基于审计数据质量控制的数据挖掘应用

陈爱林　黄淑燕

（九江学院　江西九江　332005）

【摘要】逻辑结构与信息本身相分离的特性，使得电子审计数据在真实性、完整性、一致性和有效性等方面难以满足审计工作对数据质量的要求。而利用分类、聚类、关联规则等数据挖掘技术可以控制和提升审计数据质量，提高电子审计效率。

【关键词】数据质量控制　数据挖掘　关联规则　聚类

一、应用数据挖掘技术可以提升审计数据质量

数据挖掘，也称为数据库中的知识发现，它可以从大量冗余的、不完全的、模糊的和随机的数据中提取尽可能多的、事先不为人知的但又是潜在有用的隐藏信息和知识。数据挖掘是一种特定的数据分析过程，它通过对数据进行统计、分析、综合和推理，以发现更多的知识和信息，既可以对已有的事实进行评估，又可以对未来的活动进行预测，从而为做出正确的判断提供基础。

数据质量具体表现在数据的真实性、完整性、一致性和有效性等几个方面。电子审计数据是以电子形式存在的可为审计使用的知识和信息。这种以电子形式存在的数据，由于逻辑结构与信息本身相分离，使得其在许多方面都有别于传统的审计数据，如数据来源更加难以确定、信息的变化更加难以把握等。从数据质量方面来考量，电子审计数据的无形性和易篡改性的确给审计工作带来了一些特定的风险。运用一定的技术来控制和提升审计数据质量变得异常迫切，数据挖掘技术的应用恰好可以满足这一要求。

审计信息化的发展对我国电子审计技术方法和质量管理都提出了更高的要求。数据挖掘技术的应用适应审计对象信息化的发展形势，它可以从庞大的数据库系统中提取更多有用的审计信息，以控制和提高审计数据质量，保证和提升审计数据的及时性、真实性、正确性和完整性，增强电子审计证据的证明力，从而提高电子审计的质量和效率。下面主要从数据挖掘的一般分析方法及其在审计数据质量控制中的应用展开分析。

二、数据挖掘的一般分析方法

数据挖掘涉及机器学习、模式识别、智能数据库、数据可视化、专家系统等技术，在许多需要处理大量数据的领域得到了广泛应用，有的领域应用得非常成功。数据挖掘技术的关键在于挖掘算法，数据类型不同，挖掘算法也各异，发展相对成熟的挖掘算法主要有分类、聚类、估值、关联规则以及描述和可视化等几种。运用这几种算法，数据挖掘系统可以发现和提供典型的知识和信息，以帮助日常工作和决策。

1. 分类。数据分类是指根据数据一定的特性建立相应的分类模型，按照分类模型对数据库的各个对象进行分类。构建上述的分类模型，需要运用一定的统计方法从数据库中选出已经分好类的样本数据库作为训练集，在该训练集上运用数据挖掘分类的技术建立分类模型，对于没有分类的数据进行分类。例如对银行的信贷业务进行审计时可将各种业务分类为低、中、高风险三类，然后将各笔业务分配到预先定义的业务分片。分类就是要达到“物以类聚”的目的，分类规则一旦确立，各种数据都可自动通过数据挖掘系统来归类聚集。

2. 聚类。聚类通常是数据挖掘的第一步。有别于分类分析的是，聚类分析面对的是一组未明确分类的数据，它的任务是把这些数据按相似特征归成若干类，基本要求是属于同一个类别的数据之间的相似性尽可能大，而不同类别数据之间的相似性尽可能小，从而发现数据的分布模式和数据属性间的关系。聚类分析可以采用的技术方法有统计方法、人工神经网络方法、模糊技术等。例如，企业财务数据的变化反映的是企业经营业务的变化状况，如果财务数据的变动存在着偏离企业经营业务变化的异常情况，表明这些数据很可能存在某些虚假成分，其中很可能隐藏了审计需要的重要信息。以对应收账款、应付账款和摊销的实质性测试为例，运用聚类技术可将具有相似性的会计数据进行聚类分组，从中可以发现金额明显异于其他月份或其他时期的账目，这些异常构成了审计的重点领域。从该例可以看出，数据挖掘技术可以明显提升审计数据的真实性和一致性。

3. 关联规则。各事件之间总存在着一定的相互联系，关联规则分析总结了一组事件与其他事件之间的这种联系。通过关联规则分析能寻找到数据库中大量数据的相关性，以概率形式描述甲事件和乙事件在多大程度上会同时出现或先后出现。关联规则分析常用的两种技术为关联规则和序列模式。关联规则是分析一个事件与其他事件之间的相互关联性，序列模式重点分析事件之间的前后因果关系。我们知道，会计科目之间具有很强的相关性，存在着严格的数据勾稽关系。审计人员可通过关联规则挖掘技术对审计对象数据库中的数据进

行分析,找出数据库中各数据之间的相互联系,发现某些数据之间的异常联系,以此为基础,寻找审计线索,发现审计疑点。例如,利用关联规则分析,可以发现一个企业的原材料消耗量、职工工资总额、生产量、销售费用、销售额和应纳增值税额或消费税额的关联性,通过查找相关企业这些数据的对应关系,或许能发现该企业在缴纳增值税或消费税方面存在的问题。

4. 估值。通过估值,可以测算出一些连续性变量的值。例如,根据个人或家庭的购买模式,可以估计个人或家庭的收入水平;通过与个人或家庭的正常收入水平相比,或许能找出个人或家庭收入方面的一些问题。对某个企业或单位,也可以按此逻辑来分析其收入或支出等方面数据的正常性。一般情况下,估值可以作为分类的前期工作,输入一些特定的数据,通过估值分析,得到其他难以直接获取的变量的值,然后根据预定的分类规则进行分类。例如,对于银行的个人消费信贷业务,就可以运用估值分析给各个客户打分,然后根据一定的分类标准将客户按级别分类。

5. 描述和可视化。描述和可视化是对数据挖掘的结果进行表示的方法,它有利于人们更清晰地了解和进一步分析这些数据。描述是对分析对象的内涵进行表述并概括出它们的相关特性;可视化数据分析技术增强了传统图表的表述和分析功能,可以更清晰地分析数据。

三、数据挖掘在审计数据质量控制中的应用

在会计信息化、电子商务和电子政务日益发展的今天,审计工作的质量和效率在很大程度上取决于对电子审计数据的质量控制。电子审计中的数据挖掘可看做审计部门对审计数据进行准备、分析和评价等的过程,此过程中各步骤的工作内容大体如下:

1. 明确审计的目标和内容要求,确定业务对象。数据挖掘是为一定的业务目的服务的,贯穿于具体的业务工作之中,因此认清数据挖掘的目的和要求是数据挖掘成功的第一步。在该环节,必须根据审计分析需要,明确定义审计问题,并将其转化为数据挖掘问题。不同的审计目的和要求下所要准备的数据和选择的数据挖掘算法不一样,因此其分析方法和分析模型也不一样。

2. 数据准备。了解和确定数据的来源和形式,从被审计数据库中选择适用于数据挖掘的知识和信息,对有关数据进行清理和转换,控制数据质量。数据准备包括数据选择、数据预处理和数据转换三方面的工作。数据选择就是在数据库中提取数据挖掘的目标数据项;数据预处理是对数据进行再加工,以保证数据的完整性、一致性和有效性;数据转换的目的是将数据转换成适用于一定挖掘算法的分析模型,这是数据挖掘成功的基础。

3. 数据挖掘。按照审计的目的和任务要求,根据数据的类型和特点选择合适而有效的数据挖掘分析方法,对上一环节准备的数据进行数据挖掘操作,该工作可由数据挖掘系统自动完成,最终给出数据挖掘的结果。在该环节中,根据不同的数据挖掘方法建立数学分析模型是数据挖掘的核心内容。如要预测企业的盈利能力,由于其影响因素很多,就可以运用人工神经网络方法建立分析模型。还要明确的一点是,数据挖掘不仅是一项应用技术,还要理解为一个技术应用的过程。例如,我们可以利用一定的数据挖掘软件包对信用卡的使用进行持续不断的实时监测,从而可以在大范围内侦察到信用卡持有者和商家在交易中的欺诈行为,并且通过分析还可以获知每一个交易者进行欺诈的可能性。

4. 分析和评估结果。对数据挖掘的结果进行分析和评估,并将其转换成能够最终为被审计部门和被审计单位共同理解和接受的信息和知识。该环节通常会使用到描述和可视化技术。该过程还是一个反馈过程,如对模型进行分析,发现其结果并不令人满意,就可以重新运用数据挖掘工具进行分析、建模,直至结果令人满意为止。

5. 知识的组织和运用。将数据挖掘分析得到的知识集成到审计业务信息系统的组织结构中去,使审计人员能在随后的审计工作中组织和运用这些审计知识,以提高其数据分析能力和业务水平。

主要参考文献

1. 李春宏. 基于数据挖掘技术的财务管理信息系统模型.云南民族大学学报(自然科学版),2006;10

2. 韩金红.应用数据挖掘技术提升财务分析质量.合作经济与科技,2007;1

3. 闫建红.《数据库系统概论》的教学改革与探索.山西广播电视大学学报,2006;15

4. 张英俊,谢斌红,赵红燕.基于关联规则挖掘的数据质量提高方法研究.太原理工大学学报,2008;1