孤立点数据挖掘技术在审计信息化中的应用研究

周 喜,曾 丽①
(湖南商学院 会计学院,湖南 长沙 410205)

[摘 要] 孤立点分析技术是数据挖掘的重要内容之一,可广泛应用到电信、信用卡欺骗检测、贷款审批、客户分类、气象预报和网络入侵检测等领域。在审计工作中,也可采用孤立点检测算法对审计数据进行判断和检测,帮助审计人员及时发现隐藏的审计线索,提高审计效率,孤立点数据挖掘技术比发现规律性的其他挖掘技术具有更好的现实应用价值。

[关键词] 审计信息化; 孤立点分析; 数据挖掘
[中图分类号] F239 [文献标识码] A [文章编号] 1673 - 0755(2011)05 - 0055 - 03

随着信息技术的不断发展及企业数据库管理信息系统的数据海量增加,传统以查账为主审计方法将遇到计算机技术的挑战,让审计人员不得不重新调整作业方法,选择计算机审计方式检查被审计单位的经济活动,发挥现代审计监督的作用[1]。目前,虽然有许多通用软件公司开发了相关的审计软件,也加快了我国审计信息化的步伐,但由于这些软件功能的局限性,只是把传统审计方法及流程计算机化而已。如何充分利用先进信息化技术,如孤立点分析技术去发现海量数据中隐藏或未知的信息,让"智能"的数据处理方法帮助审计人员迅速发现异常交易或事项,快速确定审计事项及重点,提高审计效率,降低审计成本及风险,是未来审计信息化研究的重点和难点[2]。目前,国内学者在审计软件、审计信息化和数据挖掘在审计中的应用的研究并不多,主要有陈伟副教授对审计软件现状及发展趋势、基于数据匹配技术的审计证据获取方法和信息系统审计新的安全服务模式的研究;吕新民教授对信息化环境下审计项目管理及数据挖掘在审计数据分析中的应用方面的研究;陈丹萍教授和辛金国等人对基于数据挖掘技术的联网审计的研究;张炳才等人对基于欧式距离孤立点挖掘方法在审计中的应用相关的研究等。

一 数据挖掘技术

数据挖掘是从大量的、不完全的、随机的、模糊的和在噪声的实际应用数据中发现趋势、规则和模式的过程,他融合了现代统计、决策理论、数据库管理和机器学习等多学科的知识,这门广义的交叉学科汇聚了不同领域的研究者,如数据库、并行计算、数理统计、可视化和人工智能等方面的学者和工程技术人员。数据挖掘技术一般分为聚类分析、分类分析、关联分析、序列分析、时间序列分析、依赖关系分析、偏差分析和孤立点分析等。数据挖掘基本过程分为:问题定义、数据收集、数据预处理、数据挖掘和结果解释及评估。数据挖掘主要算法及方法包括神经网络(Neural Networks)、序列模式分析(Sequential Pattern)、决策树(Decision Tree)、遗传算法(Gentic Algorithous)、模糊算法(Fuzzy Algorithous)、聚类分析(Cluster Analysis)、粗糙集规则(Rough Set Rule)、关联分析(Assoliantion Analysis)等[3]。

二 孤立点分析定义及方法

孤立点分析(Outlier Detection)是指数据集中可能包含一些不符合数据一般模型与行为的对象,如部分极端值等。孤立点分析也是数据挖掘中一个重要的研究方向。如在金融行业里,可利用基于孤立点分析的欺诈模型对每个信用卡客户近期及历史用卡行为进行分析,如检测到不寻常的信用卡使用情况,就拟确定为交易有欺诈行为,及时与持卡人联系确认交易是否存在欺诈,银行是否予以授权、是否冻结对方资金等操作。

孤立点分析方法包括基于统计(分布)的孤立点检测、基于距离的孤立点检测、基于密度的孤立点检测、基于聚类的孤立点检测、基于偏离的孤立点检测、基于深度孤立点检测等六种方法。

(一)基于统计(分布)的孤立点检测方法

统计方法是先假设在给定的数据集合有一个分布或概率模型,然后采用不一致性检验来定义和发现孤立点。基于统计(分布)的方法虽然易于理解,实现起来也较为方便,但只对数据分布满足某种概率分布的数值型单变量(属性)数据才有效,不适合用于多维空间的孤立点检测。因此,基于统计(分布)的孤立点检测方法应用范围受到大大的限制。

[收稿日期] 2011 - 07 - 11
[基金项目] 湖南省教育厅科学研究项目资助(编号:11C0735);湖南省教育科学"十二五"规划项目资助(编号:KJX011CGD026)
[作者简介] 周喜(1980 -),男,湖南永州人,湖南商学院会计学院讲师。
①湖南商学院会计学院讲师。

(二) 基于距离的孤立点检测方法

为了有效的避免基于统计(分布) 方法中的数据分布适应性的限制,拓宽多个标准分布的不一致检测的思想,Knorr 和 NG 引入了基于距离的孤立点的概念,他们认为如果某个点与数据集中大多数点之间的距离都超过了某个阀值,这个点就是孤立点。当数据集不满足任何分布标准时,该方法仍能有效地发现孤立点,也能处理多维的数据,唯一不足的就是要求用户合理地设置阀门,人工的介入及干预的因素较大。目前常用的基于距离的孤立点算法有: Index-based 算法、Nested-loop 算法、Cell-based 算法等[4]。

(三) 基于密度的孤立点检测方法

M. Breuning 提出一个数据集的不同部分表现出不同的特征,孤立点的检测需要将一个点与其相邻点的特征与其他点的特征做为重点参考因素。基于密度的方法也是在基于距离的方法基础上建立起来的,其重要参数是数据点之间的距离参数和某一给定范围内的数据点的个数参数,将这两个参数结合起来就是密度的概念。基于密度的方法能够检测出基于距离孤立点检测方法所不能识别的局部孤立点,也不会遗漏周围的孤立点数据,这种检测方法放弃了以前绝对孤立点观点,并纳入局部孤立点的内容,也更贴近 Hawkins 的孤立点定义。

(四) 基于聚类的孤立点检测方法

传统的大部分聚类算法如 Sting、Dbscan、Clarans 等都具有异常数据的处理能力,这些聚类数据挖掘算法主要目标是在相似的基础上收集数据来分类,产生对人们有意义的聚类信息,孤立点的产生只是副产品而已。在聚类数据挖掘处理过程中,这些算法将数据集中异常的事项作为噪音而忽略或容忍,虽然不利于异常信息的检测,但其最大的优点就是扫描数据集的效率较高,适应于大规模数据集。

(五) 基于偏离的孤立点检测方法

基于偏离的孤立点检测不采用统计(分布) 和基于距离的度量值来发现和确定异常数据,它是通过对检测数据集的主要特征来确定孤立点的,所有与给出的主要特征描述"偏离"的数据集都被认为是孤立点。基于偏离的方法主要检测技术有序列异常技术和 OLAP 数据立方体技术。

(六) 基于深度的孤立点检测方法

基于深度的孤立点检测方法中,数据集中的每一个数据都被映射为维空间中的一个点,同时也定义了其深度,根据不同的深度将这些数据划分成不同的层次。整个划分过程中,异常数据一般都是被划分到较浅层次的数据,这些数据是孤立点的可能性较大。此方法存在的缺陷是对四维及四维以上的数据处理效率较低,只适合对二维和三维空间上的数据检测。

三 孤立点技术在审计信息化中的应用

(一) 基于孤立点分析的审计数据挖掘模型

基于孤立点分析的审计数据挖掘模型主要包括: 数据预处理模块、孤立点检测模块、孤立点算法库、孤立点规则库等。基于孤立点分析的审计数据挖掘模型原理可简述如下,首先将采集到的原始审计数据进行预处理,目的是过滤无用的数据和将原始数据转换为孤立点数据挖掘算法能识别的格式; 然后从孤立点算法库中选择相应的算法对转换后的数据执行孤立点数据挖掘,并得到隐含孤立点; 再将这些隐含孤立点与孤立点规则库中的模式进行比较,把数据分成正常数据和可疑数据; 通过孤立点挖掘发现的可疑被审计数据应结合传统审计专业判断进行下一步确认,再将最终结果用于模型有效性与准确性的验证及审计决策。

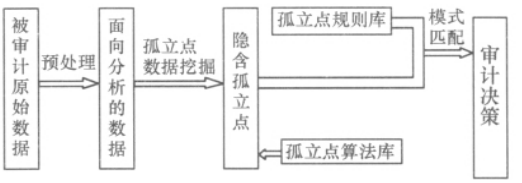

图 1 基于孤立点分析的审计数据挖掘模型

(二) 基于孤立点分析的审计信息化流程

1、审计数据采集。审计数据采集是指将符合审计要求的电子数据从本审单位的会计信息系统中提取出来,并提供给分析模型使用,它是基于孤立点检测审计应用的一个重要初始环节,它为模型提供最原始的数据,数据的好坏也将直接影响到审计的成功与否。审计数据的采集一般要满足两个条件: 一是采集的数据应符合审计模型的要求; 二是采集审计数据时应充分了解被审计单位的信息系统及业务流程。其主要方式有: 联网被审单位报送电子数据; 直接上门采集被审单位信息系统数据; 使用审计数据采集前置机实时采集被审单位电子数据。

2、审计数据存储。由于被审单位经济业务的快速增长,审计部门和人员面对两个重要的任务: 一是审计数据的整理和统一格式转化; 二是采集到的审计数据如何存储。对于采集到的审计数据从存储容量可以分为海量数据和小型数据,海量数据可采用数据仓库技术进行存储,而小型数据可采用传统的数据库存储技术。审计数据存储的模式可分为两种: 集中存储模式和分布式存储模式[5]。集中式存储模式有数据统一存储,数据集中管理,可扩展性好,数据存储量大,数据库间互操作方便和数据存储稳定及安全等优点。分布式存储模式有投入资金低,软件部署灵活和软件升级方便等优点。实际审计信息化工作中,审计数据应采用集中和分布式存储相结合的模式。

3、审计数据分析

(1) 基于偏离的孤立点检测技术算法及应用

序列异常技术(sequential exceptiom technique) 是基于偏离的孤立点检测方法中的一个重要算法及方法,它模仿了人类从一系列推测类似的对象中识别异常对象的方式[6]。它利用隐含的数据冗余。给定 n 个对象的集合 S,它建立一个子集合的序列,$\{S_1, S_2 \cdots, S_m\}$,这里 $2 \leqslant m \leqslant n$,满足 $S_{j-1} \subset S_j$,其中 $S_j \subseteq S$,序列中子集合间的相异度被估算。这个技术引入了下列的关键术语: 基数函数(cardinality functiom): 这一般是给定的集合中对象的数目; 平滑因子(smoothing factor): 这是一个为序列中的每个子集计算的函数,它估算从原始的对象集合中移走子集合可以带来的相异度的降低程度,该值

由集合的势依比例，平滑因子值最大的子集是异常集；异常集(exceptiom set)：它是偏离或孤立点的集合，被定义为类对象的最小子集，这些对象的去除会产生剩余集合的相异度的最大减少；相异度函数(dissimilarity function)：该函数不要求对象之间的度量距离，它可以是满足如下条件的任意函数：当给定一组对象时，如果对象间相似，返回值诺不较小。对象间的相异度越大，函数返回的值就越大。一个子集的相异度是根据序列中先于它的子集增量计算的。给定一个 n 个对象的子集合{ X_1，…，X_n}，可能的一个相异函数是集合中对象的方差：

$$\frac{1}{n}\sum_{i=1}^{n}(X_j - \bar{x})^2$$

这里$\bar{x}$是集合中 n 个数的平均值。对于字符串，相异度函数可能是模式串的形式(例如包含通配符)，它可以用来覆盖目前所见的所有模式。当覆盖 S_{j-1} 中所有字符串的模式不能覆盖在 S_j 中却不在 S_{j-1} 中的任一字符串时，相异度增加[4]。

例如，某单位今年的小车使用相关费用，如小车保险费、年检费、汽油费和小车保养维修费远远高于集合中的平均数$\bar{x}$，模型将被初步确认为该对象为孤立点，并供审计人员做常规审计方法处理。如通过盘存法对企业账面资产进行查实是否该单位私立“小金库”购买账外小车用于领导消费。再比如某公司总经理的工资远远高于集合中的平均数$\bar{x}$，成为一个固有的数据变异孤立点，我们就把该对象排除在外。

(2) 结合孤立点检测结果和常规审计方法判断可疑孤立点

基于“偏离”孤立点检测技术算法对采集到的电子数据进行初步检测后，得到的数据是否是真正有效的可疑的孤立点数据，还需由审计人员采用常规的审计方法对这些数据进行进一步确认，排除那些因固有的数据变异的孤立点。常用的常规审计方法有盘存法：对可疑孤立点相关的各种财产物资进行计量和清点，从而确定该孤立点是否为真正的进一步审计对象；审阅法：对可疑孤立点相关的各种书面资料进行审阅，从而确定该孤立点是不是真正的问题数据；复算法：对检测出的可能孤立点进行重新计算，可最终确认该孤立点是不是真正的待审计数据；函证法：对检测出的孤立点通过函件邮寄给相关单位和人员，以证实该孤立点是不是真正的问题数据。这些方法可进一步确定可疑的孤立点是否存在错弊，也为审计人员确定审计方向及重点、提出审计结论和作出审计报告提供了详细的、有效的数据支撑。

基于孤立点数据挖掘技术的审计判断方法，不但帮助审计人员及时发现异常的交易或事项，缩小了审查的数据量，提高了审计工作效率，而且具有较强的通用性，对行业知识依赖较少和易发现数据的隐藏信息等优点。通用软件开发商可混合多种数据挖掘技术，开发和集成相关的审计“智能化”功能，提高审计部门和审计人员的工作效率。

[参考文献]

[1] 吕新民，王学荣. 数据挖掘在审计数据分析中的应用研究[J]. 审计与经济研究，2007(6)： 35-37.

[2] 陈丹萍. 数据挖掘技术在现代审计中的运用研究[J]. 南京审计学院学报，2009(2)：57-61.

[3] RamaswamyS，RastogiR，Shmi K. Efficientalgorithms for mining outliers from large data set[M] // In: ChenW，Naughton JF，Bernstein PA，eds. Proceedings of the 2000ACM SIGMOD InternationalConference onManagement ofData. Madison: ACM Press，2000: 427-438.

[4] Jiawei Han，Micheline Kamber. 数据挖掘概念与技术[M]. 机械工业出版社，2004：258.

[5] 辛金国，柯　芳. 基于数据挖掘技术的联网审计探索[J]. 新会计，2010(6)：63-65.

[6] 辛金国. 数据挖掘技术在经济统计中的应用探索[J]. 统计与决策，2009(9)：24-26.

Application of auditing Informationization with the Outlier Data Mining Technology

ZHOU Xi，ZENG Li

(*Hunan Business College*， *Changsha* 410205，*China*)

Abstract: Outlier analysis technology is an important problem in data mining，which can be widely applied to telecommunications，credit card fraud detection，loan approval，customer classification，weather forecast and network intrusion detection and other fields. In the audit work，outlier detection algorithm can also be wed for audit data judgment and detecting the abnormal presence，helping auditors find hidden audit trail and improve audit efficiency. Outlier data mining technology has better practical value than discovering regularities in the other mining technology.

Key words: auditing informationization; outlier analysis; data mining