ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.6, June 2008, pp.1413−1421
DOI: 10.3724/SP.J.1001.2008.01413

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

基于能量和频繁模式的数据流预测查询算法*

陈安龙[1+], 唐常杰[2], 傅 彦[1], 廖 勇[3]

[1](电子科技大学 计算机科学与工程学院,四川 成都 610054)

[2](四川大学 计算机学院,四川 成都 610065)

[3](四川大学 数学学院,四川 成都 610065)

An Algorithm for Predictive Queries over Data Stream Based on Energy and Frequent Pattern

CHEN An-Long[1+], TANG Chang-Jie[2], FU Yan[1], LIAO Yong[3]

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

[2](College of Computer Science and Engineering, Sichuan University, Chengdu 610065, China)

[3](College of Mathematics, Sichuan University, Chengdu 610065, China)

+ Corresponding author: E-mail: chenanlong@uestc.edu.cn

Chen AL, Tang CJ, Fu Y, Liao Y. An algorithm for predictive queries over data stream based on energy and frequent pattern. *Journal of Software*, 2008,19(6):1413−1421. http://www.jos.org.cn/1000-9825/19/1413.htm

Abstract: A new predict model was contrived, which involves local stream energy prediction, the energy distribution pattern mining, the predictive series reconstruction and measurement method of stream energy. A new method was designed to forecast stream by energy regression and wavelets decomposing based on frequent pattern, and extended to multi-streams with strong coincidence. The concept of the nearest maximum frequent pattern was proposed to decompose local stream energy. The validity of new algorithm was demonstrated by extensive experiments on real data.

Key words: data stream; stream energy; predictive query; wavelet analyze; frequent pattern

摘 要: 设计了数据流预测查询的新模型,包括局域流能量预测、能量分布模式挖掘及预测序列的重构和数据流能量的度量方法;设计了融合数据流能量回归与基于频繁模式的小波分解预测新方法,并将新算法推广到强偶合多数据流的预测查询;提出了最近最频繁序列模式的新概念,并应用于局域流能量分解;在真实数据上的模拟实验,验证了算法的有效性.

关键词: 数据流;流能量;预测查询;小波分解;频繁模式

中图法分类号: TP311 **文献标识码**: A

网络通信与传感器技术的发展催化了数据流在诸多领域的大量产生和广泛应用,数据流具有无限连续产生、快速变化、在时间语义约束下服从全序关系等特征,如在股票交易、地震检测、军事目标监测、网络监控及传感器网络等领域产生的数据.由于传统数据库的查询技术难以满足高速变化的数据流查询需要,因此,数据

* Supported by the National Natural Science Foundation of China under Grant Nos.60473071, 10476006 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2006AA01Z414 (国家高技术研究发展计划(863))

Received 2007-04-25; Accepted 2007-08-03

流查询的研究是数据挖掘领域的重要研究内容之一.现有的研究工作主要集中在数据流的局域实时信息查询和压缩处理,但对数据流变化趋势的预测查询研究较少.文献[1]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[2]研究了在滑动窗口中的数据流的统计查询问题.文献[3]使用了直方图的方法,研究了数据流的近似聚类查询问题.文献[4]研究了数据流的实时在线查询,如何预知数据流的未来信息,达到把握现在、预知未来具有重要意义.但文献[1–4]都没有研究如何查询数据流未来趋势.文献[5]运用多元回归方法,研究了预测查询数据流聚集的方法,做了具有创新性的工作,但该方法有先天性的不足,仅能预测数据流在一定时间区间内聚集查询,如 AVG 平均与 SUM 求和,但不能预测在局域区间上的数据序列,也不能利用多数流的相互关系,对多数据流的变化规律进行预测查询.我们在文献[6,7]已经研究了运用小波技术挖掘数据流之间的同步偶合和异步偶合关系,并在文献[8]中研究了融合小波技术和偶合特征对多数据流进行压缩处理,但没有研究多数据流的预测查询问题.为了弥补已有研究的不足,本文继承了对数据流偶合关系的研究成果,分别研究了单数据流和多偶合数据流预测查询方法,提出了新的融合小波和能量回归的预测查询方法.新方法不仅适用于单数据流的预测查询平均值与 SUM 值,而且可以预测查询数据流的序列分量,并推广到多偶合数据流的预测查询.

1 本文的主要贡献

本文在继承现有研究成果的基础上,研究了融合数据流局域能量回归和频繁模式分解的预测查询新方法,主要工作如下:

(1) 提出融合能量回归和频繁模式分解的数据流趋势预测新模型.新预测模型主要包括 3 部分:预测局域流能量、确定流能量的分布模式及预测序列的重构;并给出了数据流序列在局域窗口上流能量的度量方法.

(2) 提出了最近极大频繁模式的概念,并用最近极大频繁模式表示预测区间的能量分布规律.相对于远期变化规律,近期变化规律对数据流的未来变化影响更大,因此,采用最近时期出现频率最高的小波系数模式来描述将来变化规律.

(3) 设计基于能量回归与频繁模式的数据流预测查询算法.新方法相对于传统线性回归预测方法、预测频繁变化的流序列具有显著优势,通过计算局域窗口的流能量,消除了数据单元变化的振荡性,适合于某些非规律性变化,经过分段求平方和后,而呈现规律性变化的数据序列.

(4) 采用最近极大频繁模式的小波系数对流能量进行分解.在局域区间上,数据流能量反映了在时间区间上的累积效应,最近极大频繁小波系数模式反映了预测能量的分布规律;使用最近极大频繁小波系数模式对流能量进行分解,可得到预测窗口上的流序列值.

(5) 将数据流预测查询新算法推广到多数据流的预测.在多数据流环境中,存在强偶合关系的数据流,使用多元回归预测流的局域能量,使用强偶合流的最近极大频繁小波系数模式对预测流的局域能量进行分解,即可得到预测流在预测窗口上的流序列.

(6) 设计了预测模型的自适应调整策略.当预测失败的次数大于预先给定的阈值时,系统自动重构预测模型,以降低预测误差.理论分析与实验结果表明:新算法具有较高的预测精度.

2 数据流的预测模型

本文将继承文献[6–8]对数据流的研究成果,在消除数据流的噪声基础上,融合能量回归和频繁模式的数据流趋势的预测查询.其预测模型分为 3 步:(1) 预测局域流能量;(2) 挖掘流能量的分布模式;(3) 重构预测序列.将采用小波算法重构预测序列,使用频繁序列模式作为能量分布模式,使用多元回归方法预测数据流能量,下面我们详细介绍频繁序列模式的挖掘算法以及数据流能量预测的理论基础.

2.1 频繁序列模式的挖掘算法

本文用滑动窗口将数据流划分为若干数据单元数相同的子序列,每个子序列都呈现某种变化模式,假定某

种模式在数据流中最近时间内频繁出现,简称频繁序列模式,则认为该模式最有可能出现在未来窗口.在实际应用中,很难找到模式特征完全一致的两个子序列,当两个序列模式的相似度达到给定值时,可近似认为两个序列的模式相同,流序列的相似度使用文献[6,7]的偶合度表示.为了便于理解,下面将给出一系列定义和算法描述.

定义 1(ε-相似序列). 设 X_1 和 X_2 为等长流序列,对于给定的域值$\varepsilon(0\le\varepsilon\le1)$,如果偶合度[6,7]$Corr(X_1,X_2)\ge\varepsilon$,则称序列 X_1 和 X_2 互为ε-相似.可近似认为,两个ε-相似序列的变化模式相似.

定义 2(ε-相似频度). 设数据流 S 中所有等长子序列构成集合 Sub 且 $X_i\in Sub$,对于给定的域值$\varepsilon(0\le\varepsilon\le1)$,$\Omega=\{X_j|\forall X_j\in Sub$ 且 $Corr(X_i,X_j)\ge\varepsilon\}$,则称 $f(X_i|\varepsilon)=|\Omega|/|Sub|$为 X_i 在流 S 中的ε-相似频度.

定义 3(频繁流序列). 设数据流 S 中所有等长子序列构成集合 Sub 且 $X_i\in Sub$,对于给定的域值α和$\varepsilon(0\le\alpha\le1,0\le\varepsilon\le1)$,如果 $f(X_i|\varepsilon)\ge\alpha$,则称 X_i 在流 S 中为频繁流序列;如果 $f(X_i|\varepsilon)=\mathrm{Max}\{f(X_i|\varepsilon)|f(X_i|\varepsilon)\ge\alpha$且 $X_i\in Sub\}$,则称 X_i 在流 S 中为最大频繁流序列.

由于数据流具有无限性,在有限的内存里难以计算出在全局意义上的频繁序列;随着时间的推移,最近一段时间内的变化模式在将来更有可能出现,因而挖掘最近时期内的频繁序列更为合理.

定义 4(最近ε-相似频度). 设数据流 S 中最近 k 个等长子序列构成集合 Sub_k 且 $X_i\in Sub_k$,对于给定的域值$\varepsilon(0\le\varepsilon\le1)$,$\Omega_k=\{X_j|\forall X_j\in Sub_k$ 且 $Corr(X_i,X_j)\ge\varepsilon\}$,则称 $f_k(X_i|\varepsilon)=|\Omega_k|/|Sub_k|$为序列 X_i 的最近ε-相似频度.

定义 5(最近频繁流序列). 设数据流 S 最近 k 个等长子序列构成集合 Sub_k 且 $X_i\in Sub_k$,对于给定的域值α和$\varepsilon(0\le\alpha\le1,0\le\varepsilon\le1)$,如果 $f_k(X_i|\varepsilon)\ge\alpha$,则称 X_i 为最近频繁流序列;如果 $f_k(X_i|\varepsilon)=\mathrm{Max}\{f_k(X_i|\varepsilon)|f_k(X_i|\varepsilon)\ge\alpha$且 $X_i\in Sub_k\}$,则称 X_i 为最近最频繁流序列.

由于数据流具有无限性,本文采用嵌套滑动窗口方法挖掘最近频繁流序列和最近最频繁流序列.假设主滑动窗口内含有连续数据单元个数为 $k2^p$,即窗口的宽度为 $k2^p$;并将滑动窗口划分为 k 宽度为 2^p 的局域窗口,即子序列的长度为 2^p;主滑动窗口每次滑动步长为 2^p.称这样的滑动窗口为嵌套滑动窗口,如图 1 所示.

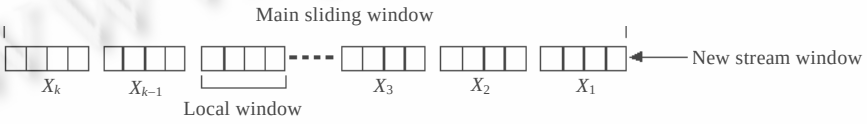


Fig.1 The nesting sliding windows

图 1 嵌套滑动窗口

最近频繁流序列最有可能呈现在未来局域窗口中,因此,选用最近 k 个局域窗口内的最近最频繁流序列模式作为未来局域窗口的特征模式.当最近 k 个局域窗口内不存在最近最频繁模式时,计算在最近 $k=\lceil k/2\rceil$个局域窗口内的极大频繁模式;若仍然不存在极大频繁模式,用同样的方法,计算在最近 $k=\lceil k/2\rceil$个局域窗口内的最近最频繁序列;照此进行下去,...,如果进行到 k 小于给定的域值ζ,则用当前局域窗口的序列模式作为预测窗口的序列模式,其主要算法思想如下:

算法 1. 挖掘频繁序列的窗口衰减算法:StreamFPattern( ).

输入:主滑动窗口含有 k 个长度为 2^p 的子序列,偶合域值ε,局域窗口数域值ζ,频度域值α;

输出:最近最频繁流序列.

(1) 将数据流的数据单元充满 k 个局域窗口,对 k 个子序列按图 1 的方式编号;

(2) 对于任意 X_i,X_j 分别计算子序列之间的偶合度 $Corr(X_i,X_j)$;

(3) if $Corr(X_i,X_j)>=\varepsilon$ then 邻接矩阵 $M[i,j]=1$,表示 X_i,X_j 之间相似;否则,$M[i,j]=0$;

(4) 分别计算邻接矩阵 $M[k,k]$所有行的和 Sum, $f_k(X_i|\varepsilon)=Sum/k$;

(5) if 存在 X_i 满足 $f_k(X_i|\varepsilon)>=\alpha$ then return 最近最频繁流序列 X_i else $k=\lceil k/2\rceil$;

(6) if $k>=\zeta$ then 重复执行步骤(4) else return 当前局域窗口的流序列 X_1.

算法 1 给出了在有 k 个局域窗口的主滑动窗口中计算最近最频繁流序列的方法.假定最近最频繁流序列的变化模式最有可能在将来出现.如果在主滑动窗口上找不到满足条件的频繁序列模式,则在更近时间内搜索.

2.2 数据流能量预测的理论基础

在含有 k 个局域窗口的滑动窗口中,每个局域窗口 i 中包含 2^p 个数据单元 $X_{i,1},X_{i,2},...,X_{i,m}$,其中,$m=2^p$.在局域窗口 i 中,定义流序列能量 $E_i=X_{i,1}^2+X_{i,2}^2+...+X_{i,m}^2$,简称流能量;流能量与数据单元个数之比为流序列的平均流能量,简称平均能量,记为 $\hat{E}_i=E_i/m$.为了方便讨论,本文作如下假设:

(1) 设 $X_{i,1},X_{i,2},...,X_{i,m}$ 为来自相同总体的独立样本,则 $X_{i,1}^2,X_{i,2}^2,...,X_{i,m}^2$ 也为相同总体的独立样本.

(2) $X_{i,1}^2,X_{i,2}^2,...,X_{i,m}^2$ 的方差为 δ_i^2,数学期望为 μ_i.

定理 1. 设 $X_{i,1},X_{i,2},...,X_{i,m}$ 为独立同分布的数据流序列,且 $X_{i,j}^2$ 的数学期望和方差分别为 μ_i 和 $\delta_i^2\neq 0$,平均能量 $\hat{E}_i=(X_{i,1}^2+X_{i,2}^2+...+X_{i,m}^2)/m$,当 $m\to\infty$ 时,$\hat{E}_i\sim N(\mu_i,\delta_i^2/m)$.

证明:因为 $X_{i,1},X_{i,2},...,X_{i,m}$ 为独立同分布的数据流序列,则 $X_{i,1}^2,X_{i,2}^2,...,X_{i,m}^2$ 满足独立同分布.又因任意 $X_{i,j}^2$ 的数学期望和方差分别为 μ_i 和 $\delta_i^2\neq 0$;由中心极限定理可知:当 $m\to\infty$ 时,平均能量 $\hat{E}_i\sim N(\mu_i,\delta_i^2/m)$. □

为了方便讨论,假定若干个局域窗口上的流能量构成随机变量序列为 $\hat{E}_1,\hat{E}_2,...,\hat{E}_t,...$,时刻 t 的最近 k 的能量序列 $\hat{E}_{t-k},\hat{E}_{t-k+1},...,\hat{E}_{t-1}$ 记为 $\hat{E}_1,\hat{E}_2,...,\hat{E}_k$,则随机变量 $\hat{E}_{k+1}$ 表示数据流在未来局域窗口内(简称预测窗口)的平均能量,令 $Y=\hat{E}_{k+1}$,使用 $\hat{E}_1,\hat{E}_2,...,\hat{E}_k$ 预测 Y.假设 $\hat{E}_1,\hat{E}_2,...,\hat{E}_k$ 和 Y 的估计值 $\hat{Y}$ 满足线性关系 $\hat{Y}=\lambda_0+\lambda_1\hat{E}_1+\lambda_2\hat{E}_2+...+\lambda_k\hat{E}_k$ 其中,$\lambda_0,\lambda_1,\lambda_2,...,\lambda_k$ 是待定参数,下面给出定理 2,证明 $\hat{Y}$ 服从正态分布.

定理 2. 设在预测窗口上平均能量的估计值为 $\hat{Y}=\lambda_0+\sum_{i=1}^{k}\lambda_i\hat{E}_i$,则 $\hat{Y}\sim N\left(\lambda_0+\sum_{i=1}^{k}\lambda_i\mu_i,\frac{1}{m}\sum_{i=1}^{k}\lambda_i^2\delta_i^2\right)$.

证明:比较容易证明,在此略. □

使用 $\lambda_0+\lambda_1\hat{E}_1+\lambda_2\hat{E}_2+...+\lambda_k\hat{E}_k$ 预测 Y 的值可能存在误差,因此,引入参数 ε 进行修正,则有线性关系为 $Y=\lambda_0+\lambda_1\hat{E}_1+\lambda_2\hat{E}_2+...+\lambda_k\hat{E}_k+\varepsilon$,$\varepsilon$ 为随机误差,下面将证明随机误差 ε 为正态分布.

定理 3. 设随机误差 $\varepsilon=Y-(\lambda_0+\sum_{i=1}^{k}\lambda_i\hat{E}_i)$,则 $\varepsilon\sim N(\mu,\delta^2)$,且 $\mu=\mu_{k+1}-\left(\lambda_0+\sum_{i=1}^{k}\lambda_i\mu_i\right)$,$\delta^2=\frac{1}{m}\sum_{i=1}^{k+1}\lambda_i^2\delta_i^2$.

证明:比较容易证明,在此略. □

定理 3 揭示了用估计值 $\hat{Y}$ 代表 Y 的预测值存在的误差 ε,且误差 ε 服从正态分布 $N(\mu,\delta^2)$.但线性回归的数学原理要求 Y 与其预测值的误差服从数学期望为 0 的正态分布,因此需要误差 ε 进行调整,构造新的误差量 ε',使得 ε' 符合线性回归的要求,下面的性质 1 给出了 ε' 的构造方法.

性质 1. 设随机变量 ε 服从正态分布 $N(\mu,\delta^2)$且 $\varepsilon'=\varepsilon-\mu$,则 $\varepsilon'\sim N(0,\delta^2)$.

证明:比较容易证明,在此略. □

性质 1 表明,预测关系表示为 $Y=\mu+\lambda_0+\lambda_1\hat{E}_1+\lambda_2\hat{E}_2+...+\lambda_k\hat{E}_k+\varepsilon'$;由于 $\mu+\lambda_0$ 和 ε' 均为待定常数,所以仍然用符号 λ_0 表示 $\mu+\lambda_0$,符号 ε 表示 ε'.设 $\hat{E}_1,\hat{E}_2,...,\hat{E}_k$,$Y$ 的 m 组观测值,基于上述定理和性质,有下列关系成立:

$$Y_1=\lambda_0+\lambda_1\hat{E}_{1,1}+\lambda_2\hat{E}_{1,2}+...+\lambda_k\hat{E}_{1,k}+\varepsilon_1,$$

$$Y_2=\lambda_0+\lambda_1\hat{E}_{2,1}+\lambda_2\hat{E}_{2,2}+...+\lambda_k\hat{E}_{2,k}+\varepsilon_2,$$

$$...$$

$$Y_m=\lambda_0+\lambda_1\hat{E}_{m,1}+\lambda_2\hat{E}_{m,2}+...+\lambda_k\hat{E}_{m,k}+\varepsilon_m.$$

记 $$Y=\begin{bmatrix}Y_1\\Y_2\\\vdots\\Y_m\end{bmatrix},\hat{E}=\begin{bmatrix}1&\hat{E}_{1,1}&\hat{E}_{1,2}&...&\hat{E}_{1,k}\\1&\hat{E}_{2,1}&\hat{E}_{2,2}&...&\hat{E}_{2,k}\\\vdots&\vdots&\vdots&\vdots&\vdots\\1&\hat{E}_{m,1}&\hat{E}_{m,2}&...&\hat{E}_{m,k}\end{bmatrix},\lambda=\begin{bmatrix}\lambda_1\\\lambda_2\\\vdots\\\lambda_m\end{bmatrix},\varepsilon=\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_m\end{bmatrix},$$

将上述方程组表示为矩阵运算:

$$Y=\hat{E}\lambda+\varepsilon \tag{1}$$

称式(1)为预测方程,$\lambda_0,\lambda_1,\lambda_2,...,\lambda_k$ 是预测参数,若式(1)有解,则必须满足 $m\geq k+2$.为确定参数$\lambda_0,\lambda_1,\lambda_2,...,\lambda_k$的值,使用最小二乘法求解,则 Y 的实际值与预测值的误差平方和 $Q_e(\lambda_0,\lambda_1,\lambda_2,...,\lambda_k)=\sum_{i=1}^{m}\left(Y_i-\left(\lambda_0+\sum_{j=1}^{k}\lambda_j\hat{E}_{i,j}\right)\right)^2$,欲使 $Q_e(\lambda_0,\lambda_1,\lambda_2,...,\lambda_k)$达到最小,分别对$\lambda_0,\lambda_1,\lambda_2,...,\lambda_k$ 求偏导数构成方程组,解方程组知 $\lambda=(\hat{E}'\hat{E})^{-1}(\hat{E}'Y)$,其中,$\hat{E}'$ 为 $\hat{E}$ 的转置矩阵.由 $(\hat{E}'\hat{E})^{-1}(\hat{E}'Y)$ 确定的参数$\lambda_0,\lambda_1,\lambda_2,...,\lambda_k$满足 $Q_e(\lambda_0,\lambda_1,\lambda_2,...,\lambda_k)$最小.

2.3 预测模型检验与误差确定

本文借用文献[5,9]的检验方法,主要是检验 Y 与 $\hat{E}_1,\hat{E}_2,...,\hat{E}_k$ 的线性相关性,如果线性相关,则有$\lambda_0,\lambda_1,\lambda_2,...,\lambda_k$不全为 0.假设 $\boldsymbol{H}_0$:$\lambda_0=\lambda_1=\lambda_2=...=\lambda_k=0$ 与 $\boldsymbol{H}_1$:$\lambda_0,\lambda_1,\lambda_2,...,\lambda_k$ 不全为 0.考虑 $U=\sum_{i=1}^{m}(\hat{y}_i-\overline{y})^2$ 与 $Q_e=\sum_{i=1}^{m}(y_i-\hat{y}_i)^2$,其中,$\hat{y}_i$ 是 y_i 的预测值,$\overline{y}=\frac{1}{m}\sum_{i=1}^{m}\hat{E}_i$.文献[8]证明了在 $\boldsymbol{H}_0$ 成立的情况下,有分布 $U/\delta^2\sim\chi^2(k)$和 $Q_e/\delta^2\sim\chi^2(m-k-1)$成立,则 $F=\frac{U/k}{Q_e/(m-k-1)}\sim F(k,m-k-1)$ 成立;在显著水平α下,如果 $F=\frac{U/k}{Q_e/(m-k-1)}>F_{1-\alpha}(k,m-k-1)$,则拒绝 $\boldsymbol{H}_0$,接受 $\boldsymbol{H}_1$,说明模型成立.对于显著水平α,如果预测误差满足 $P(|Y-\hat{Y}|\leq B_\alpha)=1-\alpha$,则认为预测成功.文献[9]给出了计算 B_α值的方法为 $B_\alpha=t_{1-\alpha/2}(m-k-1)\sqrt{Q_e/(m-k-1)}$,$Y$ 的在置信水平为 $1-\alpha$时的置信区间为$[\hat{Y}-B_\alpha,\hat{Y}+B_\alpha]$.

2.4 预测模型的调整策略

当预测模型的预测精度不能满足需要时,则需对模型的回归参数进行调整.本文采用文献[9]中的 F 检验法,分别对参数$\lambda_0,\lambda_1,\lambda_2,...,\lambda_k$ 进行检验,如检验的参数λ_i 是否为 0,统计量 $F=\frac{U_i}{Q_e/(m-k-1)}\sim F(1,m-k-1)$,其中,$Q_e=\sum_{i=1}^{m}(y_i-\hat{y}_i)^2$ 为剩余平方和,$U_i=\lambda_i^2/c_{ii}$ 且 c_{ii} 为矩阵 $(\hat{E}'\hat{E})^{-1}$ 的对角线(i,i)上的元素.对于给定的显著水平α,如果满足 $F=\frac{U_i}{Q_e/(m-k-1)}>F_{1-\alpha}(1,m-k-1)$,则表明$\lambda_i$ 对回归曲线影响较大,保留参数λ_i;否则,放弃参数λ_i.

本节探讨了表示预测区间的能量分布模式的最近最频繁序列模式的挖掘算法,并研究了数据流能量预测模型的理论依据、预测误差确定以及模型有效性检验与动态调整策略,为新算法的设计奠定了理论基础.

3 基于能量与频繁模式的数据流预测算法

在上述研究最近最频繁序列模式的挖掘算法以及能量回归预测理论依据的基础上,设计了基于能量与频繁模式的数据流预测算法,主要的算法思想如下:

(1) 使用缓存中存储的局域窗口的能量构造 $m\times(k+1)$的矩阵 $\hat{E}$ 以及 $m\times1$ 的矩阵 Y.

(2) 计算 $\lambda=(\hat{E}'\hat{E})^{-1}(\hat{E}'Y)$ 的值.

(3) 对模型进行 F 分布在显著性α的有效性检验;如果回归模型 $F\leq F_{1-\alpha}(k,m-k-1)$,则使用第 2.4 节的方法调整回归参数,对回归曲线影响较小的参数$\lambda_i=0$,保留主要参数重新确定参数值,并计算 B_α值.

(4) 使用步骤(2)、步骤(3)得到的预测模型,计算预测窗口的平均能量 $\hat{Y}=\lambda_0+\lambda_1\hat{E}_1+\lambda_2\hat{E}_2+...+\lambda_k\hat{E}_k$.

(5) 使用发现频繁流序列模式的算法 StreamFPattern( )计算最近频繁模式.

(6) 使用文献[7]的小波变换 SubwinDwt( )消除频繁模式中的噪声,并对小波系数进行单位化,称为单位化系数.

(7) 根据频繁模式的单位化系数,将能量预测值的平方根乘以单位化系数.

(8) 用小波算法重构预测窗口的流序列.

为了清楚地表达算法,根据上述思想分解出算法 2 和算法 3,算法 2 确定流能量的回归预测参数;算法 3 在回归参数的基础上预测未来局域窗口的能量,并按照频繁序列模式变换生成预测窗口的预测序列.

算法 2. 能量模型参数回归算法:PmodelRegress( ).

输入:数据流的能量,显著性水平α;

输出:回归参数$\lambda_0,\lambda_1,\lambda_2,...,\lambda_k$及预测误差 B_α.

(1) 使用流能量构造 $m\times(k+1)$的矩阵 $\hat{E}$ 以及 $m\times1$ 的矩阵 Y;

(2) 计算 $\lambda=(\hat{E}'\hat{E})^{-1}(\hat{E}'Y)$ 和 $F=(U/k)/(Q_e/(m-k-1))$;

(3) if $F\le F_{1-\alpha}(k,m-k-1)$ then

(4) {for each $\lambda_i\in\{\lambda_0,\lambda_1,\lambda_2,...,\lambda_k\}$

(5) if $F=(U/k)/(Q_e/(m-k-1))\le F_{1-\alpha}(1,m-k-1)$ then $\lambda_i\neq0$;}

(6) 计算 $B_\alpha=t_{1-\alpha/2}(m-k-1)(Q_e/(m-k-1))^{1/2}$;

(7) Return $\lambda_0,\lambda_1,\lambda_2,...,\lambda_k,B_\alpha$.

算法 2 描述了使用缓存的局域流能量确定回归预测模型参数的方法.语句(4)、语句(5)对模型进行显著性检验,如果不满足要求,则对待定参数$\lambda_0,\lambda_1,\lambda_2,...,\lambda_k$ 进行检查,去掉对预测值影响较小的参数.语句(6)通过剩余平方和以及 t 分布确定误差,确定模型预测值的置信区间.下面给出结合算法 2 对预测窗口的能量预测算法.

算法 3. 数据流预测算法:PredStream( ).

输入:数据流序列,显著性α;

输出:流序列的预测值.

(1) call for 文献[7]的小波变换算法 SubwinDwt( ),消除当前滑动窗口中各子窗口的噪声并计算能量;

(2) call for PmodelRegress( )计算回归参数;

(3) 计算平均能量的预测值 $\hat{Y}=\lambda_0+\lambda_1\hat{E}_1+\lambda_2\hat{E}_2+...+\lambda_k\hat{E}_k$,并计算预测窗口流能量的预测值 $2^p\times\hat{Y}$;

(4) call for StreamFPattern( )计算频繁流序列,作为预测序列的能量分布模式;

(5) 将预测序列的小波系数进行单位化后,用能量预测值开平方乘以各小波系数;

(6) 使用小波算法重构预测窗口的流序列;

(7) return 预测窗口的流序列.

算法 3 描述了融合能量回归和频繁模式的流序列预测算法.相对于传统回归预测方法,该方法预测频繁跳变的数据序列具有明显优势,适合于无显著变化规律,经过分段求平方和后,表现出某种变化规律的数据序列.文献[5]研究了用多元回归的方法预测数据流的局域平均值的方法,而算法 3 不但可以预测平均值,而且能够预测在局域窗口上的数据点.可以将算法 3 推广到多数据流的预测,文献[6,7]研究了多数据流的偶合关系、强偶合数据流变化规律的相似性,使用强偶合流的最近最频繁流序列模式作为预测流的能量分布模式,类似算法 3 的能量回归和频繁模式的小波重构方法,可推广到多数据流进行预测,记为 PredMultiStream 算法(略).

4 实验及性能分析

在 Window 2000 操作系统和 SQL SERVER2000 数据库环境下,用 JAVA 编程语言实现了算法,在 CPU 为 PIII600 和内存为 256M 的计算机上实现.模拟实验数据使用了深圳证券交易所和上海证券交易所的股票交易的 1 000 只股票价格数据,数据量大约为 250 万条记录.对不同情况下的平均相对误差 MRD(mean relative deviation)进行了研究.假设局域窗口所含数据单元的个数为 n,在第 i 次预测实验中,局域窗口的第 j 个数据点的实际值为 v_{ij},预测值为 v'_{ij}.在实验中预测 w 次,则 w 次预测实验后的平均相对误差为 $\frac{1}{nw}\sum_{i=1}^{w}\sum_{j=1}^{n}(|v_{ij}-v'_{ij}|/v_{ij})$.我们进行了如下几方面的实验.

4.1 局域窗口宽度对平均误差的影响

实验研究了局域窗口宽度对预测值的平均相对误差的影响.在实验中,局域窗口宽度分别取 8,16,32,64,128,256,512 等,使用算法 3 分别对 100 个流序列进行了实验.实验观察表明,当局域窗口宽度增大时,平均相对误差 MRD 呈先下降,然后逐渐增大的趋势,如图 2 所示.主要原因是由局域窗口宽度和回归模型综合作用的结果.

(1) 局域窗口宽度较小,最近最频繁序列模式更能表示未来的变化模式,因而误差较小;反之,误差较大;

(2) 在 $m\to\infty$时导出的统计量 $\hat{E}_i$ 服从正态分布 $N(\mu_i,\delta_i^2/m)$,一般要求 $m\geq 50$,从而使得在较大窗口时可能有较小的误差;反之,误差较大.这两方面因素的综合作用,平均相对误差表现为图 2 所示的变化趋势.

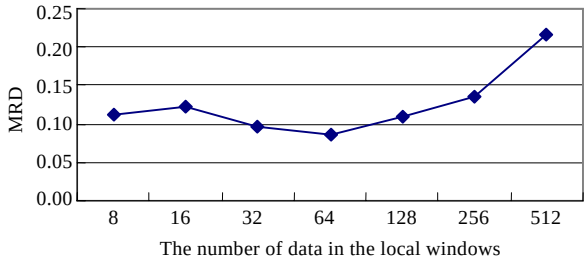


Fig.2 Mean relative deviation vs local windows

图 2 平均相对误差与局域窗口

4.2 不同预测方法的平均相对误差比较

对不同预测方法所产生的误差进行对比实验:(1) 传统的线性回归方法(简记为 LineRegress),如:$y=ax+b$,对主滑动窗口的数据点不划分为局域窗口,在主窗口上使用最小二乘法确定系数 a 和 b;(2) 在文献[5]中介绍了预测区间聚集值的算法(简记为 P_AVG).由于算法 P_AVG 是预测区间内的 AVG 平均值或 SUM 求和,为了使比较具有公平性,在主滑动窗口取不同宽度时,将等宽预测区间上平均值的平均相对误差进行对比实验.如图 3 所示,LineRegress 算法的平均相对误差相对较大,本文的 PredStream 算法次之,P_AVG 算法最小.但 PredStream 算法和 P_AVG 算法比较接近,主要原因是 LineRegress 算法不适合波动性较大序列的预测.

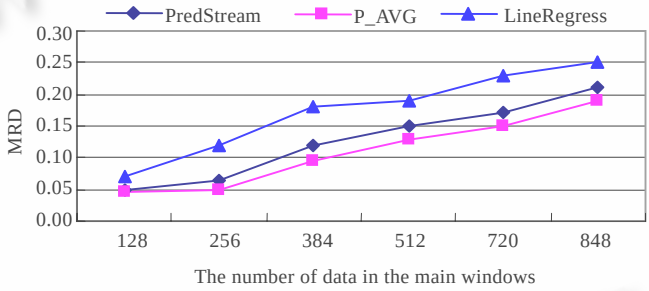


Fig.3 Mean relative deviation vs prediction algorithms

图 3 平均相对误差与预测算法

4.3 预测模型的正确率

使用了 F 检验判别能量预测模型的正确性,计算 F 统计量和在给定显著水平 α 下 $F_{1-\alpha}(k,m-k-1)$的值,若满足 $F>F_{1-\alpha}(k,m-k-1)$,则认为当前系统数据缓冲区中数据适合使用线性回归模型来预测,即模型的正确性;模型的正确次数与总检验次数的比值称为模型正确率 MRR(model right ratio).该实验选用了 100 个数据流序列,显著水平 $\alpha=0.05$,k 为回归参数的个数,分别对 k 取不同值时进行了实验.图 4 给出了在不同的 k 下,模型平均正确率的变化情况.从图 4 可以看出,平均正确率随 k 的增加而增加,增加到一定时候,正确率有所下降,说明多元线性回归模型的参数数量选择要适量.

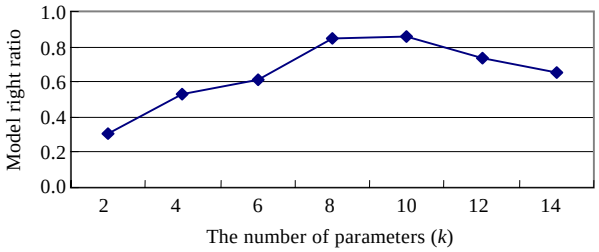


Fig.4 Model right ratio vs. the number of parameter

图 4 模型正确率与参数数量 k

4.4 多数据流与单数据流的比较

将多数据流预测算法 PredMultiStream 的预测平均相对误差与单数据流预测算法 PredStream 的预测平均相对误差进行了比较.局域窗口的宽度分别取 8,16,32,64,128,256,512 等,分别对 100 个流序列实验算法 3 进行了实验.通过实验观察发现,PredMultiStream 算法的预测平均相对误差微大于 PredStream 算法的预测平均相对误差,如图 5 所示.主要原因是 PredMultiStream 算法使用偶合特征流的频繁模式表示预测流的变化模式.

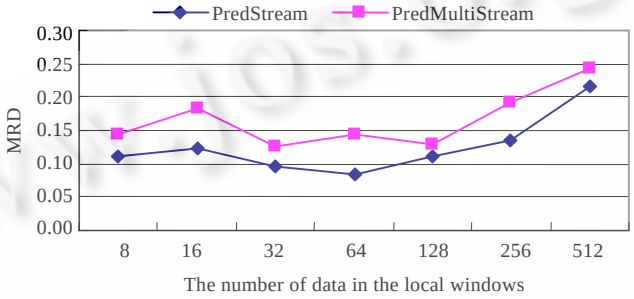


Fig.5 Multiple data-streams vs. single data-stream

图 5 多数据流与单数据流的比较

5 结 论

本文研究了数据流能量回归与最近频繁模式挖掘相融合的预测新思路.相对于传统的回归预测方法,新方法预测频繁变化的数据流具有显著优势,通过在局域窗口上的数据值平方取平均,消除了数据变化的振荡性,适合于某些非规律性变化,经过分段求平方和后,呈现某种规律性变化的数据序列.在多数据流环境中,如果知道数据流之间的偶合特性以及局部能量变化,就可以通过一个数据流预测另一个数据流的变化规律.

References:

[1] 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.

[2] 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.

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

[4] 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 WAIM 2005. Berlin, Heidelberg: Springer-Verlag, 2005. 404−415.

[5] Li JZ, Guo LJ, Zhang DD, Wang WP. Processing algorithms for predictive aggregate queries over data streams. Journal of Software, 2005,16(7):1252−1261 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/16/1252.htm

[6] 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. Proc. of the Advances in Computer Science: The 10th Asian Computing Science Conf. Kunming: Springer-Verlag, 2005. 270−271.

[7] Chen AL, Tang CJ, Yuan CA, Peng J, Hu JJ. Anti-Noise algorithm for mining asynchronous coincidence pattern in multi-streams. Journal of Software, 2006,17(8):1753−1763 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/1753.htm

[8] 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 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/18/177.htm

[9] Yuan ZF, Zhou JY. Multianalysis and Statistics. Beijing: Science Press, 2003. 158−187 (in Chinese).

附中文参考文献:

[5] 李建中,郭龙江,张冬冬,王伟平.数据流上的预测聚集查询处理算法.软件学报,2005,16(7):1252−1261. http://www.jos.org.cn/1000-9825/16/1252.htm

[7] 陈安龙,唐常杰,元昌安,彭京,胡建军.挖掘多数据流的异步偶合模式的抗噪声算法.软件学报,2006,17(8):1753−1763. http://www.jos.org.cn/1000-9825/17/1753.htm

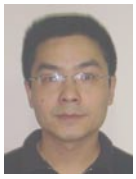
[8] 陈安龙,唐常杰,元昌安,朱明放,段磊.基于小波和偶合特征的多数据流压缩算法.软件学报,2007,18(2):177−184. http://www.jos.org.cn/1000-9825/18/177.htm

[9] 袁志发,周静芋,主编.多元统计分析.北京:科学出版社,2003.158−187.

陈安龙(1971－),男,四川仪陇人,博士,讲师,主要研究领域为数据库与知识工程,数据挖掘.

傅彦(1962－),女,教授,博士生导师,主要研究领域为数据库与知识工程,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库与知识工程,数据挖掘.



廖勇(1968－),男,博士生,主要研究领域为信息安全,数据挖掘.