DOI：10.19699/j.cnki.issn2096-0298.2020.20.055

基于数据挖掘的金融审计数据分析研究

长春财经学院 赵浏洋

摘 要：针对由于初始变量数据过多，导致金融审计数据挖掘精度不足的问题，进行基于数据挖掘的金融审计数据分析研究。本文首先从被审计的金融机构信息系统中提取所需数据，并对其进行预处理，包括缺失值处理、重复数据处理、噪声数据处理、数据变换等，其次利用主成分分析方法解决初始变量数据过多问题，降低数据维度，最后选取聚类算法作为挖掘方法，实现金融审计异常数据分析。结果表明：与神经网络算法、支持向量机、最近邻算法相比，本方法精度更高，以期为后续研究提供参考。

关键词：数据挖掘；金融审计数据；聚类算法；数据维度

中图分类号：F239.65　　**文献标识码**：A　　**文章编号**：2096-0298(2020)10(b)-055-03

审计机关是推动完善公共治理的重要力量，它可以通过提供以证据为基础的解决系统性问题的意见及建议，促进、完善政策和方案，发挥审计的监督、洞察和前瞻功能，推动改善公共治理[1]。近年来，审计机关通过开展金融审计工作，有效地发挥了监督、洞察与前瞻功能，在促进防范和化解金融风险，提高金融服务实体经济质量和效益，完善金融监管体制，推动金融体制改革和金融领域反腐倡廉等方面发挥了重要作用。然而，在审计机构进行金融审计时，面临海量的数据，仅仅依靠传统的数据检索机制和方法是远远不够的，存在审计效率低下等问题。

在上述背景下，数据挖掘技术应用成为审计领域研究的重点课题。利用数据挖掘技术寻找数据间潜在的关联，关键在于挖掘算法的选择上。常用的挖掘算法有很多，如Desaietal利用神经网络分类挖掘算法对3000个观察数据进行分析；孙北伐、张高煜、徐倩蓉等在《大数据环境下数据挖掘在审计中的分析与应用》中介绍决策树算法和随机森林算法的数据分析过程。此外，数据挖掘还包括聚类算法、神经网络算法、支持向量机、最近邻算法等[2]。本文在已有研究经验的基础上，以聚类算法为基础，进行数据挖掘研究。研究过程如下：首先进行金融审计数据采集并进行预处理，提高数据质量，然后选取聚类算法作为挖掘算法，设置挖掘条件，进行模式匹配，找寻可疑数据。通过这些可疑数据，审计机构可以以此为依据进行追踪，探寻企业违法犯罪行为，为金融领域反腐倡廉工作提供依据。

1 基于数据挖掘的金融审计数据分析方法

随着计算机网络以及信息技术的不断发展，各行各业对信息系统的依赖程度越来越深，其中金融审计就是其中一个。金融审计就是在海量数据中寻找异常数据，从而发现问题，得出各种金融机构的经营状况，以便避免出现金融风险，揭露隐藏的违法违规行为，推进反腐倡廉工作。现代金融审计人员面对的不再是简单的纸质账目，而是存储在计算机系统数据库中的种类繁多的电子数据，要想通过人工完成数据分析是不现实的，因此本文通过数据挖掘技术进行金融审计数据分析。

1.1 金融审计数据采集

金融审计数据采集进行数据挖掘的第一步，也是后续进行数据分析的基础和前提。金融审计数据采集是指审计人员在利用计算机审计时，需要根据审计要求从被审计的金融机构信息系统中提取数据文件的过程[3]。目前，采集方式主要有三种，即直接读取数据的方式、数据库连接性的方式以及数据传输的方式，三种方式特点比较如表1所示。

表1 三种金融审计数据采集方式特点对比

项目	直接读取数据的方式	数据库连接的方式	数据传输的方式
了解对方数据库结构	Yes	Yes	No
双方数据格式一致	Yes	No	No
直接访问对方数据库	Yes	Yes	No

1.2 金融审计数据预处理

从各个金融机构系统数据采集得到的数据受到人为因素、系统因素等的影响，数据质量并不高，若直接用于后续数据挖掘，将会导致数据分析准确性降低。为此，需要对采集到的数据进行预处理，具体包括缺失值处理、重复数据处理、噪声数据处理、数据变换等[4]。

1.2.1 缺失值处理

采集得到的原始数据极有可能存在缺失值，但是缺失值并不意味数据有错误。缺失值在整个数据集若是一个关键值，就需要进行填补。缺失值填补的方法有很多，如人工填写、平均值填充、最邻近方法填充、期望值最大化方法填充、贝叶斯Boostrap方法填充、回归法填充等。

1.2.2 重复数据处理

在采集到的原始金融审计数据中还存在一些重复记录的数据，这些数据也被称为冗余数据。冗余数据的存在会增加后续数据挖掘和分析的计算量，降低数据分析效率。对于重复数据的处理，需要进行记录排序，即根据关键字、词等进行排序，然后识别重复记录，将重复的数据进行合并。对于重复记录识别，可以通过简单的模糊匹配或各个角度的相似度计算来完成。

作者简介：赵浏洋 (1980-)，男，汉族，黑龙江绥化人，博士研究生，副教授，主要从事互联网金融方面的研究。

1.2.3 噪声数据处理

由于各种原因导致属性值不正确或不一致的数据被称为噪声数据。对于噪声数据，处理方法主要有三种：分箱、回归以及聚类，如表2所示。

表2 三种噪声数据处理方法

项目	原理
分箱	通过考察数据的“近邻”（即周围的值）来光滑有序的数据值
回归	通过用一个函数将数据拟合到一个多维面，来光滑数据
聚类	通过聚类将类似的值组织成群或“簇”，没有在群或“簇”当中的值就是噪声值

1.2.4 数据变换

采集到的原始金融审计数据可能来自被审计金融机构的不同类型的数据库，而不同类型的数据库的数据是不同的，无法进行比较分析，因此需要将不同形式的金融审计数据转换成适合的审计分析软件处理所需要的形式，将数据规范成相对统计的形式，去除量纲，即数据规范化。数据规范化方法主要有以下三种。

(1)Min-Max标准化(Min-Max normalization)

Min-Max标准化，也称为最小-最大规范化，基本原理是对原始金融审计数据进行线性变换，使变换后的结果落到[0,1]区间内。Min-Max标准化表达式如下：

$$x' = \frac{x - \min}{\max - \min} \tag{1}$$

其中，x'为规范化后数据；x为原始数据；$\max$为样本数据的最大值；$\min$为样本数据的最小值。

(2)Z-score标准化(zero-mean normalization)

Z-score标准化，也称为标准差规范化，基本原理是让经过处理的原始金融审计数据符合标准正态分布，即均值为0，标准差为1。Z-score标准化表达式如下：

$$x' = \frac{x - a}{b} \tag{2}$$

其中，a为对应特征均值；b为标准差 。

(3)小数定标标准化(Decimal scaling)

小数定标标准化基本原理是通过移动数据属性值的小数点位置来进行标准化，标准化的结果最终落到[−1,1]区间内。小数点移动多少位取决于属性取值中的最大绝对值。小数定标标准化表达式如下：

$$x' = \frac{x}{10^k} \tag{3}$$

其中，k为满足条件的最小整数。

1.3 金融审计数据降维

以往利用数据挖掘算法进行直接金融审计数据分析时，往往存在计算量大、分析不准确的问题，而导致这一现象的主要原因是后续输入到数据挖掘算法中初始变量数据过多。对于上述问题，将数据降维十分必要。

数据降维是指在保证原始数据损失量最小的前提下，优化数据组成，降低数据维度，减少数据规模。

对于数据降维问题，解决方法有很多，大致分为线性映射和非线性映射方法两大类。在本文中选择主成分分析方法进行金融审计数据降维。具体过程如下：

步骤1：假设待分析原始金融审计数据的形式是一个X包含n个样本的样本集。

步骤2：对样本集中的数据进行标准化处理，处理方法见1.2.4中数据变换处理方法，最后得到的标准化矩阵Z。

步骤3：计算标准化矩阵Z的相关系数矩阵R。

步骤4：用雅克比方法解矩阵R的特征方程，得特征根和特征向量(主成分)。

步骤5：计算主成分累计贡献率，一般选择超过85%贡献率的主成分作为重要主成分。

步骤6：重要主成分即为降维后数据。

1.4 金融审计数据挖掘分析

数据挖掘的概念是在20世纪80年代提出的，其定义是指从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中找寻价值信息和知识的过程，因此数据挖掘也被称为“知识发现”，一般分为以下几个过程，如图1所示。

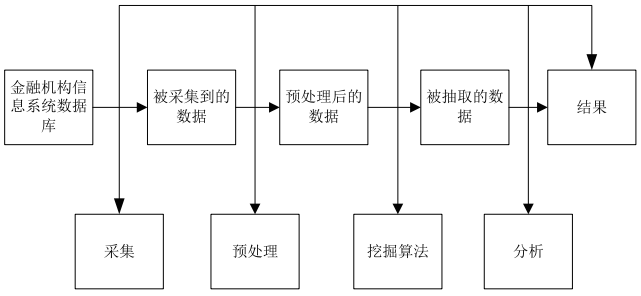

图1 数据挖掘一般过程

从图1中可以看出，前文已经介绍了数据挖掘的准备环节，现进行数据挖掘分析。在数据挖掘中，关键在于算法的选择上。数据挖掘算法有很多，如聚类算法、神经网络算法、决策树算法、遗传算法等。选择聚类算法进行数据挖掘，其理由是该算法具有可伸缩性，图形数据集的划分效果良好，并且十分高效。聚类算法是指按照某种数据特征进行分类，同一类的数据具有相同的特征，从而从中分辨出不同特征的数据，即异常数据。具体过程大致分为以下五个步骤。

步骤1：从金融审计数据中随机选定K个样本数据点作为初始聚类中心；

步骤2：计算这K个初始聚类中心到其他样本数据点的距离，包括闵可夫斯基距离、欧式距离、曼哈顿距离等；

步骤3：根据距离计算结果对每一个样本数据点进行分类；

步骤4：重新计算每个类的类中心；

步骤5：判断得到的新聚类中心是否与原来的初始聚类中心相同。若相同，聚类结束，输出聚类结果；否则回到步骤2，重新进行聚类，直到目标函数收敛。

2 实验分析

本文为测试基于数据挖掘的金融审计数据分析方法的性能，以神经网络算法、支持向量机、最近邻算法等数据挖掘方法

作为对比项，进行实验分析。

2.1 实验数据

保险公司是我国重要的金融机构之一，关系国计民生，关系社会生活中的每一分子，因此本文就选取某保险公司业务及管理费的核算数据作为仿真实验数据。利用本文章节1.1数据采集方法，从该公司财务系统中获取2008年与业务及管理费相关的全部凭证明细。

据统计，获取的某保险公司2008年与业务及管理费相关的全部凭证明细数据共计1268825条，全部为正常数据，因此本实验通过手动篡改其中1562条数据，作为异常数据，用于后续数据挖掘。

2.2 数据挖掘工具

本文数据预处理、降维以及挖掘分析都通过SPSS公司开发的Clementine12.0数据挖掘软件来实现。通过Clementine12.0，可以将数据放到软件上进行分析，从而利用内置的强大算法以及图形功能预测未来数据的走势，提前定制公司项目计划，定制未来开发的具体流程，并且可以将分析结果建立模型或流程图，方便整个数据挖掘过程将数据部署到企业开发计划上，从而完善企业后期决策计划。

2.3 数据挖掘评价指标

数据挖掘目标是从正常数据中找出异常数据，因此数据有两类，即为正例(positive)和负例(negtive)，构建混淆矩阵如表3所示。

表3 混淆矩阵

项目	Yes	No	总计
Yes	TP	FN	P(实际为 Yes)
No	FP	TN	N(实际为 No)
总计	P’(被分为 Yes)	N’(被分为 No)	P+N

根据混淆矩阵，计算数据挖掘精度，公式如下：

$$\text{precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (4)$$

2.4 结果分析

利用Clementine12.0数据挖掘软件进行数据挖掘，并统计挖掘结果，如表4所示。

表4 数据挖掘结果

方法	精度 /%
本文方法	96.38
神经网络算法	92.36
支持向量机	90.25
最近邻算法	88.14

从表4中可以看出，利用本文方法进行数据挖掘后，从1268825条正常数据中找寻1562条异常数据的精度达到96.38%，高于神经网络算法、支持向量机、最近邻算法三种挖掘方法，证明了本研究的有效性。

3 结语

综上所述，面对海量的金融审计数据，如何有效地从中挖掘出有价值的潜在信息对于审计部门来说至关重要，为此本文基于数据挖掘进行金融审计数据分析研究，并取得了一定的成果，但是由于个人能力有限，还存在不足之处，如在进行仿真研究时，样本数量并不符合数据挖掘对大量数据的要求，且财务指标不够全面，因此得出的结果可能与实际结果存在一定的误差，因此有待进一步的探讨和研究。

参考文献

[1] 赵圣伟,吴雨横.基于金融审计大数据的证券市场异常交易模型探讨[J].审计研究,2018,205(05).

[2] 朱蕊,田晨,高岑.基于数据挖掘的熨法干预肩周炎药物使用规律研究[J].时珍国医国药,2018,29(09).

[3] 陈伟,勾东升,徐发亮.基于文本数据分析的大数据审计方法研究[J].中国注册会计师,2018,234(11).

[4] 钟若武,王惠平.基于数据挖掘的高校云计算管理系统中特定数据查询技术[J].现代电子技术,2018,41(02).