编者按:随着经济管理与科学技术的不断结合,现代审计已经远远超出了仅对财务会计进行审查的狭窄范围,不断向管理领域和技术领域渗透。IT 审计是技术审计的一个典型,它实质上是对计算机软件和硬件及整个信息系统的审计。近年来,我国对 IT 审计人才的需求大幅增长,IT 审计理论及 IT 审计人才培养问题逐渐成为学界研究的热点之一。为将研究推向深入,并推动人才培养和学术成果交流,本刊特在"审计"栏目下开辟"IT 审计"专栏,以期为研究者提供交流和探讨的平台。热忱欢迎广大专家学者不吝赐稿。

数据流挖掘及其在持续审计中的可用性研究

谷瑞军,陈圣磊

(南京审计学院 信息科学学院, 江苏 南京　211815)

摘　要:随着企业信息化程度的提高和互联网的普及,每天都会产生海量的实时数据,而数据流挖掘则为分析海量数据提供了一种新途径。数据流挖掘中的聚类、分类、离群点检测等算法的研究取得了进展,为在持续审计中应用数据流挖掘提供了可行性。本文提出的一种基于数据流挖掘的持续审计模型,克服了传统持续审计模型对审计端的存储能力要求高、占用大量硬件资源、联机分析时间长、对异常数据的发现滞后等缺点。

关键词:数据流挖掘;持续审计;审计模型;聚类;分类;离群点检测

中图分类号:TP391　**文献标识码**:A　**文章编号**:1672－8750(2011)01－0036－05　**收稿日期**:2010－11－01

作者简介:谷瑞军(1979—　),男,山东菏泽人,南京审计学院信息科学学院讲师,博士,主要研究方向为数据挖掘与计算机审计;陈圣磊(1977—　),男,山东兖州人,南京审计学院信息科学学院讲师,博士,主要研究方向为机器学习。

基金项目:国家自然科学基金(70971067/G0112);国家社会科学基金(10BGL016);江苏省高校自然科学研究项目(09KJD520006)

一、引言

网络入侵检测、股市分析、传感器网络等实时监控领域需要对大量的动态数据进行实时的、连续的数据收集与分析。由于连续到达数据的多样性、快速性、时变性等特点,形成了难以预测的无界数据流。文献[1]给出了数据流定义:数据流是一个有序数据点序列 $X_1, X_2, \cdots, X_K, \cdots$,对应着一个时间序列 $t_1, t_2, \cdots, t_k, \cdots$,表示数据点 X_k 在时刻 t_k 到达,同时规定当 $t_i < t_j$ 时,数据点 X_i 比数据点 X_j 先到达。每一个数据点 Xi 是一个 d 维向量,记作 $X_i = (x_i^1, x_i^2, \cdots x_i^d)$,分别代表数据点 X_i 的 d 个属性值。如果我们把传统的存储于数据库中的数据称为静止的数据,那么数据流就是动态的、实时数据,它的数据采集过程和数据挖掘过程是同时进行的,因而必须以最快的速度从不断到来的数据流中挖掘出用户感兴趣的模式。对流数据进行实时挖掘称为数据流挖掘,它有如下特点:第一,流数据是不停产生的,而内存的大小有限,只能实时地进行处理;第二,存储在内存中的数据都是最新产生的,必须在这些数据还没被后来的数据替代之前对它进行及时处理;第三,没有任何操作可以暂时阻塞数据流,所有的数据只能扫描一次;第四,流数据往往天生就是高维的[2]。

数据流挖掘的特点决定了它比传统的数据挖掘要复杂,近几年来,数据流挖掘已成为数据挖掘研究领域一个重要分支。另外,随着信息化程度的提高,越来越多的行业会产生数据流,因此,数

据流挖掘的应用范围也在不断扩大。持续审计中需要审计实时、动态的数据流，构建基于数据流挖掘的持续审计模型是本研究的创新。

二、数据流挖掘研究进展

数据流的研究主要包括对数据流模型的研究、数据流管理研究、对数据流查询的响应研究以及数据流挖掘研究等。目前，数据流挖掘的研究热点主要集中于数据流的聚类、分类、离群点检测和频繁模式挖掘等方面，本节主要分析数据流挖掘中的聚类、分类和离群点检测的最新研究进展。

（一）数据流聚类算法

聚类（Clustering）是指对于一个已给的数据对象集合，将其中相似的对象划分为一个或多个组（称为“簇”，Cluster）的过程[3]。同一个簇中的元素彼此相似，而与其他簇中的元素相异。与传统数据的聚类算法不同，数据流聚类算法是在一个相对较小的内存空间里，对数据流进行一遍扫描后就可以把数据集划分为一个个簇集（cluster）。

经典的数据流聚类算法包括 STREAM[4]、CluStream[5] 和 DenStream[6]。STREAM 算法是一种基于划分的聚类算法，它聚焦于解决 k－中位数问题，即把度量空间中的 n 个数据点聚类成 k 个簇，使得数据点与其簇之间的误差平方和最小。STREAM 算法实现了单次扫描，时间复杂度为 O（kn）。与传统数据的聚类算法相比，STREAM 算法有更好的性能，并能产生更高质量的聚类结果，但是，STREAM 算法没有考虑数据流的演变，聚类的结果可能受控于过期的数据点。基于层次的 CluStream 算法并不对数据流进行整体聚类分析，而是把数据流看成一个随时间变化的过程。该算法使用两个过程来对数据流进行聚类分析：首先，使用一个在线的 micro-cluster 过程对数据流进行初次聚类，并按一定的时间跨度将 micro-cluster 的结果以一种称为“金字塔时间窗口”的结构进行储存。然后，使用另一个离线的 macro-cluster 过程，根据用户的个性化要求对 micro-cluster 的聚类结果进行再次分析。CluStream 通过使用倾斜时间框架，保存了数据流演变的历史信息，在数据流变化剧烈时仍可以产生高质量的聚类结果，但是，它没有考虑历史数据的衰减问题，当被应用于高维数据流的聚类时，CluStream 算法往往表现不佳。诸如 STREAM、CluStream 等扩展划分和层次的方法，由于采用距离度量，仅在对球形的数据流进行聚类分析时表现良好，因此它们不能很好地处理任意形状的数据流。DenStream 算法沿袭了 CluStream 的处理框架，把聚类分析的过程划分为联机和脱机两部分。DenStream 算法扩展了传统数据集聚类算法中基于密度的方法 DBSCAN，着眼于处理任意形状的数据流聚类问题。同时，它还强调了孤立点检测问题，将孤立点与正常数据元素区分开来。

（二）数据流分类算法

数据流实时动态到达的特点决定了数据流分类方法基本上都是增量式的。数据流分类一般都假设样本是平稳分布的，但事实上，新数据的概念信息可能会随着时间的延续而与历史数据相比发生改变，这种改变称为概念漂移[7]。VFDT[8] 及 CVFDT[9] 是两种具有代表性的数据流分类算法。

VFDT（very fast decision tree）是一种基于 Hoeffding 不等式建立决策树的方法，分类器的性能可以渐近于传统算法生成的分类器，差异的界由 Hoeffding bound 决定：对于一个范围是 R 的随机变量 r，假设存在 n 个样本点，样本均值为 $\bar{r}$。Hoeffding bound 指 r 的真实期望 $E(r)$ 以 $1-\delta$ 的概率大于 $\bar{r}-\varepsilon$，即 $Pr[E(r)\geq\bar{r}-\varepsilon]=1-\delta$，其中 $\varepsilon=\sqrt{R^2\ln(1/\delta)/(2n)}$。该算法通过不断地将叶节点替换为决策节点生成决策树，其中每个叶节点都保存有关于属性值的统计信息。不同于传统的批处理方法，VFDT 处理每一个决策树的节点时仅依赖于整个数据的部分子样本，对整个流数据仅作一次扫描，得到一个近似的解。但是，VFDT 中没有处理连续值属性问题，同时也没有考虑概念漂移的处理方法。引入概念漂移的思想，Hulten 等人在 VFDT 的基础上提出了改进算法 CVFDT[9]，每当有新样本到达时，就把 VFDT 应用到滑动窗口上。滑动窗口（Sliding Window）模型基于这样一个事实，即“用户对于最近的数据更感兴趣”。滑动窗口可以对少量的近期数据作细节分析，而对大量的历史数据仅仅给出一个概要视图，这样就只需存储小的数据窗口，从而减少了对内存的需求。CVFDT 通过不断地把 VFDT 算法应用到固定大小的滑动窗体上的方式，从不断变化的数据流上生成决策树。该算法在叶节点可能会发生概念漂移时产生一棵备选子树，并且在新子树变得更精确时用其替代原先的子树。另外，集成分类器也是解决概念漂移的一种途径。

Wang 等人提出了一种利用加权的多个分类器挖掘概念漂移的数据流分类方法[7]。该方法首先从数据流中训练几个分类器,然后根据测试数据集上的分类精度期望进行加权。集成学习方法既提高了学习模型的效率,也提高了分类精度。

(三) 数据流离群点检测算法

离群点检测问题是数据挖掘的重要研究方向之一,它被广泛应用于网络入侵检测、信用卡恶意透支检测等领域。给定数据集 D 和阈值 ξ、σ,对于样本 $X \in D$,如果存在至多 ξ 个样本点位于 X 的 σ 距离之内,则称 X 为离群点[10]。离群点检测算法分为基于统计的方法、基于密度的方法、基于距离的方法和基于偏离的方法等。上述方法均需多次扫描数据库,所以不适合数据流挖掘。如何在有限的运行空间上对数据流进行一次或较少次数的扫描,实现高效的数据流离群点检测,这是具有重要研究意义的课题。目前,数据流离群点检测已成为国内外研究者的关注热点。

Jagadish 等人采用信息论的方法给出了时间序列中离群点的定义框架,并提出了一种在时间序列中挖掘离群点的有效算法[11]。Choy 提出了一种适合大样本、静态时间序列的基于频谱的离群点检测算法 SODA,该算法可用来挖掘定时的、类型确定的离群事件[12]。Yu 等人利用小波变换的多分解特性,从原始数据集中消除聚类,从而达到发现离群点的目的[13]。Ma 等人将一种支撑向量机方法应用于时序离群点挖掘中,其思想是先将时间序列投影到一个向量空间,然后使用 SVM 进行离群点挖掘[14]。文献[15]利用核密度估计对数据分布进行近似处理,它支持基于密度和基于距离两种离群点定义。该算法效率较高,可对实时达到的数据进行在线分布式离群点挖掘,但它只适用于分布式传感器网络数据流。文献[16]在基于密度的局部离群点挖掘算法的基础上,进行增量式改进,提出离群点挖掘算法 IncLOF。当有数据被删除或插入时,只需要重新计算受影响对象的局部离群因子即可。该算法运行时间远少于基于密度的局部离群点挖掘算法,但需要计算受影响对象的 k 距离、局部可达密度、局部离群因子,因此需对数据集进行三次扫描。基于 k 均值分区的数据流离群点检测算法(DSOKP)首先将数据流划分为区块,并对每块进行 k 均值聚类,每块得到 k 个均值参考点,然后对整个数据流的 m 个均值参考点按基于距离的离群点定义进行挖掘[17]。该算法效率很高,但由于对数据流进行分块处理,无法及时捕捉流数据的概念漂移。文献[18]则从聚类的角度研究离群点检测。首先将数据流划分成块,然后使用 k 均值将块聚集成固定数量的簇。不同于数据流聚类中只保留摘要信息的做法,该方法同时保留候选离群点和每个簇的均值,供后续的固定大小的数据流块使用,以确保检测到的候选离群点是真正的异常值。

三、一种基于数据流挖掘的持续审计模型

近几年来,随着信息化程度的提高,数据流在人们的日常生活中很常见,例如股票交易数据流、web 网站访问数据流等。数据流挖掘的应用范围在不断扩大,下面将尝试把数据流挖掘拓展到持续审计领域。本节首先提出持续审计的技术要求,然后分析数据流挖掘在持续审计中的可用性,最后提出一种基于数据流挖掘的持续审计模型。

(一) 持续审计的技术要求及实现模型

随着信息技术的发展,企业等经济组织对信息及时性的要求越来越强,因而对审计信息的时效性也提出了更高的要求。传统审计主要是基于一定时间段内实施的审计,审计信息的及时性和可靠性较差,难以适应复杂多变审计环境的要求。“持续审计”(Continuous Audit,简称 CA)正是为了适应信息社会发展需要而产生的新方法,根据 AICPA/CICA 研究报告对持续审计的定义,“持续审计是独立审计师用以对委托项目的相关事项以一系列实时或短时间内生成的审计报告,对其提供书面鉴证的一套审计方法”[19]。在整个审计过程中,信息技术起到了关键性的作用。持续审计是依托于信息技术发展的审计方法创新,因此持续审计技术可行性研究非常重要。CA 的主要特征包括:第一,审计过程的实时性要求;第二,审计手段的电子性要求;第三,强调风险评价和控制;第四,保留审计“独立性”基本特征;第五,范围包含内部和外部审计[20]。其中,审计过程的实时性要求表现在三个环节,即信息证据收集的实时性、监控和分析的实时性和发布审计报告的实时性。可以看出,实时性是持续审计最重要的特征,也是对实现技术的第一要求。

从技术途径上看,CA 方法分为三类:应用嵌入式技术、应用代理技术和混合式技术[21]。CA

的实现模型也有多种，实现模型是 CA 系统整体实现的理论模式，以技术实现为外在逻辑形式，内含审计各要素排列以及审计契约各方关系所形成的抽象逻辑框架。其中，基于数据仓库和数据集市的数据挖掘持续审计模型及实现技术为当前审计和计算机交叉领域的研究热点之一。文献[22]详细描述了数据仓库和数据集市技术在持续审计中的应用，如图 1 所示。

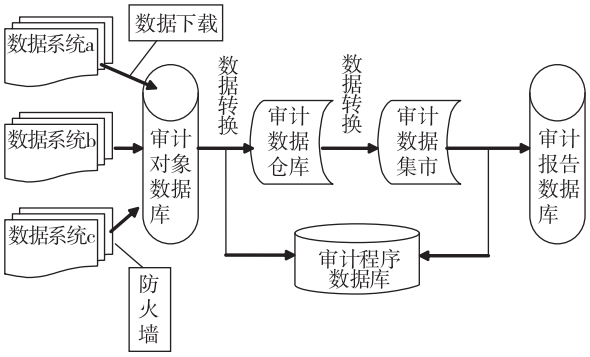

图 1　基于数据仓库和数据集市持续审计模型[22]

（二）一种新的持续审计模型

从图 1 可以看出，在外部持续审计中，审计端需下载被审计端的数据库，并存储在审计端专用的数据库里。对于银行、税务、海关等单位，每天产生的数据量巨大，这对审计端的数据采集、存储、转换和处理能力都提出很高的要求。数据仓库的特点是先存储、转换，然后作联机分析或数据挖掘。虽然利用数据仓库技术可以对被审计端进行连续审计，但尚存在两个不足：一是对审计端的存储能力要求高，数据集是多个被审计端的合集，会占用大量的硬件资源；二是随着数据量的加大，联机分析耗时增长，对异常数据的发现会有滞后。

事实上，像银行、证券等金融行业的被审计端，每天产生的数据十分巨大，所以可以视为数据流，它具备不停产生、动态变化等特点。对审计端来说，没有必要存储全部的被审计数据，只需实时监控，存储可疑的异常数据，为审计实务提供审计线索。本文结合持续审计实现模式和数据流挖掘技术，提出一种基于数据流挖掘的持续审计模型，如图 2 所示。在基于数据流挖掘的持续审计模型中，可综合使用聚类、分类和离群点检测技术，目的在于检测数据流的异常变化，及时发现可疑数据，并把这些数据存储到审计线索数据库，还可设置报警器，及时通知审计人员根据专业知识判断审计意义的异常点。

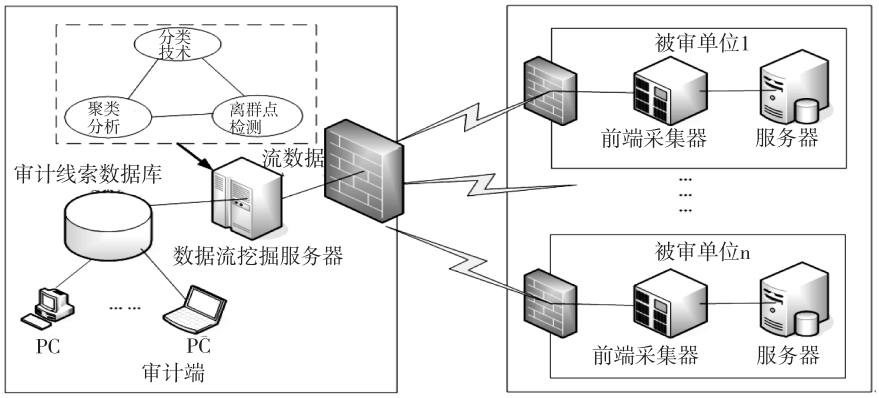

图 2　基于数据流挖掘的持续审计模型

四、总结

持续审计是依托于信息技术发展的审计方法创新，持续审计的发展依赖于信息技术的发展，数据流挖掘则为分析持续审计中产生的海量数据提供了一种新方法。本文在对数据流挖掘的最新进展进行综述的基础上，分析了在持续审计中应用数据流挖掘的可行性，并提出一种基于数据流挖掘的持续审计模型，拓展了数据流挖掘的应用领域，也为审计实践中开展持续审计提供了一种新思路。

参考文献：

[1] Golab L, Ozsu M T. Issues in data stream management [J]. ACM SIGMOD Record, 2003, 32:5 - 14.

[2] Marascu A, Masseglia F. Mining sequential patterns from temporal streaming data[EB/OL]. [2010 - 09 - 09]. http://www.di.uniba.it/ ~ malerba/activities/mstd/.

[3] Han Jiawei, Kamber M. Data mining: concepts and techniques[M]. 2nd ed. San Francisco: Morgan Kaufmann Publishers, 2006.

[4] Guha S, shra N, Moweani R, et al. Clustering data

streams: theory and practice[J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15: 515 - 528.
[5] Aggarwal C, Han Jiawei, Wang Jianyong, et al. A framework for clustering evolving data streams[C]. Proc of Int Conf on Very Large Data Bases(VLDB03). San Francisco:Morgan Kaufmann Publishers,2003:81 - 92.
[6] Cao Feng, Ester M, Qian Weining, et al. Density-based clustering over an evolving data stream with noise[C]. Proc of the SIAM Conference on Data Ming. Philadelphia: Society for Industrial and Applied Mathematics, 2006: 328 - 339.
[7] Wang Haixun, Fan Wei, Yu Philip S, et al. Mining concept-drifting data streams using ensemble classifiers[C]. Proc. of SIGKDD. New York: ACM, 2003:226 - 235.
[8] Pedro D, Geoff H, Mining high-speed data streams[C]. Proc of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2000:71 - 80.
[9] Hulten G, Spencer L, Domingos P. Mining time changing data streams[C]. Proc. of the ACM Int1 Conf on Knowledge Discovery and Data Mining. New York: ACM, 2001:97 - 106.
[10] Hawkins D M. Identification of outliers[M]. London : Chapman and Hall, 1980.
[11] Jagadish H V, Koudas N, Muthukrishnan S. Mining deviants in a time series database[C]. Proc. of the 25th VLDB. San Francisco: Morgan Kaufmann Publishers, 1999: 102 - 113.
[12] Choy K. Outlier detection for stationary time series[J]. Journal of Statistical Planning and Inference, 2001, 99: 111 - 127.
[13] Yu Dantong, Sheikholeslami G, Zhang Aidong. Findout: finding outliers in very large datasets[J]. Knowledge and Information Systems, 2002, 4:387 - 412.
[14] Ma J, Perkins S. Time-series novelty detection using one-class support vector machines[C]. Proc. of the International Joint Conference on Neural Networks. Los Alamitos :IEEE, 2003: 1741 - 1745.
[15] Subramaniam S, Palpanas T, Papadopoulos D, et al. Online outlier detection in sensor data using non-parametric models[C]. Proc. the 32nd International Conference on Very Large Data Bases. New York: ACM, 2006: 187 - 198.
[16] Pokrajac D, Lazarevic A. Incremental local outlier detection for data streams[C]. Proc. of the 2007 IEEE Symposium on Computational Intelligence and Data Mining. Los Alamitos: IEEE, 2007:504 - 515.
[17] 倪巍伟,陆介平,陈耿,等. 基于k均值分区的数据流离群点检测算法[J]. 计算机研究与发展, 2006(9): 1639 - 1643.
[18] Elahi M, Li Kun, Nisar W, et al. Efficient clustering-based outlier detection algorithm for dynamic data stream[C]. Proc. of the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery. Los Alamitos:IEEE,2008:298 - 304.
[19] AICPA/CICA. Continuous auditing, research report [R]. The Canadian Institute of Chartered Accountants, Toronto, Ontario, 1999.
[20] 陈良华,张越,陈小燕. 连续审计的概念特征和实现模型研究[J]. 审计研究,2007(3):72 - 76.
[21] 陈伟,张金城. 计算机辅助审计原理及应用[M]. 北京:清华大学出版社,2008.
[22] Rezaee Z, Sharbatogh lie A, Elam R, et al. Continuous auditing: building automated auditing capability[J]. Auditing: A Journal of Practice and Theory, 2002, 21: 147 - 163.

(**责任编辑**:黄　燕　许成安)

Research on Data Stream Mining and Its Availability to Continuous Audit

GU Rui-jun, CHEN Sheng-lei

Abstract: With the development of enterprise informatization and the popularity of the Internet, massive real-time data are being produced every day. Data stream mining provides one novel approach to analyzing massive real-time data. In this paper the sate-of-art in this field is presented, and its availability to continuous audit is discussed. Finally, based on data stream mining, one continuous audit model is proposed, which overcomes the disadvantages of huge storage capacity requirements, long-time online analysis and the delayed finding of abnormal data.

Key words: data stream mining; continuous audit; auditing model; clustering; classification; outlier detection