软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(7):1655−1670 [doi: 10.13328/j.cnki.jos.005033]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种多源感知数据流上的连续真值发现技术*

李天义, 谷 峪, 马 茜, 李芳芳, 于 戈

(东北大学 计算机科学与工程学院,辽宁 沈阳 110819)

通信作者: 谷峪, E-mail: guyu@mail.neu.edu.cn

摘 要: 真值发现作为整合由不同数据源提供的冲突信息的一种手段,在传统数据库领域已经得到了广泛的研究.然而现有的很多真值发现方法不适用于数据流应用,主要原因是它们都包含迭代的过程.针对一种特殊的数据流——感知数据流上的连续真值发现问题进行了研究.结合感知数据本身及其应用特点,提出一种变频评估数据源可信度的策略,减少了迭代过程的执行,提高了每一时刻多源感知数据流真值发现的效率.首先定义并研究了当感知数据流真值发现的相对误差和累积误差较小时,相邻时刻数据源的可信度变化需要满足的条件,进而给出了一种概率模型,以预测数据源的可信度满足该条件的概率.之后,通过整合上述结论,实现在预测的累积误差以一定概率不超过给定阈值的前提下,最大化数据源可信度的评估周期以提高效率,并将该问题转化为一个最优化问题.在此基础上,提出了一种变频评估数据源可信度的算法——CTF-Stream(continuous truth finding over sensor data streams), CTF-Stream 结合历史数据动态地确定数据源可信度的评估时刻,在保证真值发现结果达到用户给定精度的同时提高了效率.最后,通过在真实的感知数据集合上进行实验,进一步验证了算法在处理感知数据流的真值发现问题时的效率和准确率.

关键词: 多源;数据流;感知数据;真值发现;数据源可信度

中图法分类号: TP311

中文引用格式: 李天义,谷峪,马茜,李芳芳,于戈.一种多源感知数据流上的连续真值发现技术.软件学报,2016,27(7):1655−1670. http://www.jos.org.cn/1000-9825/5033.htm

英文引用格式: Li TY, Gu Y, Ma Q, Li FF, Yu G. Technique for continuous truth discovery over multiple-source sensor data streams. Ruan Jian Xue Bao/Journal of Software, 2016,27(7):1655−1670 (in Chinese). http://www.jos.org.cn/1000-9825/5033.htm

Technique for Continuous Truth Discovery Over Multiple-Source Sensor Data Streams

LI Tian-Yi, GU Yu, MA Qian, LI Fang-Fang, YU Ge

(School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China)

Abstract: As a method of assessing validity of conflicting information provided by various data sources, truth discovery has been widely researched in the conventional database community. However, most of the existing solutions of truth discovery are not suitable for applications involving data streams, mainly because their methods include iterative processes. This paper studies the problem of continuous truth discovery in a special kind of data streams-sensor data streams. Combining with the characteristics of sensor data itself and its application, a strategy is proposed based on changing the frequency of assessing source reliability to reduce the iterative processes, and therefore to improve the efficiency of truth discovery in multiple-source sensor data streams. First, definitions are provided on when the relative errors and accumulative errors are relatively small, and the necessary conditions of the variation on source reliability from

* 基金项目: 国家重点基础研究发展计划(973)(2012CB316201); 国家自然科学基金(61433008, 61472071, 61272179); 中央高校基本科研业务费(N140404013)

Foundation item: National Key Basic Research Program of China (973) (2012CB316201); National Natural Science Foundation of China (61433008, 61472071, 61272179); Fundamental Research Funds for Central Universities (N140404013)

收稿时间: 2015-09-25; 修改时间: 2016-01-12; 采用时间: 2016-02-22; jos 在线出版时间: 2016-03-22

CNKI 网络优先出版: 2016-03-22 13:23:24, http://www.cnki.net/kcms/detail/11.2560.TP.20160322.1323.002.html

© 中国科学院软件研究所 http://www.jos.org.cn

adjacent time points. Next, a probabilistic model is given to predict the probability of meeting these necessary conditions. Then, by integrating the above conclusions, maximal assessing period of source reliability is achieved, under the condition that the cumulative error of prediction is smaller than the given threshold in a certain confidence level of probabilities, in order to improve efficiency. Thus the truth discovery problem is transformed into an optimization problem. Furthermore, an algorithm, CTF-Stream (continuous truth finding over sensor data streams) is constructed to assessing source reliability with changeable frequencies. CTF-Stream utilizes the historic data to dynamically determine the time needed to assess the source reliability, and finds the truth with a certain accuracy given by customers while improving the efficiency. Finally, both efficiency and accuracy of the presented methods for truth discovery in sensor data streams are validated by conducting the extensive experiments on real sensor dataset.

Key words: multiple-source; data stream; sensor data; truth discovery; source reliability

当多数据源对同一实体具有不同描述时,通过整合这些描述信息获取该实体的实际描述,这个过程被称为真值发现(truth discovery).近年来,真值发现在传统数据库领域得到了广泛研究[1−13].然而,随着移动计算和在线应用的快速发展,数据流的应用模型已出现在众多领域,例如金融应用、网络监视、通信数据管理、传感器网络数据处理等.在数据流应用中,不仅数据是时刻变化的,数据的真值也随着时间不断演化,即每一时刻都需要对新到达的数据进行真值发现.但是,由于数据流规模宏大且需要实时处理[14,15],因此在数据流上进行真值发现时,其核心问题是尽可能地提高算法的效率,确保能够连续、高效地获取流数据的真值.而传统的真值发现方法都是基于迭代的策略,其迭代过程具有较高的时间复杂度;并且,在对流数据进行真值发现时,基于迭代的方法需要每一时刻都遍历从初始时刻到当前时刻的一些历史信息.这两点都使得基于传统数据库的真值发现方法无法被简单地移植到数据流应用中,即难以满足数据流连续真值发现的需要[16].综上,如何快速、高效地处理数据流上的真值发现依然是亟待解决的问题.

本文针对一种特殊的数据流——感知数据流上的连续真值发现问题进行研究.随着无线通信技术、微电子技术及嵌入式计算技术的快速发展,无线传感器网络技术已经成为研究的热点.传感器网络由分布于特定区域的很多传感器节点构成,每个节点均具有一定的计算、存储和通信能力,用户通过向无线传感器网络发布查询获取被监测区域的信息以满足各种应用需求,大量的感知数据随之产生.虽然人们已经提出了很多方法以面向不同的感知数据应用[17],但是它们都忽略了感知数据流上的真值发现问题.

以图 1 为例,在一个较小的室内空间布有若干个传感器(黑色圆点),每一个传感器都能采集到一系列随着时间变化的感知数据.在实际应用中,若将这个较小的室内空间看作是一个整体,则可以认为每一个传感器探测的都是该室内空间的物理信息,即每一个传感器在每一时刻都会采集同一监测对象(该室内空间)的多维属性(温度、湿度等).因此,需要在每一时刻都整合由多个传感器返回的冲突数据,获取该室内空间相应物理量的最为准确的信息.上述问题即是一个感知数据流的连续真值发现问题.同理,在一片划分了网格的海域中,整合每一时刻每一网格获取的水文信息(水位、流速等)也是一个感知数据流的连续真值发现应用.此外,由上述例子可知,每一个感知数据源往往会从多个维度对同一监测对象进行描述,即对同一实体探测得到的感知数据,通常包含该实体的多维属性.因此,在对感知数据流进行真值发现时,不仅要考虑数据源的多样性,也要考虑属性的多样性,即需要在每一时刻同时整合由多数据源提供的关于多维属性的观测值,获取每一维属性的真值.

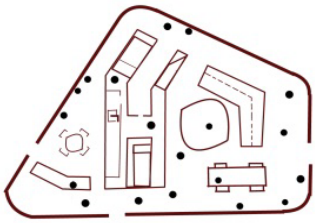


Fig.1 Distribution of sensors in an indoor space

图 1 室内空间中传感器分布

感知数据流不仅包含了数据流的连续到达、规模大等特性,还具有感知数据的几个重要特点:(1) 具有时空

© 中国科学院软件研究所 http://www.jos.org.cn

相关性[18,19],即相邻时刻数据的“跃动”较小;(2) 主要以连续型数据为主,即使是如光照强度这种需要基于照度标准值进行分类处理的数据,最初也是以连续型数据的形式进行采集[20].此外,从感知数据的实际应用角度来看,如车间光照度监测、水质监测等,用户并不需要获得一个精确值,而仅需要一个近似值,判断其是否符合相应的监测标准或属于监测标准中的哪一级别[19].虽然依据这种思想,可以将一些基于传统数据库的真值发现方法应用到感知数据流的连续真值发现问题中,但是它们的方法都是基于迭代的策略,无法实时地处理连续到达的感知数据流[16],即会产生数据过载的现象.同时,结合感知数据流在相邻时刻“跃动”较小的特点,每一时刻都更新数据源的可信度也是没有必要的.

依据上述感知数据流自身及应用方面的特性,本文提出了一个基于累积误差预测的数据源可信度更新策略,处理多源多属性感知数据流的连续真值发现问题.首先从感知数据流在相邻时刻“跃动”较小的特点出发,研究数据源可信度在相邻时刻的变化对“相对误差”和“累积误差”的影响,即给出当一段时间内每相邻时刻数据源可信度的变化满足一定条件时,累积误差的最大值;其次,本文给出预测数据源可信度在相邻时刻的变化是否满足该条件的概率模型;之后,将累积误差的预测问题转化为一个最优化问题,并在此基础上给出一种平衡准确率和效率的算法——CTF-Stream(continuous truth finding over sensor data streams),变频更新数据源的可信度.CTF-Stream 在一定概率保证和误差允许范围内利用某一时刻的数据源的可信度代替一段时间内数据源的可信度,在减少评估数据源可信度的次数的同时省去了迭代过程的执行,实现了感知数据流的高效实时处理;最后,通过实验验证了 CTF-Stream 在保证运行结果以一定概率满足用户给定精度的前提下,可以提高感知数据流真值发现的效率.

综上,本文的主要贡献有:

- 研究多源多属性感知数据流上的连续真值发现问题,结合感知数据及其应用特点,定义并研究了感知数据流真值发现的相对误差和累积误差,并分别给出当相对误差和累积误差较小时,相邻时刻数据源可信度变化应满足的条件.
- 设计了一种基于 Bernoulli 分布的概率预测模型,预测一段时间内感知数据源可信度的变化满足上述条件的概率.
- 整合上述结论,将感知数据流的数据源可信度的更新问题转化为最优化问题,在以一定概率限制了累积误差上界的前提下,最小化数据源可信度的评估周期以提高效率.
- 在此基础上,提出了一种变频评估数据源可信度的算法——CTF-Stream,处理感知数据流的连续真值发现问题.CTF-Stream 结合历史数据,动态地确定数据源可信度的更新时间,同时保证真值发现的结果满足用户给定的精度.
- 在真实数据集上通过大量实验验证了本文算法的准确性及有效性.

本文第 1 节讨论相关工作.第 2 节给出问题定义.第 3 节分析数据源可信度在相邻时刻的变化对相对误差和累积误差的影响,并给出一种概率模型.第 4 节详细介绍本文提出的数据源可信度更新策略,并给出 CTF-Stream 算法.第 5 节给出实验结果及分析.最后第 6 节对本文进行总结.

1 相关工作

目前的真值发现研究仍主要集中于传统数据库领域.文献[1]率先提出真值发现这一概念,针对 Web 环境下多数据源提供的数据存在冲突这一问题,就客观实体的单一属性进行了真值发现研究,提出了 TruthFinder 算法.文献[2]提出了一种评估数据源可信度的模型.该模型主要包括 3 种算法:Consine,2-Estimates 和 3-Estimates,可同时估计数据源可信度和真值.文献[3]是第 1 个针对连续型数据的真值发现算法.文献[4]首次综合考虑数据源的灵敏度和特指度以评估数据源的可信度,文献[4]虽然提出了一种增量方法,但在实际处理数据流的真值发现问题时并不可行[16],并且所提方法面向的都是离散型数据,不适合感知数据应用.文献[5]针对异构数据的真值发现问题提出了 CRH 算法,该方法所包含的迭代过程具有较高的收敛率和准确率.文献[6]在文献[5]的基础上进一步考虑数据源的“长尾效应”,结合区间估计的思想,提出了 CATD 算法.上述方法认为数据源之间相互独立.

© 中国科学院软件研究所 http://www.jos.org.cn

对于数据源不独立的情况,文献[7]提出在对数据进行真值发现时,有必要考虑数据源存在复制的情形,并给出了一种基于 Bayesian 分析的方法,判定数据源之间的复制关系.文献[8]在文献[7]的基础上,考虑数据源复制状态的转移问题,采用 Hidden Markov 模型推测某一时刻各数据源的复制状态,从数据的新鲜度、覆盖率和精确度这 3 个方面评估数据源的可信度.文献[9]进一步考虑了数据的复制关系,给出了一种全局复制检测算法,该算法能够识别多数据源同步复制、多数据源传递复制等情况.文献[10]给出了一个判定数据源复制关系的原型演示系统.文献[11]以上述工作为基础,提出了一个多层概率模型,该模型区分了错误是由于信息本身还是由网页提取器所造成,进一步提高了 Web 数据的可用性.文献[12]通过定义联合准确率和联合召回率进一步研究了数据源之间正相关非复制和负相关的情形.文献[13]通过实验研究了部分基于 Web 的真值发现方法.上述考虑数据源之间关系的真值发现研究面向的都是离散型数据.

在数据流应用方面,文献[16]中的 StreamTF 算法虽然可以处理数据流上的真值发现问题,但是同样面向离散型数据,而本文提出的真值发现方法针对的是连续型数据,因此二者存在一定的区别.文献[21]同样研究了数据流上的真值发现,将真值发现问题中常见的最优化模型近似转化为概率模型,增量地学习数据源的可信度,然而其估计的数据源可信度需要经过一段时间才能收敛到对应的最优解,即在提高了效率的同时牺牲了准确率;而本文提出的方法在着重考虑感知数据特性的同时,确保在更新点处获得的数据源可信度的值为其对应的最优解,同时可以通过调节参数的设置,调整真值发现结果的准确率和效率,即本文与文献[21]中的工作存在一定的区别.

2 问题定义

本文的研究内容为:在多源多属性感知数据流连续到达的过程中,如何动态地最大化数据源可信度的评估周期,即减少迭代过程的执行,提高感知数据流真值发现的效率,同时还要保证真值发现的结果满足用户给定的精度.

2.1 问题描述

如图 2 所示是本文的研究框架.首先,本文依据感知数据在相邻时刻"跃动"较小的特点,考虑相邻时刻数据源可信度的变化情况,进而定义并研究了感知数据流真值发现的相对误差和累积误差,这两种误差都是由于利用历史时刻的数据源的可信度,代替当前时刻数据源的可信度进行真值发现所造成的,我们研究并证明了这两种误差较小时(不超过给定阈值ε或关于ε的函数时),相邻时刻数据源的可信度的变化应满足的条件,并给出了预测这一条件的概率模型,这样预测一段时间内的累积误差是否不超过给定的阈值,实质上就等价于预测该段时间内的数据源可信度在相邻时刻的变化是否满足相应的条件;其次,整合上述结论,将累积误差的预测问题转化为一个最优化问题,即已知前一评估数据源的时刻,最大化它与下一评估时刻之间的时间间隔ΔT,同时限制ΔT 内的累积误差;最后,给出了一种变频评估数据源可信度的策略,结合历史数据,动态地调整数据源可信度的评估周期.在不更新数据源可信度的时刻,利用具有线性复杂度的方法对感知数据进行真值发现.因此本文实现了在使真值发现结果满足用户给定的精度的同时,减少了迭代过程,提高了感知数据流真值发现的效率.

定义 1(数据源权值). t_i时刻第 k 个数据源的可信度为该数据源在 t_i时刻的权值,记为 w_i^k, 则 t_i时刻数据源的权值集合为 $W_i=\left\{w_i^1,w_i^2,...,w_i^K\right\}$.

定义 2(观测值). t_i时刻第 k 个数据源提供的第 m 维属性的值为该数据源在 t_i时刻对第 m 维属性的观测值,记为 $v_i^{(k,m)}$.

定义 3(真值). t_i时刻,对第 m 维属性进行真值发现获得的整合结果为真值,记为 $v_i^{(*,m)}$, 则 t_i时刻真值集合为 $V_i^{(*)}=\left\{v_i^{(*,1)},v_i^{(*,2)},...,v_i^{(*,M)}\right\}$.

定义 4(更新点). 若 t_i时刻计算了数据源的权值,则 t_i时刻为一个更新点.

© 中国科学院软件研究所 http://www.jos.org.cn

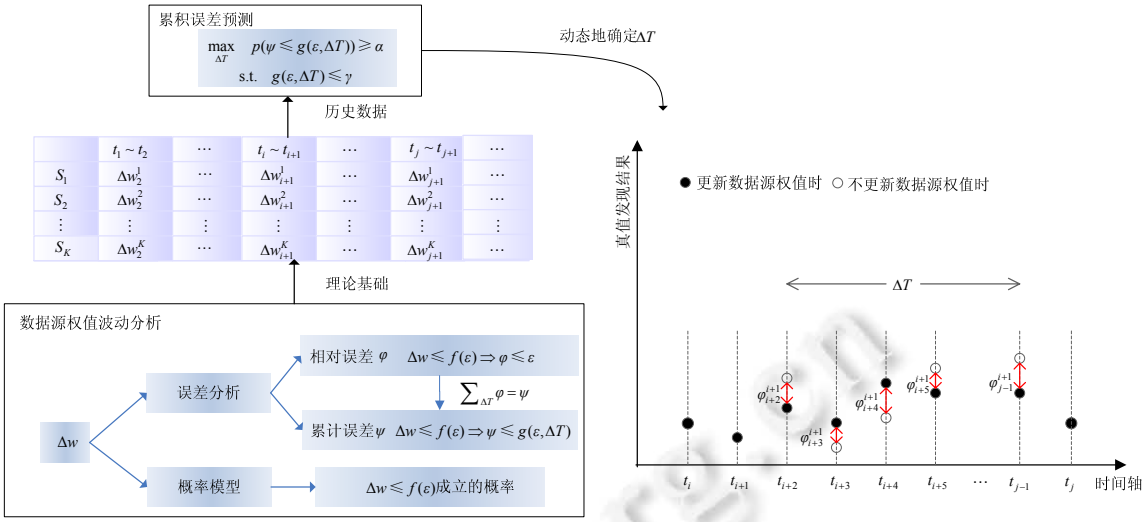


Fig.2 The research framework

图 2 研究框架

2.2 CRH算法

由第2.1节可知,本文研究的是如何在数据流上动态地最大化评估数据源可信度的周期.因此在更新点(图2所示中的 t_i、t_{i+1} 和 t_j)仍然需要评估数据源的可信度.本文利用 CRH(conflict resolution on heterogeneous data)[5] 在更新点处联合推导数据源的权值和真值.

根据 CRH 算法,在 t_i 时刻数据源权值和真值的推导框架如下所示:

$$\begin{aligned}&\mathrm{Min}\ f\left(V_i^{(*)},W_i\right)=\sum_{k=1}^{K}w_i^k\sum_{m=1}^{M}d\left(v_i^{(*,m)},v_i^{(k,m)}\right)\\&\text{s.t.}\quad \delta(W)=1,\ W\in S\end{aligned}\tag{1}$$

其中,$d\left(v_i^{(*,m)},v_i^{(k,m)}\right)$ 为损失函数,用以衡量属性的真值 $v_i^{(*,m)}$ 和属性的观测值 $v_i^{(k,m)}$ 之间的偏差;$\delta(W)$ 为数据源权值的标准化函数.本文采用平方损失函数和对数标准化函数[5]:

$$d\left(v_i^{(*,m)},v_i^{(k,m)}\right)=\frac{\left(v_i^{(*,m)}-v_i^{(k,m)}\right)^2}{\mathrm{std}\left(v_i^{(1,m)},\ldots,v_i^{(K,m)}\right)}\tag{2}$$

$$\delta(W_i)=\sum_{k=1}^{K}\exp\left(-w_i^k\right)\tag{3}$$

其中,$\mathrm{std}\left(v_i^{(1,m)},\ldots,v_i^{(K,m)}\right)$ 为 $v_i^{(1,m)},\ldots,v_i^{(K,m)}$ 的标准差.由于本文是在流数据连续到达的每一时刻,同时对多源感知数据的多维属性进行真值发现,因此需要引入每一维属性的标准差,对多维属性的平方损失函数进行标准化.

CRH 算法采用交替迭代的策略,即首先固定 $V_i^{(*)}$,更新 W_i,使式(1)达到最小,之后再固定 W_i,更新 $V_i^{(*)}$,如此反复,直至 $V_i^{(*)}$ 收敛到某个值,迭代结束,该值即为真值.

当 W_i 固定时,直接利用式(4)更新 $V_i^{(*)}$,可使式(1)达到最小[5].

$$v_i^{(*,m)}=\frac{\sum_{k=1}^{K}w_i^k\cdot v_i^{(k,m)}}{\sum_{k=1}^{K}w_i^k}\tag{4}$$

反之,当 $V_i^{(*)}$ 固定时,直接利用式(5)更新 W_i,可使式(1)达到最小[5].

$$w_i^k=-\log\left(\frac{\sum_{m=1}^{M}d\left(v_i^{(*,m)},v_i^{(k,m)}\right)}{\sum_{k'=1}^{K}\sum_{m=1}^{M}d\left(v_i^{(*,m)},v_i^{(k',m)}\right)}\right)\tag{5}$$

© 中国科学院软件研究所 http://www.jos.org.cn

本文采用 CRH 算法在更新点处对感知数据流进行真值发现的原因主要有以下两点:(1) 可以处理连续型数据;(2) 迭代过程收敛较快且算法的准确率较高[5].

定义 5(近似值). 利用 t_i 时刻数据源的权值 W_i 代替 t_j 时刻数据源的权值 W_j,根据式(4)得到的第 m 维属性的值为 t_j 时刻的近似值,记为 $v_{j/i}^{(*,m)}(i<j)$.

定义 6(数据源的权值波动). 在 t_{i-1} 和 t_i 两个相邻时刻,第 k 个数据源占所有数据源权值之和的比重的差异为第 k 个数据源在 t_i 时刻的权值波动,记为 Δw_i^k.

Δw_i^k 的表达式如下:

$$\Delta w_i^k=\left|w_i^k\Big/\sum_{k=1}^K w_i^k-w_{i-1}^k\Big/\sum_{k=1}^K w_{i-1}^k\right| \tag{6}$$

下面,本文基于式(4)和式(6),定义并研究感知数据流真值发现的相对误差和累积误差,同时证明当它们较小时,数据源权值波动应满足的条件.

3 基于数据源权值波动的误差分析

在真值发现问题中,数据源权值的评估至关重要[5],结合式(4)可知某一数据源的权值占所有数据源权值之和的比重反映了该数据源提供的观测值对真值计算的贡献程度.对于动态到达的数据流,结合式(6)可知所有数据源在相邻时刻的这一差异都较小时,显然利用前一时刻的数据源的权值代替当前数据源的权值进行真值发现会使误差很小,且省去了利用 CRH 算法迭代计算的过程.本文从这个角度出发,分别定义相对误差和累积误差,并分析当这些误差不超过某一阈值时,数据源的权值波动应满足的条件.

注意,由于我们的算法在更新点处利用 CRH 算法联合推导真值和数据源的权值,因此下文提到的真值都是在每一时刻利用 CRH 算法进行真值发现获得的整合结果.

3.1 数据源的权值波动对相对误差的影响

定义 7(相对误差). 用 $v_{j/i}^{(*,m)}$ 估计 $v_j^{(*,m)}$ 时产生的平方损失,为 t_i~t_j 时刻的相对误差,记为 $\phi_j^{(i,m)}(i<j)$. 当 $i=j-1$ 时,$\phi_j^{(i,m)}(1\leqslant m\leqslant M)$ 为相邻时刻的相对误差,简记为 φ^m.

$\phi_j^{(i,m)}$ 的表达式如下:

$$\phi_j^{(i,m)}=\left(\frac{v_j^{(*,m)}-v_{j/i}^{(*,m)}}{v_j^{(\max,m)}}\right)^2 \tag{7}$$

其中,$v_j^{(\max,m)}$ 为 t_j 时刻所有数据源提供的第 m 属性上绝对值最大的观测值,即 $\left|v_j^{(\max,m)}\right|\geqslant\left|v_i^{(k,m)}\right|(1\leqslant k\leqslant K)$,类似于对式(2)的处理,本文利用 $v_j^{(\max,m)}$ 对多维属性的相对误差进行标准化.

下面证明为使 $\varphi^m(1\leqslant m\leqslant M)$(下文简记为 φ)不超过给定阈值 ε,数据源权值波动应满足的条件.

定理 1. 当 $\Delta w_i^k\leqslant\varepsilon^{1/2}/K(1\leqslant k\leqslant K)$ 时,$\varphi\leqslant\varepsilon$,其中,$K$ 为数据源集合大小.

证明:由式(6)得,

$$\Delta w_i^k=\left|w_i^k\Big/\sum_{k=1}^K w_i^k-w_{i-1}^k\Big/\sum_{k=1}^K w_{i-1}^k\right|,$$

由式(7)得,

$$\left|v_i^{(*,m)}-v_{i/i-1}^{(*,m)}\right|\leqslant\varepsilon^{1/2}\cdot\left|v_j^{(\max,m)}\right|\Leftrightarrow\phi\leqslant\varepsilon.$$

代入式(4)到式(7),

$$\left|\sum_{k=1}^K\left(\left(w_i^k\Big/\sum_{k=1}^K w_i^k-w_{i-1}^k\Big/\sum_{k=1}^K w_{i-1}^k\right)\cdot v_i^{(k,m)}\right)\right|\leqslant\varepsilon^{1/2}\cdot\left|v_j^{(\max,m)}\right|\Leftrightarrow\phi\leqslant\varepsilon.$$

又因为有

© 中国科学院软件研究所 http://www.jos.org.cn

$$\left|\sum_{k=1}^{K}\left(\left(w_i^k/\sum_{k=1}^{K}w_i^k-w_{i-1}^k/\sum_{k=1}^{K}w_{i-1}^k\right)\cdot v_i^{(k,m)}\right)\right|\leqslant\sum_{k=1}^{K}\left|\left(w_i^k/\sum_{k=1}^{K}w_i^k-w_{i-1}^k/\sum_{k=1}^{K}w_{i-1}^k\right)\cdot v_i^{(k,m)}\right|,$$

且

$$\left|v_j^{(\max,m)}\right|\geqslant\left|v_i^{(k,m)}\right|(1\leqslant k\leqslant K),$$

则若

$$\sum_{k=1}^{K}\left(\left|w_i^k/\sum_{k=1}^{K}w_i^k-w_{i-1}^k/\sum_{k=1}^{K}w_{i-1}^k\right|\cdot\left|v_j^{(\max,m)}\right|\right)\leqslant\varepsilon^{1/2}\cdot\left|v_j^{(\max,m)}\right|,$$

有$\varphi\leqslant\varepsilon$,则当$\left|w_i^k/\sum_{k=1}^{K}w_i^k-w_{i-1}^k/\sum_{k=1}^{K}w_{i-1}^k\right|\leqslant\varepsilon^{1/2}/K$成立,$\varphi\leqslant\varepsilon$一定成立. □

由此可知,若所有数据源在 t_i 时刻的权值波动 Δw_i^k 均满足:

$$\Delta w_i^k\leqslant\varepsilon^{1/2}/K(1\leqslant k\leqslant K)\tag{8}$$

则不更新 t_i 时刻数据源的权值,直接利用 t_{i-1} 时刻数据源的权值 W_{i-1},根据式(4)计算 t_i 时刻的真值,所产生的相对误差φ一定不会超过给定阈值ε.

然而,实际上,在 t_{i-1} 时刻,t_i 时刻数据源的权值是未知的,因此也无法确定 $\Delta w_i^k(1\leqslant k\leqslant K)$ 是否满足式(8).对此,本文提出一种基于概率模型的预测方法,动态地预测数据源的权值波动满足式(8)的概率.详细内容将在第3.3节中加以讨论.

3.2 数据源的权值波动对累积误差的影响

定义 8(累积误差). t_i~$t_h(i<h\leqslant j)$时刻第 m 维属性的相对误差和,为 t_i~t_j 时刻第 m 维属性的累积误差,记为 $\psi_j^{(i,m)}(i<j)$.

$\psi_j^{(i,m)}$ 的表达式如下:

$$\psi_j^{(i,m)}=\sum_{h=i+1}^{j}\phi_h^{(i,m)}\tag{9}$$

下面证明数据源在 $t_h(i<h\leqslant j)$时刻的权值波动均满足式(8)时,$\psi_j^{(i,m)}(1\leqslant m\leqslant M)$ (下文简记为 ψ_j^i)与ε的关系.

定理 2. 当 $\Delta w_h^k\leqslant\varepsilon^{1/2}/K(i<h\leqslant j,1\leqslant k\leqslant K)$ 成立时,$\psi_j^i\leqslant\Delta T(\Delta T+1)(2\Delta T+1)\varepsilon/6$, 其中,$\Delta T=j-i$.

证明:由式(4)和式(6)可得,

$$\left(\phi_h^i\right)^{1/2}=\left|\sum_{k=1}^{K}\left(w_h^k/\sum_{k=1}^{K}w_h^k-w_i^k/\sum_{k=1}^{K}w_i^k\right)\cdot v_h^{(k,m)}\right|\Big/\left|v_h^{(\max,m)}\right|,$$

又结合定理 1,

$$\left(\phi_h^i\right)^{1/2}\leqslant\sum_{k=1}^{K}\left|w_h^k/\sum_{k=1}^{K}w_h^k-w_i^k/\sum_{k=1}^{K}w_i^k\right|.$$

根据式(6),

$$\Delta w_h^k\leqslant\varepsilon^{1/2}/K\Leftrightarrow\left|w_h^k/\sum_{k=1}^{K}w_h^k-w_{h-1}^k/\sum_{k=1}^{K}w_{h-1}^k\right|\leqslant\varepsilon^{1/2}/K,$$

则$\left|w_h^k/\sum_{k=1}^{K}w_h^k-w_i^k/\sum_{k=1}^{K}w_i^k\right|\leqslant(h-i)\cdot\varepsilon^{1/2}/K(i<h\leqslant j,1\leqslant k\leqslant K)$成立.

同时,$\sum_{k=1}^{K}\left|w_h^k/\sum_{k=1}^{K}w_h^k/-w_i^k/\sum_{k=1}^{K}w_i^k\right|\leqslant(h-i)\cdot\varepsilon^{1/2}(i<h\leqslant j)$ 成立.

于是有,$\left(\phi_h^i\right)^{1/2}\leqslant(h-i)\cdot\varepsilon^{1/2}\Rightarrow\phi_h^i\leqslant(h-i)^2\varepsilon(i<h\leqslant j)$.

又因为,

$$\psi_j^i=\sum_{h=i+1}^{j}\phi_h^i\leqslant\sum_{h=i+1}^{j}(h-i)^2\varepsilon=(j-i)(j-i+1)(2(j-i)+1)\varepsilon/6,$$

令$\Delta T=j-i$,则当 $\Delta w_h^k\leqslant\varepsilon^{1/2}/K(i<h\leqslant j)$ 成立时,有 $\psi_j^i\leqslant\Delta T(\Delta T+1)(2\Delta T+1)\varepsilon/6$ 成立.

于是,在 t_i 时刻更新了数据源权值后,当所有数据源在每一时刻 $t_h(i<h\leqslant j)$的权值波动均满足式(8)时,下式

© 中国科学院软件研究所 http://www.jos.org.cn

成立:

$$\psi_j^i \leqslant \Delta T\left(\Delta T+1\right)\left(2\Delta T+1\right)\varepsilon/6 \tag{10}$$

其中,$\Delta T=j-i$.由式(10)可知,累积误差的最大值与两个更新点间的时间间隔有关.

3.3 基于Bernoulli分布的概率预测模型

在实际应用中,对于当前时刻来说,下一时刻的数据源的权值是未知量,因此数据源的权值波动是否满足式(8)也是无法准确得知的.为此,本文提出一种基于Bernoulli分布的概率模型,预测数据源的权值波动是否满足式(8)的概率.

给定阈值ε.由于数据源的可信度独立于流数据到达的每一时刻,则所有数据源的权值波动是否平缓,在流数据到达的每一时刻都是彼此独立的.也就是说,每一时刻所有数据源的权值波动是否均满足式(8)为一系列独立随机事件,并且每一个事件都只有成立和不成立两种对立的情况.因此可以将每一时刻所有数据源的权值波动是否均满足式(8)看作是一次 Bernoulli 实验,它成立与否的概率服从 Bernoulli 分布$\xi\sim B(1,p)$.我们在实验部分也验证了这一点.下面举例说明如何估计 p.

例 1:对于 K 个数据源,假设在 $t_{i_0}\sim t_{j_0}$ 内每一时刻都采用 CRH 算法对数据源权值进行更新,统计所有数据源在每一时刻的权值波动都能满足式(8)的次数,记为 N_0,这段时间内共统计了 $M_0(M_0=j_0-i_0)$次,由统计学知识估计得到概率 $\hat{p}=N_0/M_0$.但是显然,随着时间的增加,数据源的权值波动情况可能会有所变化(外界因素、自身能耗等),因此动态地估计 p 会更准确.

例 2:在例 1 的基础上,得到初始概率的估计值 $\hat{p}$,假设在 t_{i-1} 时刻需要更新数据源的权值以使感知数据流的真值发现结果满足给定精度时,为了动态地估计 p,本文再次统计数据源的权值波动情况,即在 t_i 时刻也需要更新数据源的权值.若此时 $\Delta w_i^k \leqslant \varepsilon^{1/2}/K\,(1\leqslant k\leqslant K)$ 均满足式(8),则 $\hat{p}$ 更新为 $(N_0+1)/(M_0+1)$; 反之, $\hat{p}$ 更新为 $N_0/(M_0+1)$.

根据上述概率模型可知,若预测出在一段时间 $t_{i+1},t_{i+2},...,t_j$ 内,所有数据源的权值波动均满足式(8)的概率为 β,结合定理 2,则该段时间内的累计误差 $\psi_j^i \leqslant \Delta T(\Delta T+1)(2\Delta T+1)\varepsilon/6$ 的概率一定不小于β,其中,$\Delta T=j-i$.因此,当已知前一更新点时,可以通过预测数据源的权值波动在一段时间内满足式(8)的概率,预测累积误差的最大值,来确定下一更新点.此外,由于 $\hat{p}$ 是动态更新的,因此两个更新点间的时间间隔可能是变化的.这即是 CTF-Stream 变频的评估数据源可信度的原理.在下一节中我们会详细地加以说明.

4 基于累积误差预测的数据源可信度更新策略

根据上述结论,本文将如何确定感知数据流上数据源可信度评估的频率这一问题转化为一个最优化问题,在此基础上提出了一种基于累积误差预测的数据源可信度变频更新算法,以处理感知数据流的真值发现问题.通过减少数据源可信度的评估次数,提高感知数据流上真值发现的效率.

4.1 更新点的预测

若在 t_{i-1} 和 t_i 时刻更新了数据源的权值,可以根据数据源的权值波动情况更新概率 $\hat{p}$, 则对于 $t_{i+1},t_{i+2},...,t_j$ 这段时间内,每一时刻所有数据源的权值波动均满足式(8)的概率为 $\hat{p}^{j-i}$, 若 $\hat{p}^{j-i}$ 为一个较大的值,则 $\max\left(\psi_j^i\right)=\Delta T(\Delta T+1)(2\Delta T+1)\varepsilon/6(\Delta T=j-i)$ 的概率较大;反之,则较小,即很难限定累积误差的上界.显然,在后一种情形下,需要在 t_j 时刻重新计算数据源的权值.本文将上述所要解决的问题转化为如下最优化问题.

$$\left.\begin{array}{c}\text{Max}\quad t_j=t_i+\Delta T_0\\ \text{s.t.}\quad \Delta T_0(\Delta T_0-1)(2\Delta T_0-1)\varepsilon/6\leqslant\gamma\\ \hat{p}^{\Delta T_0-1}\geqslant\alpha\end{array}\right\} \tag{11}$$

上述最优化问题包含了下面两个约束函数:

© 中国科学院软件研究所 http://www.jos.org.cn

- $\hat{p}^{\Delta T_0-1}\geqslant\alpha$:当$\Delta T_0$ 满足 $\hat{p}^{\Delta T_0-1}\geqslant\alpha$ 时,由前面的分析可知,$t_{i+1},...,t_{j-1}$ 时刻所有数据源的权值波动均满足式(8)的概率不小于α.
- $\Delta T_0(\Delta T_0-1)(2\Delta T_0-1)\varepsilon/6\leqslant\gamma$:由定理 2 可知,当 $t_{i+1},...,t_{j-1}$ 时刻所有数据源的权值波动均满足式(8)的概率不小于α时,累积误差的最大值 $\max\left(\psi_{j-1}^{i}\right)=\Delta T_0(\Delta T_0-1)(2\Delta T_0-1)\varepsilon/6$ 的概率不小于α.由于累积误差的最大值会随着ΔT_0 的增加而增加,为避免仅由 $\hat{p}^{\Delta T_0-1}\geqslant\alpha$ 预测出的ΔT_0 使累积误差的最大值过大,我们对任意两个更新点之间的累积误差的最大值进行约束.

由于我们希望动态地更新 $\hat{p}$,所以已知 t_i 为更新点时,在 t_{i+1} 时刻也需要更新数据源的权值,即 t_{i+1} 也是一个更新点,则不需要预测 $\Delta w_{i+1}^{k}(1\leqslant k\leqslant K)$ 是否满足式(8),也不存在相对误差 $\phi_{i+1}^{i},\phi_{i+2}^{i},...,\phi_{j-1}^{i}$.

下面,对式(11)中的最优化问题进行改进,如式(12)所示,该式与式(11)相比仅在约束函数上有所改动.

$$\left.\begin{array}{c}\text{Max}\quad t_j=t_i+\Delta T\\ \text{s.t.}\quad (\Delta T-1)(\Delta T-2)(2\Delta T-3)\varepsilon/6\leqslant\gamma\\ \hat{p}^{\Delta T-2}\geqslant\alpha\end{array}\right\}\tag{12}$$

- $\hat{p}^{\Delta T-2}\geqslant\alpha$:由前面的分析可知,仅需考虑 $t_{i+2},...,t_{j-1}$ 这段时间内所有数据源的权值波动情况即可.
- $(\Delta T-1)(\Delta T-2)(2\Delta T-3)\varepsilon/6\leqslant\gamma$:由前面的分析可知,仅考虑相对误差和 $\sum_{h=i+2}^{j-1}\phi_h^{i+1}$ 即可.

4.2 CTF-Stream算法

本文提出基于累积误差预测的数据源可信度变频更新算法 CTF-Stream,处理感知数据流上的连续真值发现问题.CTF-Stream 算法主要分为如下 3 个阶段,结合图 3 加以说明.

(1) 更新数据源的权值:已知更新点 t_i、t_{i+1},调用 CRH 算法更新 t_i、t_{i+1} 时刻的数据源权值 W_i、W_{i+1};

(2) 更新概率 $\hat{p}$:由 W_i、W_{i+1} 统计 $\Delta w_{i+1}^{k}(1\leqslant k\leqslant K)$,重新估计概率 $\hat{p}$;

(3) 预测下一更新点:利用 $\hat{p}$,求解式(12)中的最优化问题,得到下一更新点 t_j.

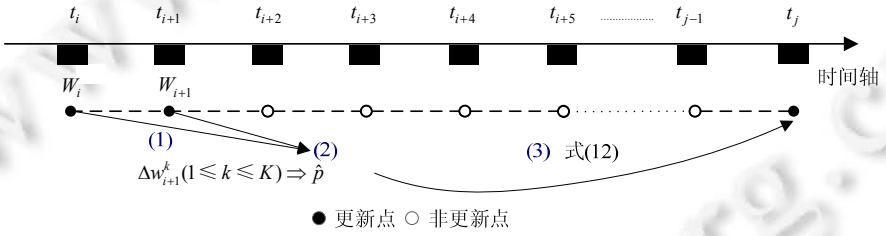


Fig.3 The process of CTF-Stream

图 3 CTF-Stream 的流程

CTF-Stream 的伪代码见表 1,算法中包含了 3 个参数:α,ε和γ,这 3 个参数即是式(12)中涉及的参数,其设置对于算法的性能具有至关重要的影响,下面分别进行说明.

- 概率阈值α.α限制了任意一段时间内,每一时刻所有数据源权值波动均满足式(8)的最小概率.α越小,结合式(12),ΔT 越大,评估数据源权值的次数就会越少.即α越小,CTF-Stream 的效率越高,准确性越低;α越大,CTF-Stream 的效率越低,准确性越高.
- 累积误差阈值γ.γ限制了任意一段时间内,预测的累积误差最大值的上界.γ越小,要求真值发现的准确性越高,结合式(12),数据源权值的评估越频繁.即γ越小,CTF-Stream 的效率越低,准确性越高;γ越大,CTF-Stream 的效率越高,准确性越低.
- 相对误差阈值ε.ε限制了任意相邻时刻的相对误差.由于相对误差是由近似值代替真值所产生的,因此,ε越大,允许的相邻两个时刻的权值波动越大;ε越小,则要求相邻两个时刻的权值波动较为平稳.然而,结合式(12)可知,CTF-Stream 的效率和准确性并不会单纯地随着ε的增大而增大或减小.当数据流上数据源的权值波动较

© 中国科学院软件研究所 http://www.jos.org.cn

大时,ε主要由式(12)中第 2 个不等式约束,即ε越大,CTF-Stream 的效率越高,准确性越低;当数据流上数据源的权值波动较小时,ε主要由式(12)中第 1 个不等式约束,即ε越大,CTF-Stream 的效率越低,准确性越高.

Table 1 The pseudo code of CTF-Stream
表 1 CTF-Stream 的伪代码

算法. CTF-Stream.
Input: $V_i,\alpha,\varepsilon,\gamma$ /*t_i 时刻的观测值集合,阈值$\alpha,\varepsilon,\gamma$*/
Output: $V_i^{(*)}$ /*t_i 时刻的真值集合*/
1: Initialize $t_j\leftarrow 1$, $M\leftarrow 0$, $N\leftarrow 0$ /*初始化前两个时刻为更新点*/
2: for $t_i=1\rightarrow\infty$ do
3: if ($t_i==t_j||t_i==t_j+1$) /*在更新点处更新数据源权值*/
4: set $W_i, V_i^{(*)}$ by call CRH
5: if ($t_i==t_j+1$) /*统计相邻时刻的权值波动情况*/
6: M++
7: for $k=1\rightarrow K$ do
8: set Δw_i^k by Formula (6)
9: if($\Delta w_i^k > \varepsilon^{1/2}/K$) /*判断数据源的权值波动是否满足式(8)*/
10: break
11: else if ($k==K$) /*所有数据源的权值波动是否满足式(8)*/
12: N++
13: $\hat{p}=N/M$ /*动态更新 $\hat{p}$ */
14: $t_i\leftarrow t_i-1$
15: set t_j by Formula (12) /*预测下一次更新数据源权值的时间*/
16: if ($t_i-t<2$)
17: $t_j\leftarrow t_i+2$
18: $t_i\leftarrow t_i+1$
19: else /*不需要更新数据源的权值*/
20: $W_i\leftarrow W_{i-1}$
21: set $V_i^{(*)}$ by Formula (4)
22: return $V_i^{(*)}$

基于上述分析,用户可以通过调节参数的设置,灵活地改变 CTF-Stream 的性能,以满足自身对真值发现准确性和效率的要求.同时,在实验部分,本文会结合实验结果再次说明这 3 个参数对算法性能的影响.

表 1 中对 $\hat{p}$ 的更新实质上采取采样的方法,即不断地在相邻两个更新点处统计数据源的权值波动情况,在对 $\hat{p}$ 值更新的同时,也积累了样本,使 $\hat{p}$ 随着时间的增加越来越准确.然而,一些过期的数据也可能会影响 $\hat{p}$ 的准确性,即 $\hat{p}$ 过多地偏离了实际的 p,对此可以采用两种方式处理:(1) 每隔一段时间重新启动算法以舍弃全部历史数据;(2) 采用滑动窗口的方式,即始终保留最新一部分的统计数据.显然,后一种方式能够同时充分地利用历史数据和新数据,因此实验过程中我们采用的是第 2 种方式更新 $\hat{p}$.此外,从表 1 中可以看出,若当前时刻不是更新点,则直接利用式(4)计算真值,时间复杂度为 $O(n)$,其中,n 为当前时刻到达的感知数据的数据量.这与执行迭代过程相比大大提高了真值发现的效率,且在计算真值的过程中不需要遍历历史数据.此外,虽然 CTF-Stream 在形式上是一个最优化问题,但实质上可以不调用任何数学程序直接计算结果.因为式(12)可以直接转化为求同时满足两个约束的ΔT 的最大值.因此可以先令 $\Delta T=\lfloor \log\alpha/\log\hat{p}+2 \rfloor$, 然后再验证$(\Delta T-1)(\Delta T-2)(2\Delta T-3)\varepsilon/6\leqslant\gamma$是否满足,若满足,则为最终结果;否则,只需令$\Delta T=\Delta T-1$ 再次验证即可.这一过程的时间复杂度是极低的.实验结果也表明,表 1 中预测更新点的部分所需要的时间即使在数据量为 $50k$ 时,也几乎为 0.

5 实验及结果分析

本文通过在真实的感知数据集合上进行实验,进一步验证了 CTF-Stream 的有效性和准确性.在此之前,本文首先验证第 3.3 节提出的概率分布假设的合理性,实验结果表明,本文提出的概率分布假设是合理的.同时 CTF-Stream 在准确性和效率方面都具有良好的性能,特别是在对连续到达的感知数据流进行真值发现时,CTF-Stream 的处理速度始终大于感知数据流的到达速度.

© 中国科学院软件研究所 http://www.jos.org.cn

5.1 实验数据

• Intel Berkeley 实验室数据:该数据集是由部署在Intel Berkeley实验室的 54 个 Mica2 传感器节点,对同一室内空间在 36 天内每隔 30s 进行一次采样所得的监测数据.本文选取其中的 25 个传感器在一天内采集到的温度和湿度两个属性进行真值发现.对于缺失数据,采用该传感器在邻近时刻的值进行填补,最终得到的数据即为本文的实验数据集.

• 天气数据:该数据集包含了 5 个知名气象网站对美国 10 个城市在 6 天内每隔 30min 进行一次观测所得到的气象数据.本文选取该数据集上的温度和湿度两个属性进行真值发现.同时,我们将 10 个不同城市的温度、湿度属性各看作是一个属性,即该数据集中包含 20 个属性.

这两个数据集在数据变化模式方面存在显著的区别:(1) 由于数据本身的特点及采集的频率等原因,相比 Intel Berkely 实验室数据,天气数据集中数据的变化幅度较大;(2) 相比 Intel Berkely 实验室数据,天气数据集中数据源的权值变化较大.下文将结合准确性和效率两方面的实验结果具体分析这两个实验数据集的区别.

特别需要指出的是,本文将数据集合的大小定义为时间戳数目×数据源数目×属性数目.在改变数据集合的大小时,本文仅通过调整选取的时间戳的总数目来调整实验过程中数据集合的大小,即对于包含的时间戳数目不同的数据集合,其每一时间戳内的数据源数目、属性数目均相同.此外,本文根据两个数据集本身的采样频率,定义实验时数据流上的时间戳,即对于 Intel Berkeley 实验室数据,本文记每 30s 为一个时间戳;对于天气数据,本文记每 30min 为一个时间戳.

表 2 系统地列出了这两个真实数据集的信息.

Table 2 Real-World datasets
表 2 真实数据集

	数据源数目	属性数目	数据集合总大小	时间戳变化范围
Intel Berkeley 实验室数据	25	2	50 000	200, 400, 600, 800, 1 000
天气数据	5	20	17 500	35, 70, 105, 140, 175

5.2 实验设置

5.2.1 对比的方法

(1) CRH 算法[5].CRH 是一种能够处理异构数据类型的真值发现迭代框架.针对不同的数据类型及用户需要,该框架可以插入不同的损失函数和标准化函数.为了增加算法的可比性,本文在实验中采用式(2)作为 CRH 的损失函数,式(3)作为 CRH 的标准化函数.

(2) GTM 算法[3].GTM 是一种基于贝叶斯网络的针对连续型数据的真值发现方法.它同样是一种迭代方法.本文参考文献[3]在实验过程中使用的参数以及选取的真实数据集合的特点,确定实验时 GTM 算法的先验参数.

5.2.2 准确性评估

对于 Intel Berkely 实验室数据集, 由于真值未知, 而 CTF-Stream 算法在更新点处执行的是 CRH 算法.因此,实验中将每一时刻都调用 CRH 算法获得的真值发现结果看作真值;对于天气数据集,选取 accuweather.com 上观测到的气象数据作为真值.本文选用均方误差 RMSE 度量 CTF-Stream 算法的准确性.显然,均方误差 RMSE 越小,算法的准确性越高.

5.2.3 效率评估

实验中,本文不仅采用算法的运行时间评估效率,还引入更新次数作为度量标准,即平均每一时间戳更新数据源权值的次数.由于 CTF-Stream 是从减少评估数据源权值的次数这一角度来提高真值发现的效率,所以引入更新次数这一度量标准十分必要.显然,运行时间和更新次数越小,算法的效率越高.

5.3 实验结果

5.3.1 概率假设验证

本文在第 3.3 节中将每一时刻所有数据源的权值波动是否均满足式(8)看作是彼此独立的 Bernoulli 实验,

© 中国科学院软件研究所 http://www.jos.org.cn

因此本文通过在 Intel Berkely 实验室数据集上多次统计 N 个时间戳内所有数据源的权值波动,来判断其均满足式(8)的频数是否服从二项分布来验证本文假设的合理性.实验中令 $\varepsilon=5\times10^{-5}$,$K=10$,$N=10$.

图 4 表明,实际统计的频数和二项分布($\xi\sim B(10,0.25)$)拟合出的频数基本相吻合,即可以认为,本文在第 3.3 节中提出的概率分布假设合理.

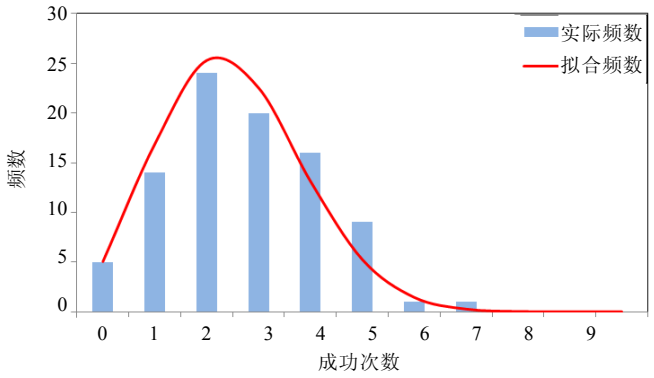


Fig.4 Hypothesis verification

图 4 假设验证

5.3.2 CTF-Stream 的效率

- 运行时间

图 5 和图 6 分别反映了当概率阈值 α、相对误差阈值 ε 和累积误差阈值 γ 发生变化,CTF-Stream、CRH、GTM 对两个不同的感知数据集进行真值发现时,运行时间存在着差异.参数设置如下:

(a) $\varepsilon=5\times10^{-4}$,$\gamma=1$,$\alpha$ 分别为 0.65,0.7,0.75,0.8,0.85;

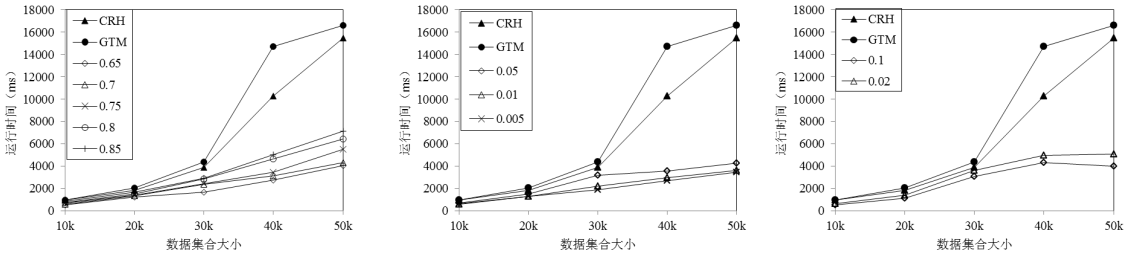
(b) $\alpha=0.7$,$\gamma=0.1$,ε 分别为 0.005,0.01,0.05;

(c) $\alpha=0.7$,$\varepsilon=10^{-3}$,γ 分别为 0.1,0.02.

对于 Intel Berkeley 实验室数据,由图 5 可以看到,当数据量较小时,不同参数下 CTF-Stream 的运行时间都要远远小于 CRH 和 GTM,即数据量越大,CTF-Stream 的效率优势越显著.此外,图 5(a)还表明,α 值较大的曲线始终位于 α 值较小的曲线的上面,因为随着 α 的增大,ΔT 减小,更新数据源权值的次数增加,运行时间随之增加.同理,图 5(c)中随着 γ 的减小,导致 ΔT 减小,更新数据源权值的次数增加,运行时间随之增加,因此 $\gamma=0.02$ 的曲线始终位于 $\gamma=0.1$ 的曲线的上面.而在图 5(b)中,随着 ε 的减小,运行时间缩短.这说明,该数据集上数据源的权值波动很小,即使是在 $\varepsilon=5\times10^{-3}$ 的条件下,每一时刻所有数据源的权值波动均满足式(8)的概率也较高,因此 ΔT 几乎不会受式(12)中的第 2 个不等式约束影响.所以,在 γ,α 固定的情况下,根据式(12)中的第 1 个不等式约束,ε 增大会导致 ΔT 的减小,进而导致运行时间有所增加.

对于天气数据,由图 6 可以看到,虽然在各个参数条件下,CTF-Stream 与迭代方法相比依然表现出效率优势,但却不如在 Intel Berkeley 实验室数据集上的效果显著.这表明,在天气数据集中,数据源的权值波动与 Intel Berkeley 实验室数据集上的数据源的权值波动相比,较为剧烈,那么,为确保真值发现的精度,需要较为频繁地评估数据源的权值.特别要指出的是,由于在天气数据集上,数据源权值波动幅度较大,因此图 6(b)中的实验结果与图 5(b)中正好相反,即类似于对图 5(b)的分析,当数据源权值波动较为剧烈时,ΔT 主要受式(12)中的第 2 个不等式约束,则 ε 增大会导致 ΔT 增加,运行时间缩短.而图 6(a)、图 6(c)运行时间随参数变化的趋势与图 5(a)、图 5(c)中相同.

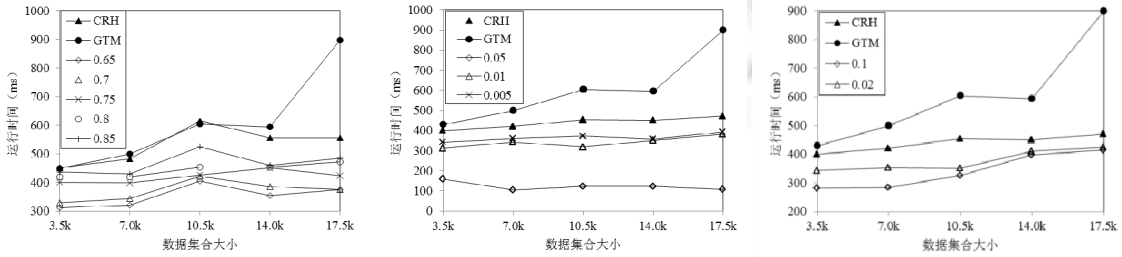
© 中国科学院软件研究所 http://www.jos.org.cn



(a) α变化 (b) ε变化 (c) γ变化

Fig.5 Intel Berkeley library dataset (1)

图 5 Intel Berkeley 实验室数据(1)



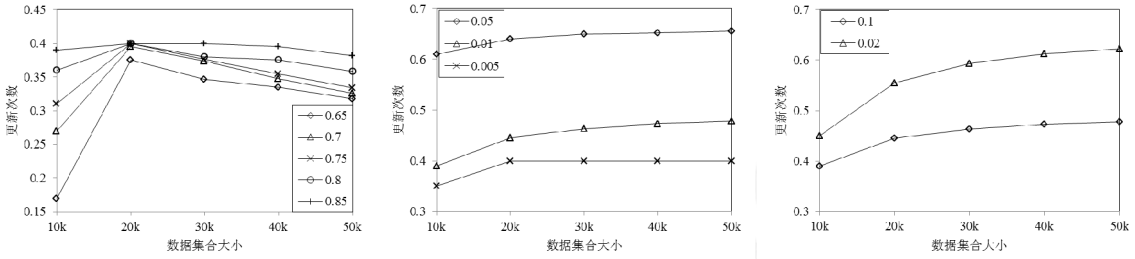
(a) α变化 (b) ε变化 (c) γ变化

Fig.6 Weather dataset (1)

图 6 天气数据(1)

- 更新次数

图 7 和图 8 分别反映了当概率阈值α、相对误差阈值ε和累积误差阈值γ发生变化,CTF-Stream、CRH、GTM 对感知数据流进行真值发现时,更新次数的差异.CTF-Stream 的参数设置与图 5 和图 6 中对应的参数设置相同.注意,由于 CRH 和 GTM 的更新次数始终为 1,因此在图 7 和图 8 中未作比较.



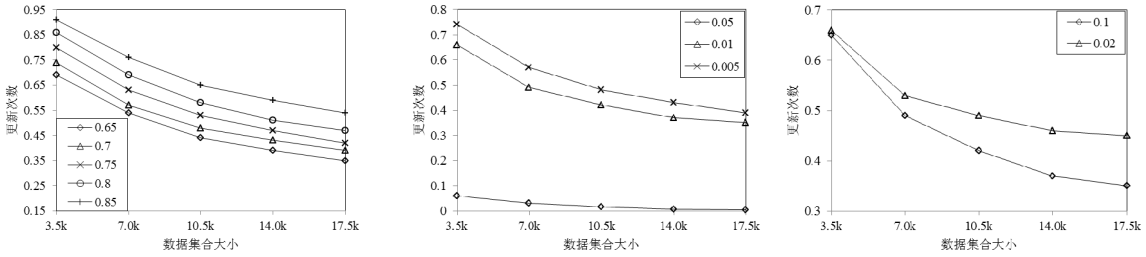
(a) α变化 (b) ε变化 (c) γ变化

Fig.7 Intel Berkeley library dataset (2)

图 7 Intel Berkeley 实验室数据(2)

对于 Intel Berkeley 实验室数据,图 7(a)表明,随着α的增大,更新次数增大,这与图 5(a)中随着α的增大,运行时间增加相吻合.同理,图 7(b)和图 7(c)的实验结果也与图 5(b)和图 5(c)中的实验结果相吻合.此外,图 7 中的更新次数随着数据量的增加,变化幅度很小.说明该数据集上数据源的权值波动随着时间戳的增加并未有较大改变.对于天气数据,图 8(a)~图 8(c)也分别验证了图 6(a)~图 6(c)中的结论.特别需要指出的是,随着数据量的增加,更新次数减小,这表明数据集后一段时间数据源权值波动与前一段时间相比幅度较小,即前一段时间对数据源的权值更新较为频繁.导致数据集合所包含的时间戳数目越多,平均更新次数越少.

© 中国科学院软件研究所 http://www.jos.org.cn



(a) α变化 (b) ε变化 (c) γ变化

Fig.8 Weather dataset (2)

图 8 天气数据(2)

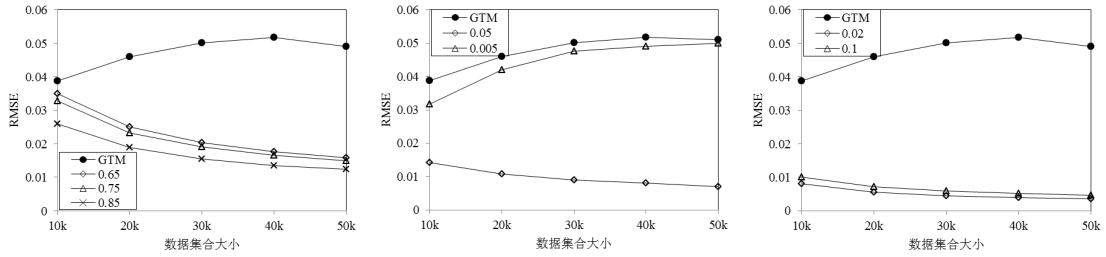
5.3.3 CTF-Stream 的准确性

图 9 和图 10 分别反映了当 CTF-Stream 的概率阈值α、相对误差阈值ε和累积误差阈值γ变化时,其准确性的变化情况,并与 GTM 算法进行了对比.参数设置如下:

(a) $\varepsilon=5\times10^{-4}$,$\gamma=1$,$\alpha$分别为 0.65,0.75,0.85;

(b) $\alpha=0.7$,$\gamma=0.1$,ε分别为 0.005,0.05;

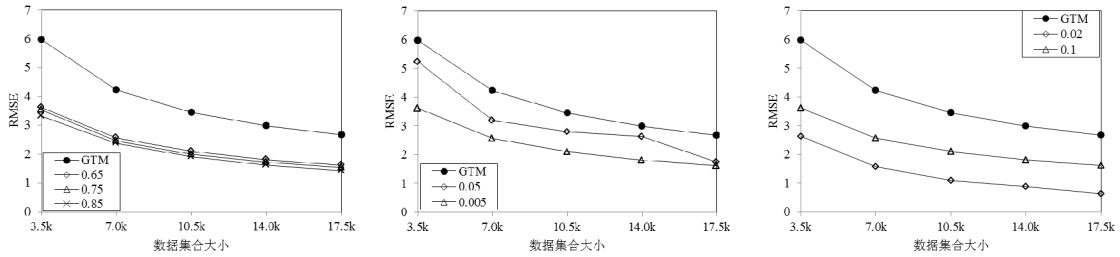
(c) $\alpha=0.7$,$\varepsilon=10^{-3}$,γ分别为 0.1,0.02.



(a) α变化 (b) ε变化 (c) γ变化

Fig.9 Intel Berkeley library dataset (3)

图 9 Intel Berkeley 实验室数据(3)



(a) α变化 (b) ε变化 (c) γ变化

Fig.10 Weather dataset (3)

图 10 天气数据(3)

对于 Intel Berkeley 实验室数据,从图 9(a)中可以看出,当γ不变时,随着α的增加,均方误差不断减小,因为α越大,更新数据源权值的次数越多,均方误差越小.类似地,因为γ越大,更新数据源权值的次数越少,均方误差越大,因此γ=0.02 的曲线始终在γ=0.1 的曲线下面.图 9(b)和图 9(c)也是同理,即对应的更新次数越多,准确率越高.

对于天气数据,其准确性随着参数的变化与更新次数随参数的变化也是一致的,即依照图 8 中的分析,对应参数所造成的更新次数越多,图 10 中准确性越高.同时,由对图 6 的分析可知,天气数据集中,数据源的权值波动

© 中国科学院软件研究所 http://www.jos.org.cn

较大,当α为一个较小值(α=0.65)时,如图 8(a)所示,更新次数也相对较多,即对于权值波动较大的数据集,α的变化对更新次数的影响相对较小.因此α在图 10(a)设置的参数范围内发生变化时,CTF-Stream 准确性的变化也较小.此外,CTF-Stream 在更新点处执行 CRH 算法.而在对静态数据集进行真值发现时,CRH 与 GTM 相比表现出很高的准确性[5].因此,即使 CTF-Stream 并没有在数据流上连续地更新数据源的权值,由于其在更新点处的准确性很高,相比于每一时刻都更新数据源权值的迭代算法 GTM,CTF-Stream 依然表现出很好的准确性.并且,在 Intel Berkeley 实验室数据集上,GTM 的均方误差随着数据集合的增大而增大,而 CTF-Stream 在大部分参数设置下其准确率都是随着数据集合的增大而减小的,这表明,CTF-Stream 在处理大规模感知数据流时具有很大优势.

上述实验在验证了 CTF-Stream 的高效性和准确性的同时,也充分说明了本文选取的两个真实数据集合具有比较显著的差异,进而说明 CTF-Stream 在处理不同类型和变化模式的数据集合时,均能表现出良好的性能.

6 结 论

真值发现作为数据集成中一种冲突消解的有效手段,在传统数据库领域已经得到了广泛的研究.但是由于时间、空间复杂度等限制,基于传统数据库的真值发现技术无法应用于一种越来越普遍的数据模型——数据流中.本文针对一种特殊的数据流——感知数据流上的连续真值发现问题进行了研究.结合感知数据自身及其应用特点,提出了一种变频评估数据源可信度的策略以平衡感知数据流真值发现的效率和准确率.本文首先定义并研究了感知数据流真值发现的相对误差和累积误差,以及它们较小时数据源在相邻时刻可信度的变化应满足的条件.进而提出一种概率模型,以预测数据源在相邻时刻可信度的变化满足该条件的概率.最后,整合上述结论,将感知数据流真值发现中的累积误差预测问题转化为一个最优化问题,在限制了累积误差的前提下,最大化数据源可信度的评估周期以提高效率.在此基础上提出了一种基于累积误差预测的数据源可信度变频更新算法——CTF-Stream,对连续到达的感知数据流进行真值发现.CTF-Stream 结合历史数据,动态地确定数据源可信度的评估周期,同时以一定概率确保真值发现结果的准确性.CTF-Stream 通过减少更新数据源可信度的次数,减少了迭代过程的执行,极大地提高了真值发现的效率.最后,本文在真实的感知数据集合上进行实验,实验结果表明,本文提出的算法在处理数据流上的真值发现问题时具有较高的准确率和效率.

References:

[1] Yin XX, Han JW, Yu PS. Truth discovery with multiple conflicting information providers on the Web. IEEE Trans. on Knowledge and Data Engineering, 2007,20(6):796−808. [doi: 10.1109/TKDE.2007.190745]

[2] Galland A, Abiteboul S, Marian A, Senellart P. Corroborating information from disagreeing views. In: Proc. of the WSDM. New York, 2010. 131−140. https://hal.inria.fr/inria-00429546/document

[3] Zhao B, Han JW. A probabilistic model for estimating real-valued truth from conflicting sources. In: Proc. of the QDB. Istanbul, 2012. http://web.engr.illinois.edu/~hanj/pdf/qdb12_bzhao.pdf

[4] Zhao B, Rubinstein BIP, Gemmell J, Han JW. A Bayesian approach to discovering truth from conflicting sources for data integration. PVLDB, 2012,5(6):550−561. [doi: 10.14778/2168651.2168656]

[5] Li Q, Li YL, Gao J, Zhao B, Fan W, Han JW. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In: Proc. of the SIGMOD. Snowbird, 2014. 1187−1198. http://hanj.cs.illinois.edu/pdf/sigmod14_jgao.pdf

[6] Li Q, Li YL, Gao J, Demirbas M, Zhao B, Su L, Fan W, Han JW. A confidence-aware approach for truth discovery on long-tail data. PVLDB, 2014,8(4):425−436.

[7] Dong XL, Berti-Equille L, Srivastava D. Integrating conflicting data: The role of source dependence. PVLDB, 2009,2(1):550−561.

[8] Dong XL, Berti-Equille L, Srivastava D. Truth discovery and copying detection in a dynamic world. PVLDB, 2009,2(1):562−573. [doi: 10.14778/1687627.1687691]

[9] Dong XL, Berti-Equille L, Hu YF, Srivastava D. Global detection of complex copying relationships between sources. PVLDB, 2010,3(1-2):1358−1369.

[10] Dong XL, Berti-Equille L, Hu YF, Srivastava D. Solomon: Seeking the truth via copying detection. PVLDB, 2010,3(1-2): 1617−1620. [doi: 10.1145/1966883.1966887]

© 中国科学院软件研究所 http://www.jos.org.cn

[11] Dong XL, Gabrilovich E, Murphy K, Dang V, Horn W, Lugaresi C, Sun S, Zhang W. Knowledge-Based trust: Estimating the trustworthiness of Web sources. PVLDB, 2015,8(9):938−949.

[12] Pochampally R, Das-Sarma A, Dong XL, Meliou A, Srivastava D. Fusing data with correlations. In: Proc. of the SIGMOD. Snowbird, 2014. 433−444. http://lunadong.com/publication/fusionWCorr_sigmod.pdf

[13] Li X, Dong XL, Lyons K, Meng W, Srivastava D. Truth finding on the deep Web: Is the problem solved. PVLDB, 2012,6(2): 97−108.

[14] Song SX, Zhang AQ, Wang JM, Yu PS. SCREEN: Stream data cleaning under speed constraints. In: Proc. of the SIGMOD. Melbourne, 2015. 827−841. http://ise.thss.tsinghua.edu.cn/sxsong/doc/15sigmod-screen.pdf

[15] Cao L, Yang D, Wang QY, Yu YW, Wang JY, Rundensteiner EA. Scalable distance-based outlier detection over high-volume data streams. In: Proc. of the ICDE. 2014. 76−87. [doi: 10.1109/ICDE.2014.6816641]

[16] Zhao Z, Cheng J, Ng W. Truth discovery in data streams: A single-pass probabilistic approach. In: Proc. of the CIKM. Shanghai, 2014. 1589−1598. http://er2004.cse.ust.hk/~wilfred/paper/cikm14a.pdf

[17] Li JZ, Li JB, Shi SF. Concepts, issues and advance of sensor networks and data management of sensor networks. Ruan Jian Xue Bao/Journal of Software, 2003,14(10):1717−1727 (in Chinese with English abstract). http://www.jos.org.cn/ch/reader/create_pdf.aspx?file_no=20031007&journal_id=jos

[18] Zhao Z, Ng W. A model-based approach for rfid data stream cleansing. In: Proc. of the CIKM. Hawaii, 2012. 862−871. http://www.cs.ust.hk/~wilfred/paper/cikm12b.pdf

[19] Cheng SY, Li JZ, Yu L. Location aware peak value queries in sensor networks. In: Proc. of the INFOCOM. 2012. 486−494. [doi: 10.1109/INFCOM.2012.6195789]

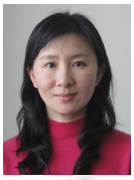
[20] Raza U, Camerra A, Murphy A, Palpanas T, Picco GP. Practical data prediction for real-world wireless sensor networks. IEEE Trans. on Knowledge and Data Engineering, 2015,PP(8):1. [doi: 10.1109/TKDE.2015.2411594]

[21] Li YL, Li Q, Gao J, Su L, Fan W, Han JW. On the discovery of evolving truth. In: Proc. of the SIGKDD. Sydney, 2015. 675−684. http://www.cse.buffalo.edu/~lusu/papers/KDD2015Yaliang.pdf

附中文参考文献:

[1] 李建中,李金宝,石胜飞.传感器网络及其数据管理的概念、问题与进展.软件学报,2003,14(10):1717−1727. http://www.jos.org.cn/ch/reader/create_pdf.aspx?file_no=20031007&journal_id=jos

李天义(1992－),女,辽宁锦州人,学士,主要研究领域为数据质量.



李芳芳(1977－),女,博士,讲师,CCF 会员,主要研究领域为数据库技术,传感器网络 CPS 数据管理.

谷峪(1981－),男,博士,副教授,CCF 高级会员,主要研究领域为图,空间数据管理.

于戈(1962－),男,博士,教授,博士生导师,CCF 会士,主要研究领域为数据管理理论与技术,分布与并行系统.

马茜(1988－),女,硕士生,CCF 学生会员,主要研究领域为感知数据管理.

© 中国科学院软件研究所 http://www.jos.org.cn