多数据流上的联机方差分析研究

王小龙,马瑞民
WANG Xiao-long,MA Rui-min

大庆石油学院 计算机与信息技术学院,黑龙江 大庆 163318
School of Computer and Information Technology,Daqing Petroleum Institute,Daqing,Heilongjiang 163318,China
E-mail:wangxiaolongdqpi@yahoo.com.cn

WANG Xiao-long,MA Rui-min.Research on on-line analysis of variances over data streams.Computer Engineering and Applications,2007,43(15):180-183.

Abstract: It is important to analyze variances over data streams.In this paper, each of single data stream that is included in data streams has same attributes set as other,the unit of single data stream is tuple,and same attributes set is included in each of these tuples.Reservoir algorithm is used to sample from these single data stream respectively,then some multiple snapshot windows are constructed,the relationship between these single data stream and multiple snapshot windows is bijective mapping, the relationship between attributes that are included in tuple and snapshot windows that are included in relative multiple snapshot windows is bijective mapping,and the relationship between basic windows that belong to same snapshot window and the attribute values that come from different single data stream is bijective mapping as well,that is,these attribute values come from same attribute that is comprised by different single data stream.Variances are analyzed based on these independent sample values. Data of these multiple snapshot windows can be processed orderly.In a word,a serialization method is used to analyze parallel data streams.The analytical and experimental results show that this analytic method is logical and effective.

Key words: data streams;snapshot window;bijective mapping;model;on-line analysis of variance

摘 要:多数据流上的联机方差分析是一个有意义的研究问题。针对以元组为单位流入的具有相同属性集的多支单数据流组成的多数据流,提出了分别对每支单数据流进行蓄水池抽样,构造一一对应于各单数据流的若干个多快照窗口,即两者之间是双射关系,可以将多快照窗口串行置于主存中,将元组包含的属性与多快照窗口中的各个快照窗口一一对应,且使得同一快照窗口中的各基本窗口与取自其对应的单数据流的属性值样本一一对应,然后对这些相互独立的样本进行方差分析。按顺序串行处理各个多快照窗口中的数据,就可以用串行化的方法来实现并行的多数据流上的联机方差分析。理论分析与实验表明,该方法是合理的和有效的。

关键词:多数据流;快照窗口;双射;模型;联机方差分析
文章编号:1002-8331(2007)15-0180-04 **文献标识码**:A **中图分类号**:TP311

1 引言

近几年,新出现了无线传感器网络(Wireless Sensor Networks,WSNs)数据管理、互联网络实时数据分析、XML 数据流管理和分析、对等计算(Peer-to-Peer/P2P computing)数据处理、普适计算(Ubiquitous/Pervasive computing)数据处理等诸多密集实时数据应用领域。在这些应用中,数据不是存储在介质上的有限的静态数据集合,而是以流的形式存在的,是由连续的、时变的、有序的、数量趋于无限的数据元素组成的数据流(Data Stream,DS)。对于连续流入的单个支流可以称为单数据流(Single data stream),对于连续并行流入的多个支流可以称为多数据流(Data Streams,DSs)。数据流可以被认为是对一个或若干个实体或对象的描述,进而可以认为其数据元素是数据流元组(Tuple of data stream),且每个元组有其属性集以及属性值集。

已有的数据流研究工作大多集中在数据流管理系统[1]、数据流上的连接[2]或聚集查询[3]、数据流挖掘算法[4]等方面,在多数据流上进行联机方差分析(on-line analysis of variance)的研究工作还较少,对多数据流进行联机方差分析可以实时地推断出数据流的一些重要统计特性,同时这也将有助于其它相关的数据流问题的解决,可以在对多数据流进行联机方差分析的基础上,对数据流进行查询或挖掘分析。将对连续流入的多支单数据流抽样得到的元组的属性与属性值分别作为随机变量与随机变量的值。方差(Variance)是随机变量的数据特征之一,它反映了随机变量取值的集中与分散程度,即反映了随机变量的值相对于其均值的偏离程度。在一些实际应用中,对这种偏离

基金项目:黑龙江省自然科学基金(the Natural Science Foundation of Heilongjiang Province of China under Grant No.F200603);黑龙江省教育厅科学技术研究项目(No.11521008)。

作者简介:王小龙,男,讲师,计算机学会会员,主要研究方向为数据流管理和挖掘技术;马瑞民,男,教授,计算机学会会员,主要研究方向为数据库理论、技术及应用。

程度进行计算和分析是重要的。总之,对多数据流进行实时的联机方差分析具有一定的研究意义。

2 模型

可以使用蓄水池(抽样)算法(Reservoir algorithm)[5]分别对各支单数据流进行连续抽样。该抽样算法仅进行一遍数据流扫描,且能使得单数据流中各个元组以相同的概率被选取。值得指出的是,既然单数据流中各个元组是以相同的概率被抽取的,那么这些数据流元组中的属性及属性值也将以同样的概率被抽取。

2.1 快照窗口的分解及串行化处理模型

在数据流上进行联机方差分析可以采用多快照窗口(Multiple Snapshot Windows,MSW)的方法顺序地进行处理,各多快照窗口一一对应(双射)于各个以元组为单位的单数据流,可以将并行的多快照窗口串行处理,这样也使得对连续流入的多数据流的处理被限制在预定的时间段内。数据流上的各多快照窗口中的每个快照窗口(Snapshot Window,SW)需要被分解为多个基本窗口(Basic Window,BW),其串行化处理的描述如图 1 所示,其中的 SW_1、SW_2 等是某个多快照窗口中的各快照窗口,每个快照窗口由多个基本窗口组成,例如 SW_1 是由基本窗口 $bw_1,bw_2,\cdots,bw_w$ 构成的,w 为该快照窗口中的基本窗口的数量,其它依次类推。同一快照窗口中的各基本窗口与取自其对应的单数据流的属性值样本一一对应,例如,快照窗口$\{\cdots,\{x_{i-1,1},x_{i\,1},x_{i+1,1},\cdots\},\{x_{i-1,2},x_{i\,2},x_{i+1,2},\cdots\},\cdots,\{x_{i-1,s},x_{i\,s},x_{i+1,s},\cdots\},\cdots\}$,其中,$\{x_{i-1,1},x_{i\,1},x_{i+1,1},\cdots\},\{x_{i-1,2},x_{i\,2},x_{i+1,2},\cdots\},\cdots,\{x_{i-1,s},x_{i\,s},x_{i+1,s},\cdots\}$分别为基本窗口 $bw_1,bw_2,\cdots,bw_w$ 中的水库抽样属性值集合,它们属于同一属性,但来自不同的单数据流,s 为多数据流包含的支流数量,例如,$\{x_{i-1,1},x_{i\,1},x_{i+1,1},\cdots\}$中的 $x_{i-1,1},x_{i\,1},x_{i+1,1}$ 为来自第 1 支单数据流某个属性的第 $i-1$ 个、第 i 个、第 $i+1$ 个属性值,$\{x_{i-1,2},x_{i\,2},x_{i+1,2},\cdots\}$中的 $x_{i-1,2},x_{i\,2},x_{i+1,2}$ 为来自第 2 支单数据流的该属性的第 $i-1$ 个、第 i 个、第 $i+1$ 个属性值,其它依次类推。

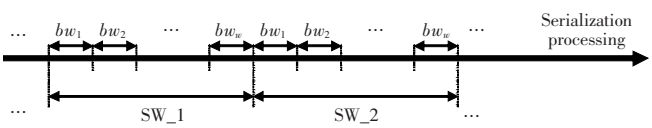

图 1 快照窗口的分解模型及串行化处理模型

2.2 多数据流上的联机方差分析模型

多数据流上的联机方差分析模型如图 2 所示,图中的多数据流包含 s 条支流,MSW_1,MSW_2 等是双射于 s 条单数据流的多快照窗口,其中的 SW_1,SW_2,… 等是双射于属性的快照窗口,其中的 m 表示快照窗口的数量,每个快照窗口都由 $bw_1,bw_2,\cdots$ 等多个基本窗口组成,每个基本窗口包含一个样本,并假设不同的基本窗口中的样本之间相互独立。说明:在逻辑上,所生成的各多快照窗口是按顺序串行进入处理器的。

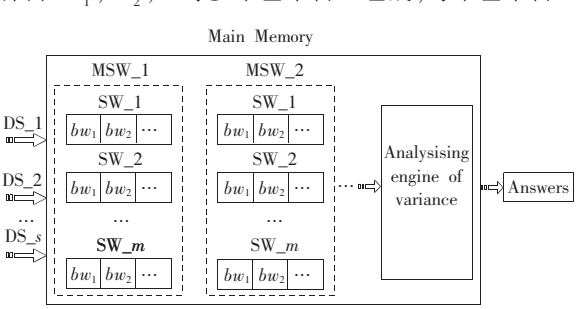

图 2 多数据流上的联机方差分析模型

2.3 数学模型

实际上,在上文中将原始的多数据流进行了三级层次的界定:多快照窗口、快照窗口、基本窗口。下文中将主要进行快照窗口层次上的联机方差分析。

已知快照窗口与属性之间是双射关系,即两者严格地一一对应。将每个属性都作为单随机变量,先将其对应的快照窗口中的属性值的集合(由其中的全部基本窗口中的属性值子集组成)作为被分析的对象。

假设每个快照窗口中各基本窗口中的样本相互独立,且取自具有相同方差 σ^2,均值分别为 $\mu_j(j=1,2,\cdots,w)$的正态数据流总体 $N(\mu_j,\sigma^2)$,σ^2 与 μ_j 均未知,则样本点的值 $x_{ij}\sim N(\mu_j,\sigma^2)$,$i=1,2,\cdots,b_j,j=1,2,\cdots,w,b_j$ 为第 j 个基本窗口中属性值的数量,w 为同一快照窗口中基本窗口的数量,进而有 $x_{ij}-\mu_j\sim N(0,\sigma^2)$,若令 $\varepsilon_{ij}=x_{ij}-\mu_j$,则 $\varepsilon_{ij}\sim N(0,\sigma^2)$,且各 ε_{ij} 相互独立。

令 $b=\sum_{j=1}^{w}b_j,\mu=\frac{1}{b}\sum_{j=1}^{w}b_j\mu_j,\delta_j=\mu_j-\mu$,可知 δ_j 为快照窗口中的第 j 个基本窗口中平均属性值与快照窗口中所有属性值的总平均值的差异,同时,可以推断出:$b_1\delta_1+b_2\delta_2+\cdots+b_w\delta_w=\sum_{j=1}^{w}b_j\delta_j=0$,进而得到进行多数据流上的联机方差分析的数学模型:

$$\begin{cases}x_{ij}=\mu+\delta_j+\varepsilon_{ij}\\ \sum_{j=1}^{w}b_j\delta_j=0\\ \varepsilon_{ij}\sim N(0,\sigma^2)\end{cases}$$

说明:以上三种模型的建立过程实际上就是构造多数据流概要数据结构(Synopsis data structure)的方法。

3 多数据流上的联机方差分析过程

为了计算 $F(F=\frac{\bar{S}_A}{\bar{S}_E})$分布的值,需要先计算 S_A、S_E(这些符号的含义请见下文的相关分析内容)以及两者的自由度,下面分别给出分析。

3.1 S_T、S_A、S_E 的分析

单个快照窗口中的全部属性值的总平均 $\bar{x}=\frac{1}{b}\sum_{j=1}^{w}\sum_{i=1}^{b_j}x_{ij}$,则能体现全部属性值之间差异的总变差 $S_T=\sum_{j=1}^{w}\sum_{i=1}^{b_j}(x_{ij}-\bar{x})^2$。单个基本窗口中的样本平均值为 $\bar{x}_j=\frac{1}{b_j}\sum_{i=1}^{b_j}x_{ij}$,则由文献[6]的分析可知:

$$S_T=\sum_{j=1}^{w}\sum_{i=1}^{b_j}[(x_{ij}-\bar{x}_j)+(\bar{x}_j-\bar{x})]^2=\sum_{j=1}^{w}\sum_{i=1}^{b_j}(x_{ij}-\bar{x}_j)^2+\sum_{j=1}^{w}\sum_{i=1}^{b_j}(\bar{x}_j-\bar{x})^2$$

取:

$$(1)S_E=\sum_{j=1}^{w}\sum_{i=1}^{b_j}(x_{ij}-\bar{x}_j)^2$$

其中,各个 $(x_{ij}-\bar{x}_j)^2$ 为单个快照窗口中第 j 个基本窗口中的各样本观察值与样本均值的差异。

(2)$S_A=\sum_{j=1}^{w}\sum_{i=1}^{b_j}(\bar{x}_j-\bar{x})^2=\sum_{j=1}^{w}b_j\bar{x}_j^2-b\bar{x}^2$

其中,各个 $b_j(\bar{x}_j-\bar{x})^2$ 为单个快照窗口中第 j 个基本窗口中的样本平均值与总平均值的差异。

则有:$S_T=S_E+S_A$

若令 $A_{\cdot j}=\sum_{i=1}^{b_j}x_{ij}$,$A_{\cdot\cdot}=\sum_{j=1}^{w}\sum_{i=1}^{b_j}x_{ij}$,则有:

$$S_T=\sum_{j=1}^{w}\sum_{i=1}^{b_j}x_{ij}^2-b\bar{x}^2=\sum_{j=1}^{w}\sum_{i=1}^{b_j}x_{ij}^2-\left(\frac{1}{n}\sum_{j=1}^{w}\sum_{i=1}^{b_j}x_{ij}\right)^2=$$

$$\sum_{j=1}^{w}\sum_{i=1}^{b_j}x_{ij}^2-b\left(\frac{1}{b}A_{\cdot\cdot}\right)^2=\sum_{j=1}^{w}\sum_{i=1}^{b_j}x_{ij}^2-\frac{A_{\cdot\cdot}^2}{b}$$

$$S_A=\sum_{j=1}^{w}b_j\bar{x}_j^2-b\bar{x}^2=\sum_{j=1}^{w}b_j\left(\frac{1}{b_j}\sum_{i=1}^{b_j}x_{ij}\right)^2-b\left(\frac{1}{b}\sum_{j=1}^{w}\sum_{i=1}^{b_j}x_{ij}\right)^2=$$

$$\sum_{j=1}^{w}b_j\left(\frac{1}{b_j}A_{\cdot j}\right)^2-b\left(\frac{1}{b}A_{\cdot\cdot}\right)^2=\sum_{j=1}^{w}\frac{A_{\cdot j}^2}{b_j}-\frac{A_{\cdot\cdot}^2}{b}$$

则 S_E、S_T、S_A 的计算公式可归纳为:

$$\left.\begin{aligned}S_E&=S_T-S_A\\S_T&=\sum_{j=1}^{w}\sum_{i=1}^{b_j}x_{ij}^2-\frac{A_{\cdot\cdot}^2}{b}\\S_A&=\sum_{j=1}^{w}\frac{A_{\cdot j}^2}{b_j}-\frac{A_{\cdot\cdot}^2}{b}\end{aligned}\right\}$$

3.2 S_E、S_A 的自由度

定理 1 S_E 的自由度为 $b-w$,其中 $b=\sum_{j=1}^{w}b_j$,$j=1,2,\cdots,w$,b_j 为第 j 个基本窗口中属性值的数量,w 为同一快照窗口中基本窗口的数量。

证明

$$S_E=\sum_{j=1}^{w}\sum_{i=1}^{b_j}(x_{ij}-\bar{x}_j)^2=\sum_{i=1}^{b_1}(x_{i1}-\bar{x}_1)^2+\sum_{i=1}^{b_2}(x_{i2}-\bar{x}_2)^2+\cdots+\sum_{i=1}^{b_w}(x_{iw}-\bar{x}_w)^2$$

根据文献[6]中的定义 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$,即 $\sum_{i=1}^{n}(X_i-\bar{X})^2=(n-1)S^2$,以及公式 $\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$,其中,$X_1,X_2,\cdots,X_n$ 是来自总体 $N(\mu,\sigma^2)$ 的样本,$\bar{X}$ 是该样本的均值,S^2 是该样本的方差(关于这些符号的含义请参阅文献[6]),可知:

(1)$\frac{\sum_{i=1}^{b_j}(x_{ij}-\bar{x}_j)^2}{\sigma^2}\sim\chi^2(b_j-1)$;

(2)因为 $\frac{S_E}{\sigma^2}=\frac{\left(\sum_{j=1}^{w}(b_j-1)\right)\left(\sum_{i=1}^{b_j}(x_{ij}-\bar{x}_j)^2\right)}{\sigma^2}$,所以 $\frac{S_E}{\sigma^2}\sim\chi^2\left(\sum_{j=1}^{w}(b_j-1)\right)$,取 $b=\sum_{j=1}^{w}b_j$,则 $\frac{S_E}{\sigma^2}\sim\chi^2(b-w)$,即 S_E 的自由度为 $b-w$。证毕。

定理 2 S_A 的自由度为 $w-1$,w 为同一快照窗口中基本窗口的数量。

证明 $S_A=\sum_{j=1}^{w}\sum_{i=1}^{b_j}(\bar{x}_j-\bar{x})^2=(\sqrt{b_1}(\bar{x}_1-\bar{x}))^2+(\sqrt{b_2}(\bar{x}_2-\bar{x}))^2+\cdots+(\sqrt{b_j}(\bar{x}_j-\bar{x}))^2$

与定理 1 同理,可得:S_A 的自由度为 $w-1$。证毕。

3.3 $\bar{S}_A$、$\bar{S}_E$、F 的分析

取 $\bar{S}_A=\frac{S_A}{w-1}$,$\bar{S}_E=\frac{S_E}{b-w}$,对于分布 $F=\frac{\bar{S}_A}{\bar{S}_E}=\frac{S_A/(w-1)}{S_E/(b-w)}$,当 $\delta_j=0$ 时,$\frac{S_A/(w-1)}{S_E/(b-w)}=\frac{\left(\frac{S_A}{\sigma^2}\right)/(w-1)}{\left(\frac{S_E}{\sigma^2}\right)/(b-w)}\sim F(w-1,b-w)$,所以当 $F_\alpha(w-1,b-w)\leqslant F$ 时,表示被分析的数据之间出现显著差异。例如,若此时取 $\alpha=0.05$,则表明 F 值在 $\alpha=0.05$ 的水平上显著,即可以 95%的可靠性(冒 5%的风险)推断:被分析的快照窗口中的属性值之间出现显著差异。

3.4 多数据流上的联机方差分析算法

多数据流上的联机方差分析算法的主要内容如下所述。

Input:Data streams;Δt;i;j;b_j;w;α.

Output:The analytical results of varience over data streams.

Steps:

Step1:To create synopsis data structure.

Step2:FOR d=1 to s //d is the number of single data stream

Step3: FOR i=1 to m

Step4: To compute $S_T,S_A,S_E,w-1,b-w,\bar{S}_A,\bar{S}_E,F$.

Step5: IF $F_\alpha(w-1,b-w)\leqslant F$

Step6: Output:①"There is a notable difference! "

②The results of $S_T,S_A,S_E,w-1,b-w,\bar{S}_A,\bar{S}_E,F$.

Step7: ELSE

Step8: Output:"There is not notable difference! "

Setp9: ENDFOR

Setp10:ENDFOR

说明:

(1)该算法中的符号的含义请参阅上文分析过程中的相关说明。

(2)终止于确定时刻的算法空间复杂度为该算法的规模;终止于确定时刻的算法时间复杂度为 O(smb),其中,s 为多数据流包含的支流数量,m 表示快照窗口的数量,$b=\sum_{j=1}^{w}b_j$,这里主要考虑了 S_A、S_E 的计算复杂度。

4 仿真实验

4.1 实验环境

OS 是 Windows 2000,CPU 为 P Ⅳ,主频为 2.4 GHz,主存为 256 MB。算法实现的工具是 VC。

4.2 部分实验数据

该实验使用小麦的监测数据[11]构造多数据流。其中的部分实验数据为:

……

SW_1:{{2.47,2.60,2.40,2.67,2.87,4.80,3.40},{3.73,4.13,4.93,3.27,4.80,4.13,3.86},{3.20,3.27,4.13,2.07,4.07,3.80,2.47}}

SW_2:{{2.07,1.67,2.07,2.80,2.07,2.73,2.33},{2.20,3.07,2.60,2.80,2.73,2.27,2.80},{2.00,2.27,2.80,2.40,2.13,2.00,1.73}}

SW_3:{{79.71,83.18,84.42,79.71,83.12,82.80,85.67},{83.26,83.18,84.42,84.42,81.41,85.68,80.17},{82.65,83.19,82.91,84.01,83.15,85.49,85.52}}

表 1 取 α=0.10 时的实验计算结果

Snapshot windows	S_T	S_A	S_E	$w-1$	$b-w$	$\overline{S}_A$	$\overline{S}_E$	F	Notable difference
SW_1	14.67	4.56	10.11	2	18	2.28	0.56	4.07	Yes
SW_2	3.06	0.83	2.23	2	18	0.42	0.12	3.50	Yes
SW_3	65.27	4.94	60.33	2	18	2.47	3.35	0.74	No
SW_4	0.65	0.02	0.63	2	18	0.01	0.04	0.29	No
SW_5	12.62	0.13	12.94	2	18	0.07	0.70	0.09	No
SW_6	6.00	0.45	5.55	2	18	0.23	0.31	0.74	No
SW_7	3.50	0.17	3.33	2	18	0.09	0.19	0.46	No
SW_8	11.14	0.34	10.80	2	18	0.17	0.60	0.28	No
SW_9	1.42	0.06	1.36	2	18	0.03	0.08	0.38	No
SW_10	16 789.45	1 993.20	14 796.25	2	18	996.60	822.01	1.21	No

SW_4:{{4.66,4.30,4.48,4.67,4.71,4.78,4.86},{4.76,4.66,4.64,4.78,4.41,4.88,4.54},{4.50,4.27,4.51,4.52,4.89,4.76,4.75}}

SW_5:{{3.73,3.33,3.47,3.67,3.73,4.00,4.07},{4.20,4.07,3.47,3.53,4.00,4.00,3.20},{3.93,3.67,4.07,4.00,4.13,3.80,3.73}}

SW_6:{{11.15,11.57,10.80,11.29,12.30,10.65,11.15},{11.15,11.30,11.49,11.57,11.14,12.30,11.47},{11.37,12.10,11.29,11.41,10.15,11.42,11.50}}

SW_7:{{10.79,11.06,10.49,10.83,11.57,11.75,11.75},{11.17,11.85,11.57,11.22,10.79,12.03,10.87},{10.91,10.87,11.22,11.44,11.08,11.59,11.05}}

SW_8:{{18.07,19.87,20.40,18.47,18.07,19.87,18.07},{18.07,18.67,19.60,18.40,19.13,18.60,19.60},{18.07,18.07,19.60,19.40,19.40,18.87,19.47}}

SW_9:{{2.30,2.16,212,2.82,2.20,2.78,2.20},{2.21,2.52,2.23,2.53,2.42,2.17,2.36},{2.44,2.52,1.97,2.68,2.04,2.79,2.82}}

SW_10:{{128.67,122.67,149.80,181.47,137.53,157.13,128.67},{128.67,201.67,154.73,144.47,181.47,178.07,141.73},{121.20,150.60,175.07,176.00,122.40,209.60,209.60}}

……

4.3 实验结果

对应于上述实验数据的实验结果如表 1 所示。

说明:

(1)快照窗口与属性是双射关系。

(2)实验中,每个快照窗口中包含 3 个基本窗口(相当于对三重数据流进行分析),每个基本窗口中包含 7 个属性值。

(3)在计算各个项的结果的过程中,中间计算结果与最后的计算结果皆保留到小数点后两位。

(4)因为 $F_{0.10}(2,18)=2.62$,所以当 $F<2.26$ 时,则 F 值在 $\alpha=0.10$ 的水平上显著,即可以以 90%的可靠性(冒 10%的风险)推断:相关的快照窗口中的基本窗口之间的差异不显著;当 $F>2.26$ 时,可以以 90%的可靠性推断:相关的快照窗口中的基本窗口之间的差异显著。

(5)在本次实验中取 α=0.10。但是,当取 α=0.05 时,$F_{0.05}(2,18)=3.55>3.50$,此时快照窗口 Snapshot_2 的实验计算结果为:差异不显著。

(6)按顺序循环处理各个快照窗口中的数据,可以有效地进行多数据流上的联机方差分析。

5 结束语

该文使用蓄水池算法、多快照窗口、快照窗口与基本窗口等方法构造了多数据流的概要数据结构,论述了多数据流上的联机方差分析过程,并进行了实验验证。值得指出的是,这其中亦包含将并行的多数据流进行串行化处理的思想。后续的研究工作包括:对于一个或若干实体或对象,以及对于不同的挖掘任务,各属性的重要性往往是不同的,可以先给属性赋予不同的权重,使得越重要的属性的权重越大,如果在进行联机方差分析的过程中发现某些权重较大或用户感兴趣的属性发生显著变化,可以继续对多数据流进行更深入的分析或挖掘。

(收稿日期:2006 年 9 月)

参考文献:

[1] Arasu A,Babcock B,Babu S,et al.STREAM:the stanford stream data manager[C]//Halevy A Y,Ives Z G,AnHai Doan.Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data.USA:ACM Press,2003:665.

[2] Golab L,Tarner M.Processing sliding window multi-joins in continuous queries over data streams[C]//Freytag J C,Lockemann P C,Abiteboul S,et al.Proceedings of 29th International Conference on Very Large Data Bases.USA:ACM Press,2003:500-511.

[3] Dobra A,Garofalakis M,Gehrke J,et al.Processing complex aggregate queries over data streams[C]//Franklin M J,Moon B,Ailamaki A.Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data.USA:ACM Press,2002:61-72.

[4] Wang Hai-xun,Fan Wei,Yu P S,et al.Mining concept-drifting data streams using ensemble classifiers[C]//Getoor L,Senator T E,Domingos P,et al.Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.USA:ACM Press,2003:226-235.

[5] Vitter J S.Random sampling with a reservoir[J].ACM Trans on Mathematical Software,1985,11(1):37-57.

[6] 盛骤,谢式千,潘承毅.概率论与数理统计[M].2 版.北京:高等教育出版社,1989.

[7] 汪璇,杨国才,武伟,等.基于记录过滤的粗糙集属性约简算法研究[J].计算机工程与应用,2005,41(36):175-178.