基于大数据挖掘分析的财务报表舞弊审计

吴勇 何长添 方君 张超

一、引言

伴随着云计算、大数据、物联网、区块链、数据分析与可视化、流程自动化以及人工智能等创新技术在会计、审计和财务领域的广泛应用，数字化信息呈爆发性增长，大数据已成为企业获取竞争优势的关键性基础，也是未来企业提高生产率、创新以及进行价值创造的重要源泉。在对大型组织实施审计的过程中，审计人员虽然可以获取客户组织内、外部的大量数据，但同时也易于被这些数据所淹没，因此系统地挖掘和分析大量数据背后公司的行为特征，特别是准确识别客户可能存在的异常行为和舞弊风险就显得尤为重要。此外，在高度自动化的环境下，财务报告使用者对报告时效性的要求越来越高，这就需要对大量自动生成、即时可访问的海量数据实施持续性审计。为此，审计人员亟须使用新的理论方法和技术工具，实现由传统人工审计到大数据分析的转化，从而进一步拓展审计的深度和广度，提高审计工作效率，提升审计质量和价值。

数据挖掘是指运用专业的工具、技术、模型和方法，对大量无序的数据进行采集、加载、分析和集成，以便从海量数据中发现隐含的、有效的、有价值的、可理解的模式、规则和知识，并对结果加以解释，从而为解决相关复杂决策问题提供依据和辅助支持。舞弊侦测是数据挖掘技术在财务报表舞弊审计中的重要应用，然而其在实际运用过程中面临着一系列挑战。一般而言，当发现某一审计客户存在舞弊行为时，审计人员会面临三方面决策：特定客户的审计计划中应涵盖哪些具体类型的舞弊行为（如收入确认、高估资产、少计负债等）？哪些数据来源（如日记账、系统工作日志、电子邮件等）可以为识别各类舞弊提供证据？采取何种数据挖掘技术（如定向技术或非定向技术）才能最有效地找到潜在的舞弊证据？

为了解决上述三方面问题，本研究在系统总结和梳理审计大数据的本质内涵与特征分类的基础上，明晰了审计大数据挖掘分析与传统数据分析的差异，探寻数据挖掘技术在财务报表舞弊审计中的最佳应用模式，建立了数据挖掘技术应用于财务报表舞弊审计中的整合性框架，以便指导审计人员将相关数据挖掘技术高效地应用于具体审计活动。

二、审计大数据的内涵

（一）审计大数据的定义

近年来，学术界和产业界基于不同的视角，对大数据做出了不同的定义。大数据作为组织中一类重要的信息资产，是与固定资产和人力资本类似的生产要素，对经济社会发展具有重要价值。大数据具有规模性（Volume）、高速性（Velocity）、多样性（Variety）和真实性（Veracity）等技术特征（被称为“4V”特征）。基于大数据资源观视角，可以认为大数据是企业的战略性资源，其来源多样、特征复杂，企业如果能够快速有效地进行大数据分析，并通过直观、可视化的方式获得大数据分析背后隐藏的知识和规律，增强管理洞察力和价值发现能力，那么大数据将成为支持企业管理决策的一类重要资源，具有重要的决策价值。但是，如果企业看不懂或不会用大数据，那么其决策有用性的价值将受限。

对于审计工作而言，通过分析、挖掘发现被审计单位大量交易数据背后隐藏的信息，特别是通过对大量交易数据的统计特征分析、分类、聚类和关联特征分析，能够有效识别潜在的异常交易和舞弊特征信息，从而为舞弊风险侦测、审计风险评估等提供有益的决策支持。

（二）审计大数据的分类

大数据环境下，企业在经营管理过程中会产生各类数据信息，大数据的形式和特征极其复杂，不仅表现在其数量规模大、来源广、形态结构多样，还表现在其状态变化和开发方式等具有不确定性。就审计工作而言，可以从数据来源、数据类型和数据获取三个方面对审计大数据进行分类。

就数据来源而言，大数据环境下审计人员能够获得企业内部众多数据资料，例如：ERP系统、财务处理系统、交易处理系统以及客户关系管理系统提供的交易数据，从企业生产制造设备、各类传感器中采集的生产、仓储、运输等生产运营过程中的业务数据，企业内部办公系统中的电子邮件、公文处理和会议档案等数据。同时，审计人员还能从外部网站及社交媒体平台中获取包括对被审计单位的各种分析评论、网络舆情以及分析研究报告。就数据类型而言，审计人员不仅能够获得传统的数值型、文本型数据，还能够获取诸如图像、音频、视频等多种类型的数据。就数据获取而言，审计人员可以从物联网平台、ERP系统、各类传感器、网络平台和社交媒体以及视频监控设备上获取数据。对于上述多种类型、多种来源、多方采集的多源异构的海量数据，需要建立数据分析处理模型，以提取出相关信息、识别潜在关系、建立内在关联，有效识别数据背后隐藏的规律性认识，增

强数据洞察力，从而为相关舞弊识别、风险评估等审计决策提供依据。

（三）审计大数据分析

审计大数据分析是指审计人员为了实现既定的审计目标，通过数据抽取、转换、装载（Extract-Transform-Load，缩写为ETL）程序获取内、外部多种类型的数据，运用大数据分析模型、方法和技术，分析全部交易及不同来源数据背后隐藏的异常情况，有效识别舞弊、错误以及违反内部控制等情形。例如，通过计算数据的平均数、标准差、最大值及最小值等统计参数，有效识别异常交易，或通过数据分类、聚类及关联分析，有效识别数据的特征及其内在关联等，从而为舞弊识别、审计风险评估、审计报告出具等提供有效的决策支持。

传统的信息系统仅仅能够获取及分析企业内部的结构化数据，编制静态报表，基于有限的数据进行有限的分析。在数据具有海量、实时与多元特征的大数据环境下，大数据分析工具已经成熟，其可以快速向下挖掘数据特性，为使用者实时提供适应各种营运变化的解决方案，采取交互式仪表盘的操作，快速洞析数据，且通过可视化分析与展示技术，以更加直观、更易于理解的方式呈现大数据分析的结果。大数据分析与传统数据分析的对比如表1所示。

表1　大数据分析与传统数据分析的对比

项目		传统数据分析	大数据分析
数据层面	数据来源	企业内部资料	企业内部资料、企业外部资料
	数据类型	结构化资料	结构化资料、半结构化资料、非结构化资料
	数据量	Peta Bytes(1015)	Zetta Bytes(1025)
分析层面	解决的问题	描述性分析：发生了什么？诊断性分析：为什么会发生？	描述性分析：发生了什么？诊断性分析：为什么会发生？预测性分析：我们可以做什么？处方式分析：我们应该做什么？我们如何做到最好？
	常见应用	报表统计分析、企业仪表盘	快速分析、行为预测、最优化模拟和仿真、机器学习
	反应方式	通过过去的数据进行反应	预测未来情况，优化决策，提前行动

虽然大数据和数据分析是两个独立的概念，但两者紧密关联，图1就表明了审计领域中两者之间的内在联系。

多年来，会计师事务所习惯于在路径A中使用传统的数据分析工具（如Excel、ACL和Case-WareIDEA）来分析会计数据的样本。美国会计学会对相关从业人员的最新调查结果显示，会计师事务所已开始进入路径B并远离抽样审计方法，数据可视化的审计工具（如Tableau）越来越受到审计人员的欢迎，但审计的重点仍然是传统的会计数据和审计程序，如查找重复的发票。调查结果

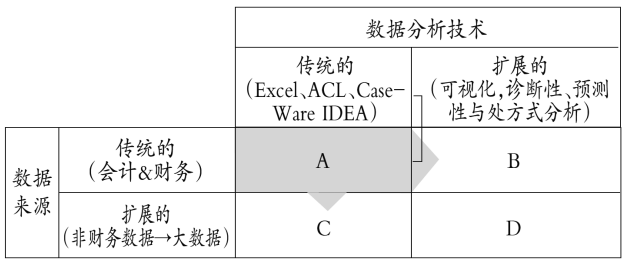

图1　财务报表审计中扩展数据分析的路径

也提及社交媒体分析已经作为审计的一部分，并开始向路径C方向发展，而路径D才是真正地将大数据和高级数据分析工具进行有机融合，在传统审计注重会计与财务数据的描述性分析基础上，引入非财务数据以及外部的宏观经济、制度环境、社交媒体等海量的多源异构大数据，更加关注现状背后的问题与成因分析、未来的趋势预测以及如何针对问题给出优化解决方案和处置策略。

三、数据挖掘分析技术的概念和分类

（一）数据挖掘技术的概念

在阐述数据挖掘的概念内涵时，需要明晰数据抽取与查询、数据分析和数据挖掘三者之间的关系。

1.数据抽取与查询。在财务报表审计中，Excel、ACL和Case-WareIDEA等计算机辅助审计工具和技术（CAATTs）是审计人员检查客户数据时最常用的工具，其内置的多种函数功能等可以进行描述性统计和简单的数据分析，有效地识别客户数据集中的可疑数据模式，而且还能作为进一步审计程序的样本选择工具。例如，如果某公司的内部控制制度规定支付给供应商超过50000元的货款需要财务总监签字，为了防止存在开出低于50000元的舞弊性支票以避免财务总监审查的情形发生，可以利用数据提取工具抽取金额在49000~49999元之间的支票，以确定是否有违反内部控制的舞弊行为存在。

2.数据分析。数据分析工具集为审计人员提供了从简单到复杂的一系列分析技术。例如，简单的数据分析包括基本的描述性统计（如计数、最小值、最大值、平均值和离散度等）以及比率分析，而相对复杂的数据分析包括单变量和多变量回归等推理统计以及相关分析等。数据分析可以进一步分为描述性分析、诊断性分析、预测性分析和处方式分析。

（1）描述性分析（Descriptiveanalysis）。描述性分析主要回答过去发生了什么，通过将过去和当前数据转换为总结性、概括性的报告、图表、数据透视表等形式，帮助审计人员全面、高效地了解被审计单位当前的经营状况和财务业绩。例如，将营业收入增长率与前期数据相比可以帮助管理会计师了解公司的成长能力，与行业基准相比可以看出公司是否保持了竞争优势。除此之外，描

述性分析在顾客、企业、员工层面也有助于管理会计师发挥职能。例如:退货率和保修索赔率可以反映客户对公司新产品的满意程度;研发费用占比可以衡量公司对开发新产品或服务的重视程度;员工技能、生产力等特征可以识别高效率的员工。

(2)诊断性分析(Diagnosticanalysis)。诊断性分析旨在分析为什么会发生,诠释当前结果的原因。例如:相比于同行业的其他企业,为什么企业的经营费用、销售费用和管理费用会增加?为什么平均有效所得税率会变化?为什么应收所得和净收益之间的差异会越来越大等?与同期相比,为什么销售收入会下降?诊断性分析可以进一步细分为两类:识别异常,以及发现两个或多个变量之间未知的连接、模式和关系。

(3)预测性分析(Predictiveanalysis)。预测性分析旨在回答未来可能会发生什么,它利用各种统计、建模、数据挖掘工具对某段时间内累积的历史数据进行研究,计算未来事件发生的可能性,从而对未来进行预测。预测性分析采用的数据大部分是定量的数据,主要的算法有分类分析、回归分析和时间序列分析等。

(4)处方式分析(Prescriptiveanalysis)。处方式分析(或称为规范性分析)是在前述分析的基础上,给出相应问题的解决方案和行动建议,主要回答如何做得更好或该朝哪个方向努力。它是在描述性和预测性分析结果的基础上,通过探寻一个或多个解决方案,并分析每个解决方案的可能结果,给出最优解决方案,以便有效地指导我们如何才能取得更好的结果。具体的分析方法主要包括情景假设分析、单变量分析、边际分析、现金流量分析、敏感性分析,以及机器学习、仿真优化等智能决策和优化算法。

3.数据挖掘。数据挖掘是基于海量数据来揭示、发现有意义的关系、规则、模式或趋势的过程,它融合了人工智能、数据库和数据仓库技术、模式识别、机器学习、统计学、数据可视化和高性能计算等多个领域的理论和技术,是数据库中知识发现的核心步骤。针对财务报表舞弊审计而言,超出正常范围或预测范围的数据挖掘结果能为审计人员提供重要的风险信号。

(二)数据挖掘技术的分类

数据挖掘的模型、工具和技术手段多种多样,按照数据挖掘技术中变量之间关系的不同,可将其分为定向(或自顶向下的方法)和无定向(或自底向上的方法)两大类。其中:定向数据挖掘用于识别感兴趣的特定目标变量,探究该变量与选定的其他变量之间的关系;而无定向数据挖掘没有特定的目标变量(因变量),可用于探寻数据总体中任何变量之间的关系。换言之,定向数据挖掘适用于检验特定的假设,而非定向数据挖掘适用于检验新的假设。

1.定向数据挖掘。定向数据挖掘技术包括:

(1)分类(Classification)。分类的核心目的是把具有某些特征的数据项映射到特定类别上,它通过对带有类标号的训练数据集的学习来建立分类模型,常以分类规则、决策树或数学表达式的形式予以表达,并利用分类模型将新对象准确划分到相对离散的类别中。分类的准确性、鲁棒性以及对分类结果的解释能力是衡量分类质量的重要指标。常用的分类技术包括决策树归纳、贝叶斯信念网络、神经网络、支持向量机和遗传算法等。例如,针对财务报表审计,可以依据财务报表相关信息、股价波动和成交量等变量的特征,将审计客户划分为低风险、中风险和高风险三大类。

(2)估计(Estimation)。估计模式的函数定义与分类模式相似,主要差别在于分类模式采用离散预测值(如类标号),而估计使用的是连续的预测值。此种观点下分类和估计都是预测问题,但数据挖掘界普遍认为:用预测法预测类标号为分类,预测连续值(如使用回归方法)则为估计。例如,针对财务报表审计而言,估计技术不是将客户风险划分为低、中、高三类,而是给出一个风险评分(例如0~10),估计值适用于对象的总体排序,并生成阈值得分,随后可以通过类似于逻辑回归的技术实现对象分类。例如,将风险评分超过7.5分的客户界定为高风险客户,对其实施更加严格的审计程序。

(3)预测(Prediction)。“描述”和“预测”是数据挖掘的两个重要目标,描述性数据挖掘的任务主要是找到描述数据的可理解模式,以便更好地刻画目标数据中数据的一般性质,例如分类和估计等数据挖掘技术通常用于揭示数据集中先前确定的变量的特征。而预测性挖掘的任务是对当前数据进行归纳以便做出预测。因此,数据挖掘领域的学者经常将预测技术与分类和估计技术区分开来。预测技术主要是利用历史数据建立模型来找出变化规律,并用此模型预测数据集中其他感兴趣的变量或字段的可能值或一个数据集合中某种属性值的分布情况。与分类技术类似,预测技术也使用训练集来构建初始模型。在财务报表审计中,若找出某一因素超出预测范围的异常值,则能为识别审计疑点和确定重点审计领域和方向提供重要支持。

2.非定向数据挖掘。非定向数据挖掘技术的典型应用主要包括:

(1)关联规则(Affinitygrouping)。关联规则是指在没有特定的因变量和自变量的情形下,寻找数据集中相关变量之间的关联关系和相关关系。最著名的关联规则挖掘算法是Apriori算法,该算法的基本思想是:统计多种商品在一次购买中共同出现的频数,然后将出现频数

多的搭配转换为关联规则。在财务报表审计应用中，审计人员可以利用数据挖掘技术分析收入与成本费用以及资产、厂房、设备和工厂维护费等分类账之间的关联规则，那些不在预期关联分组之内的分类账将是审查的重点。此外，关联规则分析还能应用于遵循特定序列（如时间模式）的数据上。例如，银行针对洗黑钱的审查，可以根据相关事件发生的顺序查找潜在的关系，有助于识别出一系列不正常的、但金额相对较小的账号间转账的舞弊行为。

（2）聚类（Clustering）。聚类是数据挖掘中用来发现数据分布和隐含模式的一项重要技术，其核心目的是根据数据集中的变量关系将数据项聚为多个子类或簇。按照“最小化类间的相似性、最大化类内的相似性”原则，使得类内的数据差异尽可能小，类间的数据差异尽可能大。聚类技术包括层次方法、基于神经网络的自组织映射和基于密度的技术。与分类模式不同的是，聚类中要划分的类别是未知的，它是一种不依赖于预先定义的类和带类标号的训练数据集的无监督学习（Unsupervisedlearning），无须背景知识，其中类的数量由系统按照某种性能指标自动确定。例如，在一组分类账集合中，审计人员可以按照销售收入、应收账款、产品成本和库存聚类为几个子类。

（3）描述和可视化（Description&Visualization）。数据挖掘的重要目标是在数据库中发现“有趣的、有价值的”知识，并利用可视化技术将其以图形、图像等易于理解的方式予以呈现，以便用户更加直观、清晰地理解数据挖掘的结果。当然，对数据挖掘结果潜在含义的解释有赖于审计人员所掌握的专业知识以及对客户业务模型的深刻理解。

（三）不同数据分析工具之间的内在关系

图2列示了数据抽取与查询、数据分析和数据挖掘三个概念之间的关系。

随着审计人员对审计技术工具的应用从数据抽取与查询转向数据分析和数据挖掘，软件在功能方面变得更加复杂，诊断性分析和预测性分析能力也更强。通过系统的文献梳理可以发现，审计人员使用以上三种技术的频率存在差异，数据抽取和查询工具的使用频率很高，而数据挖掘工具的复杂性及对审计相关知识的要求更高，使其应用受到制约。

四、基于全生命周期的大数据挖掘技术在审计中的应用

注册会计师发表的无保留审计意见是对财务报表不存在由错误或舞弊引起的重大错报的合理保证。针对财务报表舞弊审计，美国注册会计师协会（AICPA）发布的第99号审计准则《财务报表审计中对舞弊的考虑》

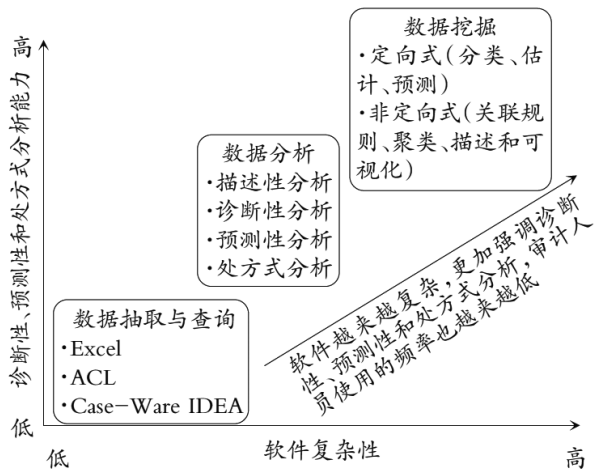

图2 不同数据分析工具之间的内在关系

（SAS第99号）将舞弊分为财务报表舞弊和资产侵占舞弊两类，虽然后一类舞弊发生的频率可能更高、涉及面更广，但其重要性并不如财务报告舞弊。SAS第99号将财务报表舞弊定义为“在财务报表中故意错报或遗漏金额，旨在欺骗财务报表使用者，导致财务报表在所有重大方面均无法按一般公认会计原则（GAAP）公允列报”。审计准则体系要求审计人员制订周密的审计计划、执行严格的审计分析以及实质性审计程序，以便检查财务报表中出现的错误，按照“审计计划—审计实施—审计报告”的审计业务全生命周期，审计各个阶段的步骤可列示为图3。

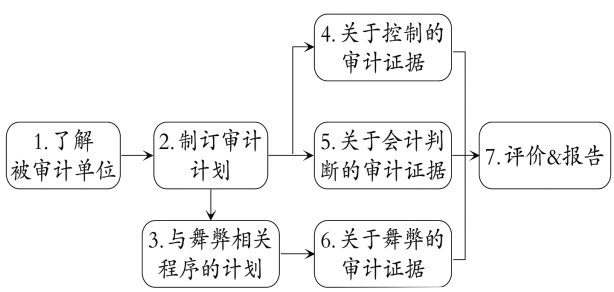

图3 审计业务全生命周期

本文基于审计业务全生命周期各阶段的地位、目的和性质，探讨数据挖掘技术在各阶段可能发挥的潜在作用。

（一）审计计划阶段

审计计划阶段应聚焦于企业利益相关者关注的领域，并衡量使用大数据分析技术、方法的可行性，同时识别企业各作业流程所存在的风险，制订合理的审计计划。

1.了解被审计单位。无论被审计单位财务报表是否出现错误或舞弊，对重大错报风险的评估都是审计计划

阶段的重要工作，而了解被审计单位是审计计划早期阶段的关键内容。审计人员必须了解各种与客户相关的风险因素，包括所有权结构和组织架构、价值创造过程的本质、与业务合作伙伴和关联方的关系，以及客户运营的监管环境。鉴于被审计单位在业务流程、组织结构以及与业务合作伙伴关系方面的复杂性，审计人员有可能利用数据挖掘工具和技术提高对被审计单位绩效、外部关系和网络数据的分析能力。通常，在预审计环节，可以利用数据挖掘技术对被审计单位的外部数据（如公布的财务报表、新闻报道、分析报告、股价、监管文件等）以及内部的财务数据等进行多维度分析，并选择行业数据进行对标分析，以有效识别风险。

2.制订审计计划。被审计单位评估的风险越高，越需要制订更详细的基于风险导向的审计计划。作为审计计划的一部分，审计人员通常会采取各种分析性程序来制订预期的账户余额，如债务水平、现金水平和应计水平。审计人员还会对财务报表核心数据进行比率分析，以得到预期值。同时，国家和行业的产出、收入和盈利水平是一个重要的比较基准，将国家和行业的外部数据分析与被审计单位内部数据相结合，利用内、外部数据集的有效组合进行数据挖掘，有助于制订更加有效的审计计划，提升审计效率和质量。

在基于大数据审计的过程中，审计计划阶段最具挑战性的工作是采集与准备所需的数据资料，包括向企业信息部门提出具体的数据需求以及从外部获取相关审计大数据。如何确保所取得数据的准确性和完整性，如何对多种来源和类型的审计大数据进行数据预处理等都会影响审计工作的有效性。

3.与舞弊相关程序的计划。舞弊风险作为审计计划的一部分，SAS第99号要求审计人员系统地解决因舞弊而产生重大错报的风险。审计人员必须以谨慎的方式（包括审计团队进行积极的头脑风暴）讨论审计计划中的财务报表舞弊风险，完成各种分析性程序。例如，针对可能存在舞弊的收入报告，在审计计划阶段，审计人员应执行与收入相关的分析性程序，如将收入、销售量与企业的生产能力进行比较，若销售量超过产能，则表明可能存在虚假销售。此外，在报告期间和报告结束后一段时间内，对月销售收入的趋势分析有助于识别出可能存在与客户签订未公开的合同，以及退回已确认收入的商品等收入账户舞弊的异常情形，这种基于时间维度的数据挖掘分析，为大数据环境下审计风险评估提供了新的有效思路。例如，不是对已知指标（如库存周转率和产品盈利能力）直接进行比率分析，而是通过对多个指标的数据挖掘开发更复杂的客户增值流程和风险概况模型，有效识别潜在的舞弊风险。

此外，审计计划阶段的另一个重要内容是考虑内部控制缺陷所带来的风险，要求审计人员评估内部控制的设计和执行情况。被审计单位通常会准备大量的会计流程文件来证明其内部控制设计和执行的有效性，审计人员可以对这些会计流程进行系统的数据分析，以评估风险和后续测试的关键控制点。

（二）审计实施阶段

在审计实施过程中，需要利用所获取的审计大数据进行审计测试，传统审计抽样及抽查等技术方法已无法有效发现错误及异常交易，审计人员需要利用大数据分析技术，基于特定审计目标，开发出相应审计测试的大数据分析处理模型和方法来分析所有交易数据，以便发现异常情形，帮助审计人员聚焦高风险审计范围和领域。

1.关于控制的审计证据。审计实施阶段的重点工作是通过符合性测试和实质性测试，收集充分、适当的审计证据。符合性测试旨在评估内部控制设计和运行的有效性，对控制过程进行过程挖掘是数据挖掘在此阶段的重要应用。过程挖掘通常在ERP系统的系统日志上运行，对控制过程进行过程挖掘是数据挖掘技术应用于内部控制测试收集审计证据的一种重要方法。

2.关于会计判断的审计证据。随着“轻资产”类型公司的大量涌现，无形资产在企业资产结构中的比重越来越高，与此相关的会计要素的确认、计量、记录和报告的过程等涉及复杂的价值评估和判断，科学合理利用大数据资源能够为相关价值判断提供更好的审计证据。

3.关于舞弊的审计证据。现场审计中，审计人员会开展大量的实质性测试，其中的一些程序涉及舞弊审计。例如，关于收入的实质性分析程序可能会利用分类数据，如按月比较产品线或业务部门报告的收入。实质性测试作为舞弊相关分析程序的一部分，对公司绩效数据、收入数据库以及其他客户数据库的数据挖掘将具有重要的潜在意义。

（三）审计报告阶段

审计人员应评估“在现场工作完成时或接近完成时”因舞弊行为导致重大错报的可能性，并做出适当回应。回应可能包括重新评估在审计计划阶段被拒绝使用的数据挖掘程序，或在审计实施阶段观察到的异常情况而开展的新数据挖掘程序。例如，实质性分析程序或其他收入测试可能表明需要数据挖掘。之后，对审计证据进行最终审查，从而决定审计报告最终的性质。

综上，本文构建的基于大数据挖掘分析的财务报表舞弊审计应用框架如图4所示。

针对上述审计计划、数据准备、审计实施过程的大数据分析程序及结果，应加强审计复核和质量监控，以确保基于大数据分析结果的可靠性。同时，充分利用可

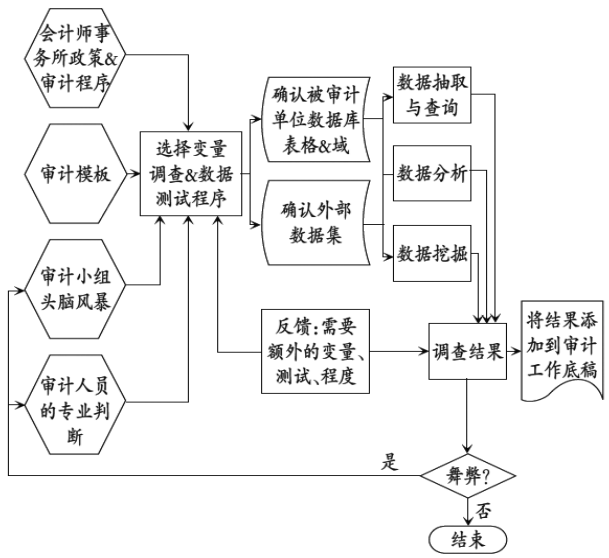

图4 基于大数据挖掘分析的财务报表舞弊审计应用框架

视化分析和仪表盘展示平台，构建风险预警机制，增强洞察力，提高审计效率和效果。

五、面向审计全生命周期的各类大数据挖掘技术的具体应用

伴随着数据量的爆炸式增长以及数据挖掘技术和工具的出现，会计师事务所亟须从客户的海量数据中发现更有价值的信息和知识，以便为制订审计计划、评估审计风险和内部控制的有效性、确定重点审计领域和方向、识别舞弊行为等提供有益的支持。相关研究表明，数据挖掘技术已成功应用于财务报表虚假信息披露、日记分类账挖掘分析、公司内部业务流程分析和角色建模分析、公司内外部文本信息的挖掘分析以及其他舞弊欺诈检测模型等。

对财务报告披露信息进行挖掘分析以便识别出可能存在舞弊的公司报告，是数据挖掘在财务报表审计领域的典型应用。Ravisankar等利用支持向量机、神经网络和遗传算法等数据挖掘技术，分析财务报表的重要数据（如净利润和毛利润）和核心比率（如长期债务/总资产），识别出虚假报告。Alden等利用进化遗传算法（EvolutionaryAlgorithms，EAs）识别财务报表中的舞弊行为。Perols的研究表明，逻辑回归和支持向量机（SVM）分类技术在财务报表舞弊识别中有很好的表现。Glancy和Yadav建立了一种新型的欺诈检测模型，能够生成具有典型数据挖掘特征的奇异值分解向量（SVD），用于分析财务报表附注中的欺诈行为。

现代会计信息系统的基础业务流程特性决定了日记分类账能够完整地列示业务的所有方面，因此，目前审计人员对大规模数据集进行最多分析的领域是客户的日记分类账。Debreceny和Gray将数据挖掘技术应用于客户日记分类账分析，挖掘了29套分类账。Argyrou使用自组织映射（SelfOrganizingMap，SOM）识别单个公司的一组会计分录中存在的可疑交易。

近年来，数据挖掘技术在审计领域的另一个重要应用是过程挖掘。该技术从企业信息系统（通常是ERP系统的工作流和角色分析）生成的事件日志中提取业务流程知识，对企业决策过程、合规与风险管理和控制结构等进行过程挖掘。了解用户在企业系统中的作用与角色是对企业内部控制和具体交易进行审计的前提，在有关角色挖掘的研究中，Colantonio等建立了角色建模方法，并成功应用于某大型企业员工舞弊行为侦测。针对员工舞弊问题，Jans等采用单变量和多变量聚类技术用于识别潜在的舞弊交易。

大数据环境下，被审计单位充斥着大量的内、外部文字信息，文本挖掘技术在审计中的应用日益受到关注。电子邮件是企业内部员工之间以及企业与外部供应商、客户等沟通的重要工具。而且电子邮件也为相关法律诉讼以及监管要求提供了重要的证据信息，组织也会更加重视保存结构良好的电子邮件信息。电子邮件涵盖日期、收件人和主题字段等信息，此种半结构化特性为电子邮件的数据挖掘添加了重要的时间和社交网络维度。Debreceny和Gray将文本挖掘和社交网络分析成功应用于电子邮件数据挖掘。Humpherys等建立了一个涵盖惯用语言、情感、单词长度和句子复杂度等特征的财务报表欺诈识别模型，实验数据结果显示模型对财务报表欺诈行为识别的准确率达到67%。

大数据环境下，众多公司利用商业分析（Businessanalytics）来服务于其战略决策和运营管理。审计人员在审计过程中，也应该拓展传统的审计分析方法，更多地采用商业分析工具来提升审计工作效率以及审计工作质量。Appelbaum等、Vasarhelyi等构建了数据挖掘分析技术在审计业务全流程的应用框架，系统总结和梳理了审计过程中应该运用哪些商业分析技术方法，以及审计过程的不同阶段应该选用哪些合适的商业分析技术方法，具体如表2所示。

在审计计划阶段，签订审计业务约定书时，审计人员可以查阅经审计的财务报表、其他公共信息以及其他外部数据来源，并从定量和定性数据中得出预期模型。在此过程中，常用的大数据挖掘分析技术主要包括：比率分析、文本挖掘、可视化、回归模型和描述性统计等。在制定计划与风险评估的过程中，审计人员可以在获取未经审计财务报表的基础上，构建可能发生和应该发生的风险预测模型。此时除了比率分析，还可以使用聚类、可视

表2 基于审计业务全流程的大数据挖掘分析技术应用框架

大数据挖掘分析技术		审计计划阶段		审计实施阶段	审计报告阶段	
		审计业务约定书	计划与风险评估	审计测试	审计复核	审计意见
描述性分析	聚类模型		√	√	√	√
	描述性统计	√	√	√		
	过程挖掘:过程发现模型		√	√	√	√
	比率分析	√	√			
	斯皮尔曼等级相关性测试		√	√	√	
	文本挖掘模型	√		√	√	√
	可视化模型	√	√	√	√	√
预测性分析	层次分析法(AHP)	√	√		√	√
	人工神经网络(ANN)	√	√	√	√	√
	自回归积分滑动平均模型(ARIMA)					√
	Bagging and Boosting模型	√	√		√	√
	贝叶斯理论/贝叶斯信度网络(BBN)	√	√	√	√	√
	本福德定律	√	√	√	√	√
	C4.5统计分类		√	√	√	√
	证据推理模型	√	√	√	√	√
	专家系统/决策支持	√	√	√	√	√
	遗传算法	√	√		√	√
	假设检验	√	√	√		√
	线性回归	√	√			
	Logistic回归		√		√	
	蒙特卡洛模拟	√	√	√	√	√
	多准则决策辅助				√	
	可行性分析模型	√				√
	过程挖掘:过程优化	√	√	√	√	√
	结构模型			√		√
	支持向量机(SVM)	√	√	√	√	√
	时间序列回归					√
	单变量和多变量回归分析					√
处方式分析	人工神经网络(ANN)	√	√	√	√	√
	自回归积分滑动平均模型(ARIMA)	√	√	√	√	√
	专家系统/决策支持	√	√	√	√	√
	遗传算法	√	√	√	√	√
	线性回归	√	√	√	√	√
	Logistic回归		√	√	√	√
	蒙特卡洛模拟	√	√	√	√	√
	时间序列回归	√	√	√	√	√
	单变量和多变量回归分析	√	√	√	√	√

化、回归模型、信度网络、专家系统和描述性统计等。

审计实施阶段的重要工作是执行审计测试,可以考虑根据客户的环境,选择将抽样测试与全样本测试的结果进行比较,并根据基准和预期结果的差异,聚焦确定符合性测试和实质性测试的范围和重点方向,如果测试结果显示有问题或预示着需要进一步调查,可能需要进一步测试和收集证据。该阶段除了会用到所有目前常用的审计检查技术,还会使用聚类、文本挖掘、过程挖掘、可视化、支持向量机、人工神经网络、专家系统、信度网络、回归模型、本福德定律、描述性统计、结构模型以及假设检验。

在审计报告阶段,进行审计复核时,需要使用不同的技术对异常结果进行交叉验证测试和分析,前述实质性测试阶段用到的所有技术方法都适用。在此过程中,还会用到专家系统、概率模型、信度网络、支持向量机、人工神经网络、遗传算法等。随着2016年审计报告准则的颁布实施,如何提高审计过程透明度、增加审计意见和审计报告的信息含量和价值有用性显得尤为重要。在确定审计意见时,主要使用的大数据挖掘分析方法包括时间序列回归、概率模型、信度网络、专家系统和蒙特卡洛模拟等。此外,如何应用数据挖掘分析技术以得到更细致、定量化的审计意见,以及拓展目前无保留审计意见和非无保留审计意见的二分法结果,也是值得探究的重要领域。

六、结论

基于大数据的挖掘分析已经在财务危机预警、财务欺诈检测、股票市场预测和量化投资决策等会计与财务领域有着广泛的应用,然而其在审计领域的应用稍显滞后。基于客户内、外部海量大数据的深度挖掘分析,为审计功能的发挥提供了一个补充的证据来源,而且审计人员能够分析客户公司财务报告相关数据的生成过程,洞察被审计单位经营管理过程中的各类问题,有助于使审计行业向价值链上游移动,真正成为客户公司的商业伙伴。基于海量数据的挖掘分析将改变审计人员的工作方式,同时,海量数据涉及的诸如信息超载、信息相关性、可靠性、模糊性等问题也会影响审计人员的职业判断,这里尤为关键的是面向具体审计领域和工作内容选择最合适的数据挖掘技术工具。

本研究通过系统的文献梳理,旨在探寻数据挖掘技术工具在财务报表舞弊审计中的最佳应用模式,构建数据挖掘技术应用于财务报表舞弊审计的整合性框架,以便指导审计人员将相关数据挖掘技术高效地应用于具体审计活动中。未来的研究中,一方面,基于大数据挖掘分析的审计范式变革,会改变审计人员的行为方式、工作流程和执业风险,而相关会计和审计准则并没有跟上技术变革的步伐;另一方面,大数据挖掘分析为审计人员提供了补充的证据来源,然而由于海量数据的多源异构特征,使得在数据质量、可靠性等方面存在巨大差异,如何依据充分性、可靠性和相关性的审计证据标准框架对其适用性进行科学评估值得探究。

基金项目:国家自然科学基金项目(项目编号:J1924003);安徽省教育厅教学研究省级重点项目(项目编号:2017jyxm0040)。

作者单位:合肥工业大学管理学院

原载《财会月刊》(武汉),2021.3.90~98