数据流挖掘研究

史金成[1,2],胡学钢[1]
(1.合肥工业大学 计算机与信息学院,安徽 合肥 230009;
2.铜陵学院 计算机系,安徽 铜陵 244000)

摘 要:上世纪末,为适应网络监控、入侵检测、情报分析、商业交易管理和分析等应用的要求,数据流技术应运而生。数据流独特的特点,对传统数据的处理方法带来了很大的挑战。介绍了数据流的有关概念及数据流挖掘的特点,讨论了数据流挖掘的研究现状。最后,举例说明了数据流挖掘的应用,并展望了数据流挖掘未来的研究方向。

关键词:数据流挖掘;聚类;分类;频繁模式
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2007)11-0011-04

Study on Data Stream Mining

SHI Jin-cheng[1,2], HU Xue-gang[1]
(1. School of Computer and Information, Hefei University of Technology, Hefei 230009, China;
2. Department of Computer, Tongling College, Tongling 244000, China)

Abstract: Since the end of last century, data stream techniques have been advanced to meet the requirements of network monitoring, in-break detecting, information analyzing, business transaction management and analyzing, etc. The unique character of data streams brings a big challenge to traditional data processing methods. Introduces some relational concepts of data streams, as well as the character and recent research of data stream mining. Finally, main applications and future research directions of data stream mining are put forward.

Key words: data stream mining; clustering; classification; frequent pattern

0 引 言

近年来,随着硬件技术的高速发展,人们获取数据的能力得到了很大的提高。经常会出现这样的情况:大量需要处理的数据以很快的速度产生。例如,一个网站一天的点击次数可以达到几千万次,电话公司的大型交换机每天可以记录高达几千万条的通话记录。由于数据量太大、而且数据产生的速度太快,如果仍然按传统的数据库应用模式来处理这些数据,则这样的任务将是不可能完成的。于是,一种新型的解决方案应运而生,其最大特点是,待处理的数据以一种动态、流式的形式出现(即数据流,data stream),对数据流中的数据只能按顺序进行一次或有限次的访问。如今,数据流尤其是数据流挖掘已成为新一代计算理论与应用的研究热点之一。

收稿日期:2007-02-10
基金项目:安徽省高等学校自然科学研究重点项目(2006KJ027A)
作者简介:史金成(1979-),男,安徽宿松人,讲师,硕士研究生,研究方向为数据挖掘和知识工程;胡学钢,教授,博士,硕士生导师,研究方向为数据挖掘和知识工程。

1 数据流有关概念

1.1 数据流

数据流就是大量连续到达的、潜在无限的数据的有序序列[1]。我们日常生活中有很多数据流的例子,如股票交易所的股票价格信息,环境温度的监测数据,电信部门的通话记录,网站的点击记录以及传感器网络中的监测数据等。和传统数据库中相对静态的数据不同,数据流有时效性、实时性、无限性和瞬时性等特点[2]。人们在处理传统数据时,可以完整、详细地收集数据,且将它们全部存储在数据库中,然后再由计算机来处理,传统的商务和事务型数据库管理系统强调维护数据的完整性和一致性,追求的目标是系统吞吐量和平均响应时间,而不是系统的自适应性和查询服务质量,更加不注重数据处理的时空限制。由于数据流的特殊性,短时间内有大量数据连续到达,这些数据具有随时间动态变化的趋势,要求快速即时响应,且这些数据往往又是高维的,所以在处理数据流时,必须注重时空限制,对数据流中的数据通常采取的是单一扫描的线性算法,用精度换时间,尽量在对数据的一次访问中获得较优的解和有用的信息。一些数据库应用中常

用的操作,如排序、找最大值或最小值、计数等 Blocking 操作在数据流的处理中都是行不通的,且一般的数据流算法是不可回溯的。

1.2 数据流模型

要想对数据流进行挖掘,首先必须建立相应的数据流模型[3]。数据流 D 可以看作是由不断到达的数据组成的动态增长数据集,即 $D = \{a_1, a_2, \cdots, a_i, \cdots, a_j, \cdots\}$,其中 a_i 为数据,如果 $j > i$,则 a_i 先于 a_j 到达;如果 $|j - i| = 1$,则 a_i 与 a_j 相邻。

根据 a_i 的描述的不同,数据流模型可以分为以下 3 类:

(1) 时间序列数据流。用来描述时间序列数据,如每分钟 NASDAQ 成交量、每 2 分钟所观测到的 IP 流量。这里的 a_i 可以定义为:

$$a_i = (j, I_i)$$

其中,I_i 为 i 时刻产生的数据。

(2)Cash Register 数据流。这种模型应用非常普遍,如超市的收银机记录、对 IP 地址的监控记录。这里的 a_i 可以定义为:

$$a_i = (j, I_i)$$

其中,$I_i \geqslant 0$ 为新产生的数据,很显然有:$a_i[j] = a_{i-1}[j] + I_i$。

(3)Turnstile 数据流。这种模型可以记录动态的插入和删除操作,如记录拥挤的地铁站中乘客的人数。这里的 a_i 可以定义为:

$$a_i = (j, U_i)$$

其中,U_i 可以大于 0,也可以小于 0。很显然,有 $a_i[j] = a_{i-1}[j] + U_i$。

1.3 数据挖掘及数据流挖掘

所谓数据挖掘,是指从海量的数据中揭示出隐藏在数据背后的经验与知识,发现其中有意义的模式、规则或异常点[4]。

对于静止的数据,它的挖掘过程可以分成两个部分:先把数据收集起来存储在数据库中,然后在数据库上应用各种挖掘技术进行挖掘。而对于流数据,它的收集过程和挖掘过程是同时进行的,必须以最快的速度,从不断到来的数据中挖掘出感兴趣的模式(interesting pattern),来响应用户的实时查询。

基于数据库的传统数据挖掘都要求数据分析处理的精确性,但是对于流数据,由于数据收集时间的有限性,分析处理速度的有限性,我们无法在精确度上做过多要求,使得挖掘结果的近似性(approximation)成了不同于传统数据挖掘的一个特点。

数据流挖掘的特点决定了它比传统的数据挖掘要复杂,因此一般在数据流挖掘中要解决以下一些主要问题:

(1) 设计低内存消耗的挖掘算法。由于数据流中的数据是连续产生的,需要的内存很大,所以只能实时地处理数据流,因此在设计挖掘算法时必须充分利用有限的内存,使得一次能处理更多的数据。

(2) 设计计算高效的挖掘算法。在数据流挖掘过程中,内存中储存的数据始终都是最新产生的数据,必须在这些数据还没被后来的数据替代前,完成对它们的处理。这就要求在设计算法时必须考虑算法的效率,如何在最短的时间内完成对数据的处理。

(3) 设计一趟挖掘算法。因为数据流是连续不停地产生的,我们也没有任何可以暂时阻塞数据流的操作,所以对所有的数据都只能扫描一趟,必须设计一趟挖掘算法。

2 数据流挖掘研究现状

目前对数据流的研究主要集中在以下几个方面:对数据流模型的研究、对数据流查询的响应研究、数据流管理研究以及数据流挖掘研究等。其中,数据流挖掘方面的研究成果主要集中在数据流的聚类、分类和频繁模式挖掘方面。

2.1 聚类算法

聚类(clustering)分析[5]也称无监督学习,其任务是将数据划分成有意义或有用的组(簇)。聚类算法可以分成划分方法(partitioning method)、基于层次的方法(hierarchical method)和基于密度的方法(density - based method)等几类,算法的选择取决于数据的类型、聚类的目的和应用。数据流的聚类算法不同于传统数据的聚类算法,必须是增量式的,对聚类的表示要简洁,对新数据的处理要快速,对噪音和异常数据必须是稳健的。因此,基于数据流的聚类算法要在一个相对较小的内存空间上,对数据流进行一遍扫描后,把数据集合分为一个个簇集。典型的用于数据流的聚类算法有 STREAM[6]算法和 CluStream[7]算法。

(1)STREAM 算法。

Jiawei Han 等人提出了基于 K - Means 的 STREAM 算法。算法以 k 为参数,随机选取 k 个对象,每个对象代表一个划分的平均值或中心。剩下的每个对象,根据其与各个中心的距离,把它赋给最近的划分,这样把 n 个对象分为 k 个划分。然后重新计算每个划分的平均值,根据新的中心来划分对象。这个过程不断重复,使划分内具有较高的相似度,而划分间的相似度较低。通常,这个算法需要大量的内存空间,还要求对数据流随机存取,这些都是和数据流挖掘的前提相

抵触的,因此需要改进以便应用于数据流。

Guha 等人对基于 K - Means 的 STREAM 算法进行了改进,提出了改进后的 STREAM 算法,这个算法只对数据流进行一次扫描,且占用很小的空间。设有 n 个对象或元组,k 为划分的个数,也表示有 k 个中心,这个算法只需 $O(nk)$ 的时间和 $O(n\varepsilon)(\varepsilon < 1)$ 的空间。

(2)CluStream 算法。

C. Aggarwal, J. Han 等人提出了数据流聚类算法 CluStream,该算法首次提出把数据流看成一个随时间变化的过程,而不是一个整体进行聚类分析,该算法有很好的可扩展性,可产生高质量的聚类结果,尤其是在数据流随时间变化较大时比其它算法产生更高质量的聚类。CluStream 算法不仅能给出整个数据流聚类的结果,还可以给出任意时间范围内的聚类结果,以及进行数据流的进化分析。它使用了两个过程来处理数据流聚类问题:首先,使用一个在线的 micro - cluster 过程对数据流进行初级聚类,并按一定的时间跨度将 micro- cluster 的结果按一种称为金字塔时间窗口(pyramid time frame) 的结构储存下来。同时,使用另一个离线的 macro - cluster 过程,根据用户的具体要求对 micro - cluster 聚类的结果进行再分析。

朱蔚恒等人进一步指出了 CluStream 算法存在的一些不足,主要有以下几个方面:首先,对非球形的聚类,micro - cluster 过程不能给出一个很好的描述;其次,micro - cluster 过程利用数据与最近子聚类中心的距离以及一个预定的距离阈值来判断数据是否属于该子聚类,但距离某一聚类中心近的点不一定就属于该聚类;最后,micro - cluster 判断一个新来的数据是否属于某子聚类的方法是根据该数据与子聚类中心的距离,因此,不能保证生成子聚类的数据最后都包含在一个与子聚类中心距离不大于 r 的圆形邻域中。针对上述 CluStream 算法的不足,朱蔚恒等人提出了基于密度与空间的 ACluStream(arbitrary - shape CluStream) 聚类算法[8]。该算法首先积累一定的数据,然后使用任意的聚类算法对它们进行聚类,再把聚类的结果划分成一个个互不相交的聚类块。记录这些聚类块的特征值向量并使用一个 Hash 表 H 来记录指向这些向量的指针。然后对每一个新来的数据点 d,如果它属于一个聚类块,那么将其加入该聚类块中,并修改该聚类块的特征值;否则,判定它是否为孤立点。当积累的数据量达到窗口大小时,对内存中的聚类块进行分析,根据实际情况或者结合、或者分解,并将保存在内存中的聚类块特征按 pyramid time frame 的结构记录下来,便于后续查询。

2.2 分类算法

分类(classification)[5]的任务就是确定对象属于哪个预定义的目标类。传统数据挖掘算法的一个目标就是在有限的内存上从大量的数据中构建数据模型,这个目标已经被很多分类算法所实现。如 SPRINT[5]算法、RainForest[5]算法等。但是,这些算法都要求对数据集进行多次扫描,因此并不适用于数据流的分类。

针对数据流的分类,Domingos 等人提出了一种高效的增量决策树算法——VFDT[9](very fast decision tree)。VFDT 能用固定的内存和固定的时间为每个样本构建一棵决策树,有效地解决了时间、内存和样本对高速数据流上的数据挖掘的限制。它利用 Hoeffding 边界来保证算法的输出模型与批量学习(batch learner)的输出模型是趋向于一致的。

VFDT 和其它大多数机器学习方法一样,假设数据是从静态分布中随机获取的,不能反映数据随时间变化的趋势。因此,P. Domingos 和 G. Hulten 引入了滑动窗口技术,对 VFDT 算法进行改进,提出了 CVFDT[10](Concept - adapting Very Fast Decision Tree) 算法,除了保留 VFDT 算法在速度和精度方面的优点外,增加了对数据产生过程中变化趋势的检测和响应,使得算法更好地适应高速数据流的分类。CVFDT 利用样本窗口来有效维护决策树的更新和模式的一致性,然而它并不需要在每个新样本到来的时候都学习新的模式,而是通过增加相应新样本的总数来更新充分统计量,相应地减少滑动窗口中旧的节点数量。

2.3 频繁模式挖掘算法

频繁模式(frequent pattern)挖掘[5]的任务是通过对一个数据集进行统计,找出出现频率最高的一个数据或一组数据,从而挖掘出频繁模式。虽然频繁模式挖掘已经被广泛研究,出现了许多经典算法,如 Apriori 算法、FP - growth 算法、CLOSET 算法、CHARM 算法[5]等,但这些算法难以增量式更新,不适合数据流挖掘。

针对数据流的频繁模式挖掘,Giannella 等人用 FP - tree[11](frequent pattern tree)为数据结构,并在此基础上提出了 FP - Stream[11~13]算法。该算法采用倾斜时间窗口技术来维护频繁模式以解决数据流频繁模式挖掘中的时间敏感问题。

FP - stream 结构包括两部分:一个用来捕获频繁和准频繁项信息的频繁模式树(FP - tree)和为每个频繁模式提供的倾斜时间窗口(tilted - time window)表。

3 数据流挖掘的应用

由于数据流在人们的日常生活中很常见,所以研

究数据流挖掘有很大的现实意义，下面介绍两种比较典型的应用。

3.1 数据流挖掘在股市交易中的应用

在股市交易过程中，产生的数据以数据流的形式出现。对股市交易所产生的数据流进行动态挖掘，可以帮助人们预测股市的起伏，甚至能捕捉转瞬即逝的个股买/卖点或在众多股票中选出应买卖的股票。目前人们已开发出许多商业的软件包，如 NET-PROPHET、Falcon 等。

数据流挖掘在这方面的另一个应用是研究股市炒作的快速检测算法和技术。纽约的 Dow Jones 市场是人工式的交易，NASDAQ 是电子交易；而中国的沪市和深市都是电子交易。这些电子交易每天产生大量的数据流，无疑已超出人工处理的能力，这就必须应用计算机算法进行智能自动监控。通过对股市交易数据流的挖掘，可以识别出哪些是合法炒作，哪些是非法炒作，在炒作之前，哪些局内人士异常地大量买进，从中渔利。这些对稳定我国的股票交易市场有很大的积极作用。

3.2 数据流挖掘在零售业务中的应用

在零售业方面，计算机的利用率已经越来越高，特别是随着条形码技术的广泛使用，商场、超市等零售企业的交易数据也是以数据流的形式出现。通过对这种数据流进行挖掘，可以起到以下几个方面的作用：

(1)进行销售、顾客、产品、时间和地区的多维分析；

(2)对促销活动有效性进行分析，以此提高企业利润；

(3)对顾客忠诚度进行分析，以留住老客户，吸引新客户；

(4)挖掘关联信息，以形成购买推荐和商品参照，以帮助顾客选择商品，提高销量。

另一方面，通过研究顾客的购买行为，分析人们的购买模式，估计他们的收入和家庭成员数目，在庞大的数据体中找出哪些人适合寄广告或折扣券，哪些人喜欢哪一类的折扣券，哪些人应给予的折扣多些，哪些人应给予的折扣少些，哪些产品摆在一起会比分别放在各自的类中卖得更快更多，这些都能增加利润。

4 结论与展望

介绍了数据流挖掘有关的概念、特点、研究现状和应用。但是，因为人们从事数据流挖掘研究的时间还很短，所以还有很多值得研究的东西。我认为，以下几个方面将是以后数据流挖掘研究的方向：

(1) 寻找新的适于数据流的数据结构和建模方法，研究有效度量数据相似性的方法；

(2) 研究针对数据流的高效异常挖掘算法；

(3) 研究数据流基于时间变化的特性，探索数据流变化的表示与建模方法，挖掘数据进化和变化的趋势；

(4) 研究数据流基于约束的聚类分析算法；

(5) 研究数据流的局部周期挖掘算法。

参考文献：

[1] Nan Jiang, Le Gruenwald. Research Issues in Data Stream Association Rule Mining[J]. SIGMOD Record, 2006, 35(1):14-19.

[2] Brian B, Shivnath B. Models and issues in data stream systems[C]//Proceedings of 21st ACM Symposium on Principles of Database Systems. New York: ACM Press, 2002:1-16.

[3] 倪志伟，黄 玲，李锋刚，等. 数据流管理与挖掘研究[J]. 合肥工业大学学报：自然科学版，2005，28(9)：1157-1162.

[4] Tan Pang-Ning, Steinbach M, Kumar V. 数据挖掘导论[M]. 范 明，范宏建译. 北京：人民邮电出版社，2006.

[5] 梁 循. 数据挖掘算法与应用[M]. 北京：北京大学出版社，2006.

[6] Guha S, Mishra N, Motwani R, et al. Clustering data streams: Theory and practice[J]. Knowledge and Data Engineering, IEEE Transactions, 2003, 15(3):515-528.

[7] Aggarwal C, Han J, Wang J, et al. A framework for clustering evolving data streams[C]// Proc of Int Conf on Very Large Data Bases (VLDB'03). Berlin, Germany: [s. n.], 2003.

[8] 朱蔚恒，印 鉴，谢益煌. 基于数据流的任意形状聚类算法[J]. 软件学报，2006，17(3)：379-386.

[9] Domingos P, Hulten G. Mining high-speed data streams[C]// Proceedings of the ACM Conference on Knowledge and Data Discovery(SIGKDD). Edmonton, Alberta: [s. n.], 2000:71-80.

[10] Gurmeet S M. Approximate frequency counts over data streams[C]// Proc VLDB Conference. Hong Kong, China: [s. n.], 2002:346-357.

[11] Giannella C, Han Jia-wei, Pei Jian, et al. Mining frequent patterns in data streamsat multiple time granularities[C]// Proc of the NSF Workshop on Next Generation Data Mining. [s.l.]: [s.n.], 2002.

[12] 刘学军，徐宏炳，董逸生，等. 挖掘数据流中的频繁模式[J]. 计算机研究与发展，2005，42(12)：2192-2198.

[13] 张 昕，李晓光，王大玲，等. 数据流中一种快速启发式频繁模式挖掘方法[J]. 软件学报，2005，16(12)：2099-2105.