基于滑动窗口密度聚类的数据流偏倚采样算法

胡志冬 任永功 杨 雪
(辽宁师范大学计算机与信息技术学院 大连 116029)

摘 要 对于移动计算领域的移动对象轨迹数据流的管理,最普遍采用的技术手段是采样技术,而传统的均匀采样易丢失一些关键的变化数据,造成信息丢失现象。针对这一问题,提出一种基于概率密度聚类的数据流偏倚采样算法。该算法在滑动窗口模型下,充分利用了轨迹数据流自身的分布特性,结合偏倚采样算法思想克服了均匀采样的数据丢失问题。算法首先采用基于数据存在密度的聚类技术将滑动窗口划分为强簇、弱簇和过度簇,然后针对不同的簇给予不同的采样率,进行偏倚采样,进而得到最终的数据流摘要。经过实际数据集的实验检测,证明算法较好地保证了采样质量,并具有较快的数据处理能力。

关键词 轨迹数据流,滑动窗口,密度聚类,偏倚采样
中图法分类号 TP301 **文献标识码** A

Bias Sampling Data Stream Based on Sliding Window Density Clustering Algorithm Research

HU Zhi-dong REN Yong-gong YANG Xue
(School of Computer and Information Technology, Liaoning Normal University, Dalian 116029, China)

Abstract In management of the mobile object trajectory data stream in the field of mobile computing, the most commonly used technical means is sampling techniques, but the traditional uniform sampling is easy to lose some of the key changes in data, resulting in the phenomenon of loss of information. To solve this problem, we proposed a data stream based on the probability density clustering bias sampling algorithm. The algorithm in a sliding window model, makes full use of the distribution of characteristics of the the trajectory data stream itself, combines a bias sampling algorithm ideology to overcome uniformly sampled data loss problems. Firstly the sliding window is divided into a strong cluster clustering techniques based on density data exists, weak clusters and excessive cluster, and then different sampling rates for different clusters biased sampling are given, thereby to obtain a final summary of the data stream. The experimental testing results of the set of actual data show that the algorithm ensures the sampling quality and has faster data processing capability.

Keywords Trajectory data stream, Sliding window, Density clustering, Bias sampling

1 引言

近年来,随着通信技术与软硬件设备的不断发展,移动计算快速普及,许多应用领域都产生了对大量数据流的处理需求,如金融证券管理、网络监控、Web日志挖掘、数据在线分析等等,对这类数据进行挖掘分析已经成为一个热点问题。不同于传统数据,这种数据流往往具有如下特点:1)数据流总量具有无限性;2)数据流自身具有明显的时变性。这些特点要求所设计的数据流分析挖掘算法必须能够对数据进行实时挖掘,即对每一个元组都能快速处理,并将处理结果实时输出。在动态环境中连续产生的数据流往往数量巨大,并能够无限地流动与快速变化,数据不能完全存储在数据库中,这种情况下,如果能够从数据流中选取出部分有效数据临时存储在内存中,并利用这些有效数据进行分析挖掘,以便获得一个近似的但又满足可信度的分析结果,显然是一种可以实施的有效方法。于是,一个首先要解决的问题就是,如何设计出筛选保留有效数据的最佳策略。

从数据流中筛选保留有效数据的过程又叫数据摘要构造过程。目前进行数据摘要常用的方法有采样法[1]、概率直方图法[2]、小波分析法[3]、草图绘制法[4]等等,每一种摘要的构造结构都有其自身的特点,依据待解决问题的不同类型,可以选择不同的与之相适合的摘要结构。移动对象的轨迹数据无论在时间上还是空间上都是连续的,具有很明显的时空特征,考虑到该类数据流的自身特性,本文探讨了基于采样技术的轨迹数据流构造问题。

文献[5,6]基于数据自身的分布密度提出了一种偏倚采样算法,该算法通过为每项数据定义一个采样率来反映原始数据的密度分布特性,能够在一定程度上保证采样的质量,但它是针对均匀采样问题的,只适用于静态数据;文献[7]针对数据流的时变特性,利用一个密切关联于时间的偏倚函数来

到稿日期:2013-04-17 返修日期:2013-07-24 本文受辽宁省计划项目基金(2012232001),辽宁省自然科学基金(201202119)资助。
胡志冬(1984—),男,硕士生,主要研究方向为数据挖掘;任永功(1972—),男,博士,教授,硕士生导师,CCF会员,主要研究方向为数据挖掘、图像处理技术等,E-mail:renyg@dl.cn;杨 雪(1987—),女,硕士生,主要研究方向为数据挖掘。

实现非均匀采样，该算法虽然考虑到了最近数据的重要性，需要保证最近的数据采样质量最高，但在一个固定的时间段内，其采样质量往往不一。

本文依据局部聚类思想，结合滑动窗口模型与概率密度聚类算法，提出了一种基于滑动窗口的概率聚类偏倚采样算法。该算法按照窗口的存在密度，对当前滑动窗口进行聚类，将其划分为强簇、弱簇和过度簇，根据各簇的不同性质，为其赋予不同的采样率，从而选取出某些保留了揭示轨迹重要变化的数据，实现对轨迹数据流摘要的线性构造。

2 相关概念

2.1 滑动窗口模型

对于数据流的挖掘处理，研究者们先后提出了许多不同的模型，滑动窗口模型即为其中之一，它基于“最近数据最重要”的数据流特征，在内存中仅保留指定时间段内到达的最近的 N 个数据项。

目前，现存的滑动窗口技术大多采用即刻更新的方法：只要有新数据到达，就立刻更新窗口，使最旧的过期数据项自动丢弃。考虑到数据流与时间密切相关，本文采用基于时间的滑动窗口模型，并约束当前窗口中的所有数据不能同时存储在内存中，且到达的先后顺序无法被记录。

假设当前时间为 T，窗口大小为 W，则包含在 $[T-W,T]$ 时间段内到达的所有数据就构成了当前的滑动窗口，该模型即为滑动窗口模型，模型中的任意时刻，窗口均只包含当前 W 时间段内到达的数据。模型如图 1 所示。

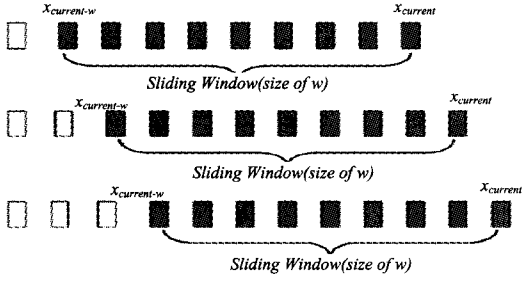

图 1 滑动窗口模型

2.2 数据的存在密度与簇评判标准

将数据流中的一个数据对象看作一个样本点，表示为 (id,p,t,v)，其中 id 为数据索引，p 为数据存在密度，t 为该对象到达窗口的时间，v 为数据对象元素值(维度为 d)，则数据流可看成一个无边界时间序列 $S=\{\langle id_1,p_1,t_1,v_1\rangle,\langle id_2,p_2,t_2,v_2\rangle,\cdots,\langle id_N,p_N,t_N,v_N\rangle\}$。一个数据对象 g_i，假设其到达窗口的时间为 t_i，窗口中包含数据序列 $\{g_1,g_2,\cdots,g_n\}$，各个数据对象的到达时间为 $(t_1,t_2,\cdots,t_n)$，则 g_i 在 t 时刻的存在密度定义为：

$$p_i=\sum_{j=1}^{n}2^{-\lambda(t-t_i)} \tag{1}$$

这里借用函数 $f(t)=2^{-\lambda t}$ 作为每个数据点对应于时间的衰减系数，$\lambda>0$。如果数据对象 g_i 到达窗口的时间为 t_c，则它在 t 时刻的衰减度值取 $2^{-\lambda(t-t_c)}$，当前窗口的存在密度定义为窗口中各数据对象存在概率的几何平均，即

$$CP=(\prod_{j=1}^{n}p_j)^{1/n} \tag{2}$$

设聚类后簇 C 的存在概率为 CP^C，则对于簇 C，如果 $CP^C<\beta$，称其为弱簇；如果 $\beta\leqslant CP^C<\alpha$，称其为过度簇；而如果 $CP^C\geqslant\alpha$，则称为强簇。这里 $\theta<\beta<\alpha<1$，且 $\beta<\min(\sqrt{\alpha\theta},\alpha^2)$，$\theta\in[0,1)$。

2.3 候选簇的选取

假设某一时刻算法维护的簇集合为 $C=\{C_1,C_2,\cdots,C_n\}$，如果此时有一个新的数据对象 g_i 到达，则在为其选择合适候选簇时，需要满足如下规则：

(1)依次扫描簇 $C_1,C_2,\cdots,C_n$，如果在第一次遇到簇 C_j 时，满足：C_j 当前为过度簇，但加入新数据对象 g_i 后变为强簇，则将该簇 C_j 作为候选簇；

(2)如果依次扫描各簇的过程中，第一次碰到的簇 C_j 本身是强簇，加入新数据对象 g_i 后依然是强簇，则将其作为候选簇；

(3)如果扫描过程中，规则(1)与规则(2)均不满足，但新数据对象 g_i 加入 C_j 簇后得到窗口的存在密度增大，则将其加入候选簇。

3 基于滑动窗口模型下密度聚类的偏倚采样——SWMDS

数据摘要的构造方法有很多，其中采样是最常用也是最为简单的方法之一，主要分为均匀采样与偏倚采样。其中均匀采样基于单遍运算，算法简单，易于实现，在数据流摘要构造中获得了广泛的应用。蓄水池采样算法即为均匀采样算法代表之一，其算法描述如下：

输入：数据集或无线数据流 N，蓄水池大小 k；

输出：以 k/H 为概率进行均匀采样所形成的样本集 Buf

```
1. 初始化存储蓄水池数据流项的数据结构 Buf;
2. While 新数据项 e_i
3.   数据流项初始化;
4.   计数器 countOfItem 加 1;
5.   if countOfItem<k//蓄水池未满
6.     将新数据项 e_i 加入 Buf;
7.   else //蓄水池已满
8.     生成随机小数 r
9.     if r<k/countOfItem
10.      计算 p=k×r
11.      用当前数据项替换 Buf[p];
12.    end
13.  end
14. end
```

通过算法分析，我们可以发现，这种基于均匀采样的构造方法极易丢失某些关键数据，尤其是对密度变化较大的数据集，这类算法可能就会丢失那些相对稀疏但又很重要的数据，同时有可能引入那些不太重要的噪声数据。文献[4]基于数据分布密度提出了一种偏倚采样算法，该算法通过为每项数据定义不同的采样率，以使样本能够真实地反映出原始数据集的分布特性，从而克服均匀采样丢失重要信息的弊端。本文算法即是在此思想基础上进行设计。

3.1 SWMDS 算法思想

文献[8]中，B. Yingyi 依据轨迹数据流局部连续的特性，提出了一种局部聚类策略，在数据流异常检测方面获得了不

错的效果。受这一思想启发，本文在采样前加入一个类似于局部聚类的数据预处理过程，在该过程中，数据流将按其存在密度划分为长度不等的基本窗口，窗口的大小反映了数据流中相似元素的多少；最后根据各窗口存在密度的大小来设置不同的采样率对数据进行采样，合并选取样本构成最终的数据摘要。

设窗口间存在密度差的阈值为 τ，时间约束的上限值为 m_b，数据流 S 中的一个数据片段定义为序列 $C=\{s_i, s_{i+1}, \cdots, s_j\}$（其中 $j-i+1\leqslant m_b$），令 B_p, B_q 为 C 中以 p, q 为结尾的数据片段，如果 $\exists s_p \in C, \forall s_q \in C$，有 $|CP_{B_p}-CP_{B_q}|<\tau$，则称 C 为一个簇，将满足 $\forall s_i \in C, |CP_{B_q}-CP_{B_i}|<\tau$ 的数据点 s_q 定义为 C 簇的中心点，$\mathrm{Max}_{s\in C}|CP_{B_q}-CP_{B_s}|$ 定义为簇 C 的半径，将到达时间大（小）于 q 的数据序列定义为 s_q 的右（左）窗口。

如果滑动窗口预选取的样本总数为 M，当前存在的基本窗口数量为 k，第 i 个窗口的大小为 B_i，则定义该窗口的采样率为：

$$\delta=\begin{cases}\dfrac{M}{\sum_{i=1}^{k}B_i}\times\mu_1', & \text{if 该窗口属于强簇}\\ \dfrac{M}{\sum_{i=1}^{k}B_i}\times\mu_2, & \text{if 该窗口属于过度簇}\\ \dfrac{M}{\sum_{i=1}^{k}B_i}\times\mu_3, & \text{if 该窗口属于弱簇}\end{cases} \tag{3}$$

其中，μ_1, μ_2, μ_3 分别表示对 3 类窗口的权值。

3.2 SWMDS 算法描述

开始执行算法时，需要由用户指定初始变量，当前窗口数量可以随着算法的运行不断变化。

算法 SWMDS(b, ω, τ, M)

输入：聚类窗口大小 b，滑动窗口大小 ω，存在密度阈值 τ，预选取样本总数 M

输出：数据流摘要样本

1. 初始化数据结构 Buf，用于当前聚类窗口不满时的数据流项缓存；
2. 初始化数据结构 SWM，以便存储滑动窗口内的所有采样摘要；
   设聚类中心点 Centre=NULL，聚类半径 R=0；
3. while 新数据流项 s_i
4. Count_Item++；//记录数据流项数
5. if Count_Sam_Win<ω//滑动窗口不满
6. if Count_Item<b//聚类窗口不满
7. 将 s_i 放入缓存 Buf；
8. else //聚类窗口已满
   if Centre=NULL//未发现聚类中心
9. 计算 $r_i=|CP_{B_i}-CP_{B_{i-1}}|$
10. if $r_i>\tau$
11. Centre=newCentre(t_i)；//更新类中心点
12. if $r_i<\tau$ and $r_i>R$
13. R=r_i；//更新聚类半径
14. else//已有类中心点
15. 计算 $r_i=|CP_{B_i}-CP_{B_{i-1}}|$；
16. if $r_i<\tau$ and $r_i>R$
17. R=r_i；//更新聚类半径
18. else
19. B_i=Buf；//构造完成当前基本窗口 B_i
20. 由采样率 δ 与 B_i 计算当前窗口的采样数 $num_i=B_i*\delta$；
21. Sample(B_i, num_i)；//对聚类窗口内数据进行采样
22. 计数器 Count_Item--；
23. 将采样摘要数据存储至 SWM；
24. Count_Sam_Win++；
25. 清空 Buf；
26. else//滑动窗口已满
27. 丢弃到达时间最早的数据
28. Count_Sam_Win--；
29. 更新 SWM；

4 实验分析

我们基于 C++ 编程实现了本文所提出的 SWMDS 算法，采用传统均匀采样算法中的蓄水池采样算法（Resample）进行实验对比，通过对一组合成数据与一组真实数据的采样实现来测试该算法的有效性。实验评价指标主要为：采样质量与运算时间。

4.1 采样质量指标

借用重构法思想，对算法采取的最终数据样本使用线性插值的方法进行数据流的重构，将重构数据流 $\hat{S}$ 与原始数据流 S 每一时刻取值的均方差 $L(\hat{S},S)$ 作为采样质量评估指标。$L(\hat{S},S)$ 的计算公式如下：

$$L(\hat{S},S)=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\|\hat{e}_i-e_i\|_2^2} \tag{4}$$

其中，$\|\hat{e}_i-e_i\|_2^2=\sqrt{\sum_{j=1}^{d}(\hat{e}_i^j-e_i^j)}$（$d$ 为数据维度）。

4.2 结果分析

对于合成数据，我们采用随机生成函数产生一个含有 10000 组的二维伪轨迹数据集，实验中，将其滑动窗口 ω 大小设置为 512，结果如图 2 与图 3 所示。真实数据则使用一组真实的 GPS 导航轨迹数据集，这一数据集记录了某人从家到办公室连续 12 天开车走过的真实路线，含有 30000 组二维轨迹数据，将其滑动窗口 ω 大小设置为 1024，实验结果如图 4、图 5 所示。

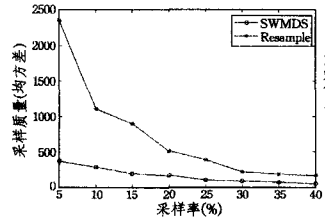

图 2 合成数据不同采样率下的采样质量

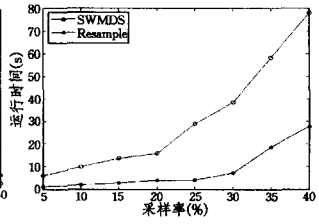

图 3 合成数据不同采样率下的运行时间

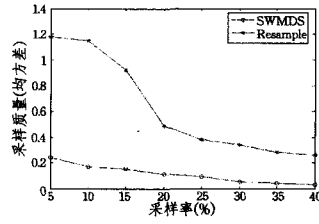

图 4 真实数据不同采样率下的采样质量

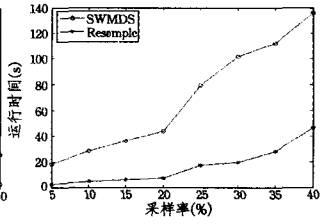

图 5 真实数据不同采样率下的运行时间

（下转第 269 页）

检索系统。本方法主要通过图像区域中的视觉属性获得词袋，通过子块集合聚类得到一个最接近图像区域的区域类型，不同于其它采用整个视觉特征来获取词袋的方法。这样更能够简单有效、准确地表示图像特征，这样使得图像特征更加靠近视觉词汇的概念。本文提出的方法在语义图像检索的实验中取得了良好的效果。实验证明，我们提出的方法优于普通的基于局部或全局描述子构造的词袋模型。

参 考 文 献

[1] Hiremath P S, Pujari J. Content based image retrieval using color, texture and shape features[C]// Advanced Computing and Communication. Gulbarga: Gulbarga University. 2007

[2] 周明全，耿国华，韦娜. 基于内容图像检索技术[M]. 北京：清华大学出版社，2007

[3] Duygulu P, Barnard K, De Freitas J F G, et al. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary[J]. Lecture Notes in Computer science, 2002(2353): 97-112

[4] Spyrou E, Tolias G, Mylonas P, et al. Concept detection and keyframe extraction using a visual thesaurus[J]. Multimedia Tools and Applications, 2009, 41(3): 337-373

[5] Manjunath B, Ohm J, Vasudevan V, et al. Color and texture descriptors[J]. IEEE Trans Circuits Syst Video Technol, 11(6): 703-715

[6] Spyrou E, Tolias G, Mylonas P, et al. Concept detection and keyframe extraction using a visual thesaurus[J]. Multimedia Tools and Applicationgs, 2009(41): 337-373

[7] Avrithis Y, Doulamis A, Kollias S. A Stochastic Framework for Optimal Key Frame Extraction from MPEG Video Databases [J]. Computer Vision and Image Understanding, 1999(75): 3-24

[8] Mylonas P, Spyrou E, Avrithis Y, et al. Using Visual Context and Region Semantics for High-Level Concept Detection[J]. IEEE Transaxtions on Multimedia, 2009, 11(2): 229

[9] 曹利华，柳伟，李国辉. 基于多种主色调的图像检索算法研究与实现[J]. 计算机研究与发展，1999(36)：96-100

[10] 张瑜慧. 基于 SVM 的语义图像检索技术的研究与实现[D]. 扬州：扬州大学，2007

[11] Chang E, Kingshy G, Sychay, et al. CBSA: content-based soft annotation for multimodal image retrieval using Bayes point machines[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2003(13): 26-38

[12] Lowe D G. Object recognition from local scale-invariant features [J]. Computer Vision, 1999(2): 1150-1157

[13] Wang Jun-qiu. Vision-based Global Localization Using a Visual Vocabulary[C]// Robotics and Automation. Beijing: Peking University, 2005

(上接第 256 页)

从实验结果图我们可以看出，随着样本采样率的增大，采样质量逐渐提高，运行时间也随之增大。通过两种算法实验结果的对比，我们能够明显看出本文所提算法 SWMDS 的采样质量远远高于传统的 Resample 算法，并且当采样率为 5%～15%时，SWMDS 算法的采样质量几乎是 Resample 算法的 5 倍。在运行时间上，由于本文算法 SWMDS 采样前加入了聚类处理，因此相比于传统 Resample 算法要花费更多的时间，但依然保持与采样率之间的线性关系。

结束语 本文针对传统均匀采样在轨迹数据流摘要构造过程中易丢失关键信息的问题，提出一种基于概率密度聚类的数据流偏倚采样算法。该算法在滑动窗口模型下，结合了偏倚采样算法思想，首先基于数据存在密度进行聚类分析，将滑动窗口划分为强簇、弱簇和过度簇，然后针对不同的簇给予不同的采样率，依据各自的采样率对窗口内数据进行偏倚采样，进而构造出更为完善的数据流摘要。算法充分考虑了轨迹数据流自身的分布特性，能够在较低的采样率下获得较高的采样质量。

参 考 文 献

[1] Kun-Ta C, Hung-Leng C, Ming- Syan C. Feature-preserved sampling over streaming data[J]. ACM trans. Knowl. Discov. Data, 2009, 2(4): 1-45

[2] 张春阳，周继恩，钱权，等. 抽样在数据挖掘中的应用研究[J]. 计算机科学，2004，31(2)：126-128

[3] Dimitris S, Antonios D, Timos S. Hierachically compressed wavelet synopses[J]. The VLDB Journal, 2009, 18(1): 203-231

[4] 余波，朱东华，刘嵩，等. 密度偏差抽样技术在聚类算法中的应用研究[J]. 计算机科学，2009，36(2)：207-209

[5] 戴东波，赵杠，孙圣力. 基于概率数据流的有效聚类算法[J]. 软件学报，2009，20(5)：1313-1328

[6] 常建龙，曹锋，周傲英. 基于滑动窗口的进化数据流聚类[J]. 软件学报，2007，18(4)：905-918

[7] 程转流，胡为成. 滑动窗口模型下的概率数据流聚类[J]. 计算机工程与应用，2011，47(4)：141-145

[8] B Ying-yi, C Lei, Wai-Chee F A, et al. Efficient anomaly monitoring over moving object trajectory streams[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, ACM, 2009