 181$ 之间,倾斜程度达 86%,表示了下降的趋势,周期性的峰值成分按 30~50 点的规律出现,表示这种周期性的行为会在未来连续出现。

表1 根据图2得到的负荷序列概要注释

特征	阶段	细节
趋势	第1阶段	$x_1 = 98.0672t + 9776.3, t < 122$
	第2阶段	$x_2 = 59.8124t + 12760, 123 < t < 180$
	第3阶段	$x_3 = 41.1373t + 10683.1, 181 < t < 288$
拐点	下降	6, 86, 130, 233
	上升	48, 77, 159, 213
方差变化	位置	131
季节项	周期	34
	峰值位置	44, 80, 120, 195, 233

完成了序列的统计信息(概要)分析之后,使用 ADWT 将这些概要成分分解到不同的尺度上,就可以重新组合形成下一天的聚集预测。我们使用前一日(2002/3/15)的“主要成分”,趋势和显著周期信息(表1),产生第二日(2002/3/16)的时间序列的相应元素,预测结果如图2所示。预测(底部细曲线)与实际时间序列(中间粗曲线)是一致的。都显示了下降的态势。在预测曲线中周期性波动看上去匹配的不好,在观测点 160 之后明显地偏差,然而 222 观测点之后预测曲线也有所改良。预测序列与实际时间序列的相关度为 54.4%,均方误差为 0.000042。

为了评估本文方法的结果,我们使用 BP 多隐层神经网络(MLPNN)执行非线性 AR 预测,MLPNN 具有如下设置,11 个输入层,12 个隐层,1 个输出层,对负荷数据流规格化的方法是: $x'_i = (x_i - \mu_x) / \sigma_x$,其中 μ_x 为算术均值, σ_x 为标准方差。定义负荷预测准确率 $\Delta L_{FA} = [1 - \text{SQRT}(1/N \times \sum \Delta L_R^2)] \times 100\%$,其中 $\Delta L_R = (L_F - L_P) / L_P \times 100\%$ 为负荷预测相对误差, L_F 为预测负荷, L_P 为实际负荷。

如图2所示(上面细线表示 MLP 预测,中间粗曲线表示实际负荷,下面细线表示 ADWT 预测曲线),MLPNN 预测曲线(上部细曲线)与原始曲线具有很大差异,均方根误差是 ADWT 方法的 12 倍,这是由于 BPNN 网具有短时相关性,即

假设下一个时间序列值只与前一个值相关,然而许多电力负荷数据流的变量之间的相关性不会以较快的速率衰减,观测值经历较长时间之后仍然具有显著的相关性。ADWT 产生的预测与实际值的相关性 +42%,比 MLP 的负相关性要好。

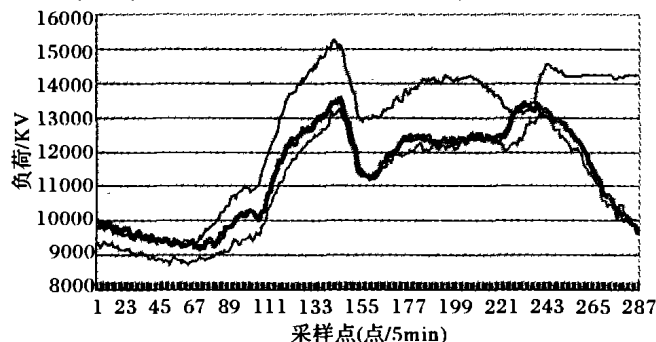


图2 预测性能比较

5 结语

本文提出基于无偏似然估计软阈值的小波大纲构建方法,在其基础上的预测方法对分解后的平稳成分 A_L 使用线性回归预测,对各个级别系数 D_i 采用 Fourier 能量谱分析,根据数据的不同成分得到统计性信息的预测。与文献[5]相比,本文提出的基于小波大纲的数据流预测方法考虑了信号中必然存在的噪声的影响,形成的小波大纲可以有效地消除噪声,同时大幅度减少了内存中保存的小波系数的空间,实验证明利用这种小波大纲得到的预测信息在计算速度上有大幅提高,而质量并未明显降低。与神经网络在预测中的应用相比,由于神经网络的隐节点数目难以确定、过度拟合、训练时间长、预测精度对训练样本的质量和数量敏感等原因,不适合快速在线预测。文中所提算法可伸缩性能好,可以快速适应趋势的变化,能够广泛地应用于数据挖掘领域。下一步工作包括实现增量更新的小波大纲预测算法,研究多数据流的相关性预测等。

(上接第 1361 页)

$\{B, D\}_2, \{B, F\}_2, \{C, F\}_2, \{D, E\}_2, \{D, F\}_2, \{B, G\}_2$ 和 $\{A, B\}_0, \{A, F\}_1, \{A, G\}_0, \{B, C\}_1, \{C, D\}_1, \{C, G\}_0, \{D, G\}_0, \{E, F\}_0, \{E, G\}_1, \{F, G\}_1$, 接着,产生新的候选项目集 \bar{C}_3 , 并通过再次扫描数据库计算它们的支持数。有 $C_3 = \{A, D, E\}_1, \{B, D, E\}_1, \{B, D, F\}_1, \bar{L}_3 = \{A, C, E\}_2$ 和 $\bar{L}_3 = \phi$, L_3 和 \bar{C}_3 分别更新为 $\{A, C, E\}_2$ 和 $\{A, D, E\}_1, \{B, D, E\}_1, \{B, D, F\}_1$, 因为 $|L_3| \leq 1$, 挖掘过程终止。

故 $M_2 = \{A, C\}_3, \{A, D\}_2, \{A, E\}_3, \{B, E\}_3, \{C, E\}_3, \{B, D\}_2, \{B, F\}_2, \{C, F\}_2, \{D, E\}_2, \{D, F\}_2, \{B, G\}_2, \{A, C, E\}_2$, 再利用大项目集生成关联规则算法可求得关联规则。

4 结语

综上所述,利用已存储信息,PSI 算法每次扫描能减少候选项目集支持数的计算。对于表 1 事务数据库 D, 表 2 表示当前最小支持度变为 2 时,PSI 和 Apriori 算法的结果比较。PSI 利用最小支持数为 3 的已存信息,只需两次扫描数据库,而 priori 需要三次扫描数据库。而且,每次扫描数据库计算支持数的候选项目集总数, priori 比 PSI 的大。最后,需要指出 PSI 算法的思想,不仅可运用到挖掘关联规则的 priori 改进算法中,而且可运用到挖掘序列模式的 prioriall 算法中,以减

参考文献:

- [1] GABER MM, KRISHNASWAMY S, ZASLAVSKY A. Ubiquitous Data Stream Mining, Current Research and Future Directions [A]. Workshop Proceedings held in conjunction with The Eighth Pacific-Asia Conference on Knowledge Discovery and Data Mining [C], Sydney, Australia May 26, 2004.
- [2] GILBERT AC, KOTIDIS Y, MUTHUKRISHNAN S, et al. Surfing wavelets on streams: One-Pass summaries for approximate aggregate queries [A]. In Proc. of the 27th Int'l Conf on VLDB [C] 2001. 79-88.
- [3] ZHENG H, ZHANG L. The factor analysis of short-term load forecast based on wavelet transform [J]. Power System Technology, Proceedings. PowerCon 2002. International Conference on, 2002, 2: 13-17.
- [4] RENAUD O, STARCK JL, MURTAGH F. Prediction Based on a Multiscale Decomposition, International Journal of Wavelets [J]. Multiresolution and Information Processing, 2003, 1(2): 217-232.
- [5] AHMAD S, TASKAYA-TEMIZEL T, AHMAD K. Summarizing Time Series: Learning Patterns in 'Volatile' Series [A]. Proceedings of the 5th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2004), Lecture Notes in Computer Science [C]. Heidelberg: © Springer-Verlag, 2004, 3177: 523-532.
- [6] BYOUNG-KEE Y, SIDIROPOULOS N, JOHNSON T, et al. Online data mining for coevolving time sequences [A]. In ICDE [C], 2000. 13-22.
- [7] DONOHO D, JOHNSTONE I. Adapting to unknown smoothness via wavelet shrinkage [J]. Journal of the American Statistical Association 1995, 90(9): 1200-1224.
- [8] STEIN CM. Estimation of the mean of a multivariate normal distribution [J]. Annals of Statistics. 1981, 9: 1135-1151.
- [9] GAROFALAKIS M, GIBBONS PB. Probabilistic Wavelet Synopses [J]. ACM Transactions on Database Systems, 2004, 29(1): 43-90.
- [10] MALLAT S. A theory for multiresolution signal decomposition, the wavelet representation [J]. IEEE Trans. Pattern Ana. and Machine Intell, 1989, 2(7).

少重复挖掘知识过程中的时间和空间上的开销。

表2 Apriori 与 PSI 的比较

	Apriori 数	PSI 数
C_1	7	
C_2	21	\bar{C}_2 6
C_3	4	\bar{C}_3 3
总数	32	9

参考文献:

- [1] AGRAWAL R, IMIELINSKI T, SWAMI A. Mining association rules between sets of items in large databases [A]. In: Proceedings of the ACM SIGMOD Conference on Management of Data [C]. Washington D. C, 1993. 207-216.
- [2] AGRAWAL R, SRIKANT R. Fast algorithm for mining association rules [A]. In: Proceedings of the 20th International Conference on Very Large Databases [C]. Santiago, Chile, 1994. 478-499.
- [3] YUGAMI N, OHTO Y, OKAMOTO S. Fast discovery of interesting rules [A]. In: Proceeding of the 4th Pacific-Asia Conference [C]. PAKDD2000, Japan, 2000. 17-27.
- [4] CHEN M-S, HAN J, YU PS. Data mining: An overview from a database perspective [J]. IEEE Transactions on Knowledge and Data Engineering, 1996, 8(6): 866-883.
- [5] TSAI PSM, CHEN CM. Discovering knowledge from large database using prestored information Information System [J], 2001, 26(1): 1-14.