基于改进孤立森林算法的审计数据异常主动预警研究

蔡玲嘉[1] 李 雄[1] 王泽涌[2] 陈哲瀚[2]
CAI Lingjia LI Xiong WANG Zeyong CHEN Zhehan

摘 要 审计数据中出现的异常值可能反映业务活动的异常状态，在进行数据质量评估时，有效异常数据越多，说明数据质量越高，因此在海量审计数据异常检测与识别时，需要对海量异常点进行分析与解释，复杂度较高。为此，文章提出一种基于改进孤立森林算法的审计数据异常主动预警研究。首先自动抓取并预处理审计相关数据，然后选取关键特征，基于选择特征构建孤立森林，最后引入四分位法改进孤立森林中正常与异常数据评分的分界线划分方式，利用改进后的孤立森林算法实现审计数据异常的主动预警。实验结果表明，在此方法下，审计数据异常主动预警结果的 G-mean 指标高达 0.96，AUC 高达 0.936，验证了该方法的有效性和优越性。

关键词 改进孤立森林算法；审计数据；数据异常；异常预警；主动预警

doi：10.3969/j.issn.1672-9528.2025.03.031

0 引言

随着企业规模的扩大和业务范围的拓展，审计数据维度复杂多样，传统的人工审计方式已难以高效、精准地发现潜在风险与异常。因此，如何有效利用现代信息技术手段，对审计数据进行自动化、智能化检测，及时发现并预警异常行为，成为审计领域亟待解决的重要问题。近年来，众多学者在这一领域开展研究，如文献 [1] 采用 BiLSTM 捕捉时序数据时间依赖，引入 Wasserstein 距离及梯度惩罚改进评估。结合重构与判别损失定义异检函数，用局部自适应阈值提升时序数据异检精度。该方法在应用过程中受到审计数据复杂性和多样性的限制，难以检测审计数据中的有效异常数据。文献 [2] 针对评价基准误差问题，提出混合选择集成法，用孤立森林预排劣基检测器，采用动态选择集成，提出假真值准度。为解决单一指标不足，引入元学习策略，转多基检测器动选为二分类问题，设多元特征训元分类器。据元分类器选优基检测器，提电力调度数据异检能力。然而，该方法在运行过程中受到了计算资源和时间复杂度的限制，无法对审计数据中的海量异常点进行全面、深入的分析与解释。针对上述局限之处，本文提出了一种基于改进孤立森林算法的审计数据异常主动预警方法，以期实现对审计数据中异常行为的精准、高效检测与主动预警，有助于推动审计领域的技术创新与发展。

1. 广东电网有限责任公司 广东广州 510699
2. 南方电网数字企业科技（广东）有限公司 广东广州 510699

1 爬取审计数据

利用网络爬虫技术，审计人员可以通过系统自动化地从多个数据源中获取数据，减少手动收集数据的时间和精力。同时，该技术还能够从多个数据源中收集数据，从而提供更全面的视角 [3]。要想获取某些访问限制或法律约束的数据源数据，需要在进行审计数据异常主动预警时，高效、准确地从多元化数据源中爬取审计相关数据 [4]。对此，本文采用网络爬虫技术进行审计数据爬取。由于审计数据的来源广泛，主要包括企业内部系统、外部数据源以及社交媒体等非传统数据源，为实现多任务并发执行与负载均衡，本文设计一个多层次的爬取策略 [5]，将异步 IO 与分布式爬虫框架集成在一起，具体审计数据爬取的关键步骤为：先根据数据源分析确定初始 URL 集合，通过内部网络或公开域名 /API 访问 [6]；再构建 HTTP 请求头，发送请求至服务器，用 HTML 解析器提取数据如财务报表、交易记录、系统日志；之后，逐一读取并解析 URL，直至满足停止条件，为后续分析预警打下基础。

2 预处理审计数据

原始爬取的审计相关数据在应用于数据异常主动预警之前，需要进行严格的数据清洗和预处理工作，以确保数据的准确性和完整性。为提升数据质量，需要删除冗余信息与噪声，以减少对后续分析的干扰 [7]。为识别并去除数据中的冗余与重复信息，本文采用分析算法获取原始审计数据之间的相关性：

$$\gamma_{ij}=\frac{C\left(X_i,X_j\right)}{\sqrt{F\left(X_i\right)}\sqrt{F\left(X_j\right)}} \tag{1}$$

式中：γ_{ij} 表示审计数据 X_i 与 X_j 之间的相关性；$C(X_i, X_j)$ 表示审计数据 X_i 与 X_j 之间的协方差；$F(X_i)$、$F(X_j)$ 分别表示审计数据 X_i 与 X_j 的方差。准确识别并剔除原始审计数据中高度相关的变量，从而精简数据集，减少冗余信息，为后续异常预警排除干扰。与此同时，原始审计数据中的噪声点也是必须要剔除的目标 [8]，本文采用基于密度的局部离群因子算法进行噪声点的检测，密度的具体计算公式为：

$$D_{\gamma_{ij}}=\frac{1}{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\left|X_i-X_j\right|+\left|Y_i-Y_j\right|\right)/2N} \tag{2}$$

式中：$D_{\gamma_{ij}}$表示密度；N 表示审计数据的点集的数量，点集定义为 (X_i, Y_i)。由此求出待识别数据点周围密度后，所求密度数据的倒数即为该数据点的离群计算数值。当某个数据点的离群计算数值显著高于其相邻数值时，即被视为孤立的噪声点，需进一步剔除。然后，需要进行数据转换，即将原始审计数据统一转换为适合分析的数值型格式。最后，为消除不同量纲对数据异常预警的影响，本文还需对原始审计数据进行归一化处理，表达式为：

$$X^{'}=\frac{X_{D_{\gamma_{ij}}}-X_{\min}}{X_{\max}-X_{\min}} \tag{3}$$

式中：$X_{D_{\gamma_{ij}}}$、$X^{'}$ 分别表示归一化前、后的审计数据；$X_{\min}$、$X_{\max}$ 分别表示原始审计数据中的最小、最大值。

3 选择审计数据特征

审计数据预处理可规范数据，去除噪声和冗余，提高质量。因审计数据维度高且复杂，直接检测可能计算量大、准确性低，故需先选特征。特征选择去冗余，提模型泛化力。本文在孤立森林模型中，用近似马尔可夫毯算法选关键特征，提高预警效率 [9]。首先，基于审计数据集，利用互信息计算特征间的依赖关系，具体表达式为：

$$\lambda_0\left(Z_u,Z_v\right)=\sum_{Z_u\in Z}\sum_{Z_v\in Z}X_{D_{\gamma_{ij}}}P(Z_u,Z_v)\log\left(\frac{P(Z_u,Z_v)}{P(Z_u)P(Z_v)}\right) \tag{4}$$

式中：$\lambda_0(Z_u, Z_v)$ 表示审计数据特征 Z_u 和 Z_v 之间的互信息；Z 表示审计数据特征集合；$P(Z_u, Z_v)$ 表示审计数据特征 Z_u 和 Z_v 的联合概率分布；$P(Z_u)$、$P(Z_v)$ 分别表示审计数据特征 Z_u 和 Z_v 的边缘概率分布。由此可以计算所有特征对象之间的互信息后，即可构建一个加权特征依赖图 T：

$$T=\left(Z,\lambda,\lambda_0\left(Z_u,Z_v\right)\right) \tag{5}$$

式中：λ 表示审计数据特征间依赖关系的集合。审计数据特征依赖图以节点表特征，边表依赖强度。构建图后，设阈值找近似马尔科夫毯：算候选特征依赖增益，选超阈值且增益大者入子集，直至条件不满足 [10]。得到最小特征子集，能隔目标与非关键特征依赖。本文用此算法选择关键特征，为孤立森林构建奠定基础。

4 改进孤立森林算法主动预警审计数据异常

审计数据特征提取可精简数据，从而减少干扰，提高异常检测效率与准确性。基于此，采用孤立森林算法预警。将选定特征输入孤立树根节点，随机选择属性分割，设定阈值左右分配特征，直至隔离或达最大高度 [11-12]。构建所有孤立树后组成森林，根据评分规则判定数据异常，生成异常评分，具体计算公式为：

$$f=2^{-\frac{T\cdot E\left[H(z)\right]}{R(m)}} \tag{6}$$

式中：f 表示单个异常审计数据样本特征 z 相对于样本集 m 的异常分数值；$E[H(z)]$ 表示遍历孤立森林中的每一棵孤立树后，得到的路径长度 $H(z)$ 的期望值；$R(m)$ 表示路径长度 $H(z)$ 在 m 个样本条件下的平均值。孤立森林算法据特征异常分数判异常点。常规算法受审计数据特征多样性和非平衡性影响，本文用四分位法改进正常与异常数据评分分界。假设上述孤立森林算法计算的各个审计数据特征样本异常分数值集合为 $f=[f_1,f_2,\cdots,f_m]$，集合中各分数值按升序排列，通过四分位法对该异常分数值集合进行四分位划分，可得到异常分数值数据的第 1/4、1/2、3/4 的数据分别为 q_1、q_2、q_3。通过公式计算四分位距：

$$\varsigma=q_3-q_1 \tag{7}$$

式中：ς表示审计数据特征样本异常分数值集合f的四分位距。基于四分位距，计算审计数据异常检测阈值范围 $[f_{\min},f_{\max}]$：

$$\begin{cases}f_{\min}=q_1-1.5\varsigma\\ f_{\max}=q_3+1.5\varsigma\end{cases} \tag{8}$$

将所求阈值范围内的审计数据判定为正常数据，而不在该阈值范围内的审计数据判定为异常数据。因此，利用四分位法改进后的孤立森林算法，能对审计数据进行高效的异常检测。

5 仿真实验

5.1 实验环境与数据

为验证所提方法的先进性，选取基于生成对抗网络的审计数据异常主动预警方法、基于卷积神经网络的审计数据异常主动预警方法作为对比方法，进行仿真对比实验。在 Linux 系统服务器中搭建本次实验环境，参数配置情况如表 1 所示。

表 1 实验环境参数设置

实验环境	参数	设置
硬件环境	CPU	Intel Core i9-9900K
	GPU	NVIDIA GeForce RTX 2080 Ti
	内存	32 GB RAM
软件环境	Python 3.8，TensorFlow 2.4，scikit-learn 0.24，NumPy 1.19	

然后，收集了某企业多个审计项目的 3 600 条数据作为实验数据，如表 2 所示。

表 2 实验审计项目数据表（部分）

ID	交易金额 / 元	交易类型	凭证编号	凭证状态	…	是否异常
1	10 000	采购	PT-20230401-001	已审核	…	0（正常）
2	25 000	销售	PT-20230402-002	待审核	…	0（正常）
3	5 000	报销	PT-20230405-003	已支付	…	1（异常）
4	8 500	工资发放	PT-20230410-004	已发放	…	1（异常）
5	3 100	预付款	PT-20230507-005	已支付	…	1（异常）
6	12 650	借款	PT-20230518-006	待审批	…	0（正常）
…	…	…	…	…	…	…
359	7 500	退货	PT-20230609-007	已处理	…	0（正常）
360	9 640	投资收益	PT-20230910-008	已入账	…	0（正常）

表 2 所示的 360 条实验审计项目数据中，正常数据 300 条，异常数据 60 条。经过缺失值填充、异常值处理等一系列预处理后，应用于实验组和对照组中审计数据异常主动预警方法的性能测试上。

5.2 实验指标

在本次实验中，为评价实验组中方法和对照组中两种方法在实验数据集上的主动预警效果，根据实验结果生成表 3 所示的混淆矩阵。

表 3 混淆矩阵

真实结果	预警结果	
	0	1
0	真正例 A 279	假负例 B 21
1	假正例 C 6	真负例 D 54

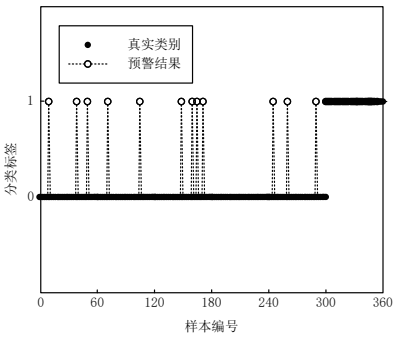
基于表 3 所示混淆矩阵，采用 G-mean 指标作为实验指标，其具体计算公式为：

$$G=\sqrt{\frac{A}{A+B}\times\frac{D}{D+C}} \tag{9}$$

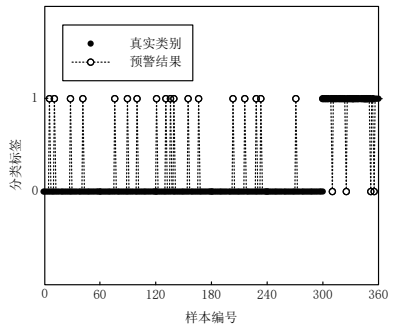
式中：G 表示 G-mean 指标，是一种用于衡量不均衡样本分类性能的综合评价指标，其值越大，说明审计数据异常主动预警效果越好。

5.3 结果分析

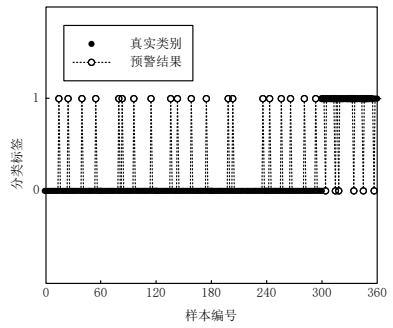
在完成实验组中方法和对照组中两种方法下的实验审计项目数据异常主动预警后，统计并整理各方法下的数据异常预警结果，如图 1 所示。



（a）所提孤立森林算法预警结果



（b）生成对抗网络方法预警结果



（c）卷积神经网络方法预警结果

图 1 审计数据异常主动预警实验结果

从图 1 看出，孤立森林算法覆盖所有异常数据，G-mean 指标最优达 0.96。这一优异表现主要得益于近似马尔科夫毯特征选择方法的应用，该方法能够有效剔除冗余特征，保留与异常检测高度相关的关键特征，从而显著提升了模型的稳

定性和准确性。此外，本文在传统孤立森林算法的基础上引入了四分位法，改进了正常与异常数据评分的分界线划分方式，使得异常数据的识别更加精准，减少了误报和漏报的可能性。这种改进不仅提高了算法的检测精度，还增强了其在实际应用中的鲁棒性。同时，所提方法还具备主动预警机制，能够实时分析审计数据，及时发现潜在风险，为企业提供决策支持，从而有效降低风险发生的可能性。综上所述，本文提出的基于改进孤立森林算法的审计数据异常主动预警方法，在特征选择、算法改进、预警机制和性能指标等方面均表现出显著优势，不仅能够提升异常检测的准确性和效率，还为企业审计提供了强有力的技术支持，具有广泛的应用前景和实际价值。

为展示审计效率，分析 3 种预警方法精确率随召回率变化，结果如图 2 所示。

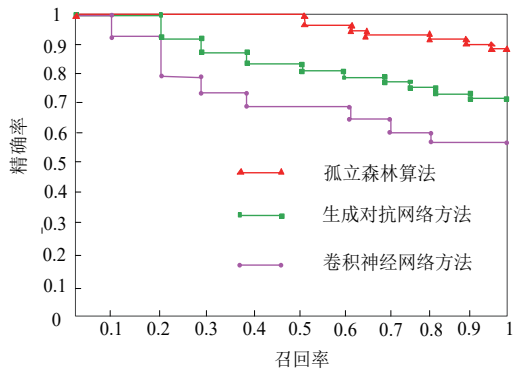


图 2 三种方法精确率 - 召回率曲线图

根据图 2 得出，孤立森林算法 AUC 达 0.936，高于生成对抗网络的 0.716 和卷积神经网络的 0.812。这表明所提方法在审计数据异常预警上更准确有效，能精确区分正常与异常数据，从而提升审计效率和准确性。

综上所述，本文提出的基于改进孤立森林算法的审计数据异常主动预警方法，不仅在性能指标上优于其他主流算法，还在实际应用中展现出更高的准确性和稳定性，为企业审计提供了更加可靠的技术支持，具有重要的实践意义和推广价值。

6 结语

本文针对审计数据异常检测问题，提出了一种基于改进孤立森林算法的主动预警方法，通过引入四分位法改进常规孤立森林算法，有效提升了审计数据异常的检测性能，为审计领域的智能化、自动化发展提供了新思路和技术支持。

参考文献：

[1] 王德文，潘晓飞，赵红博 . 基于改进生成对抗网络的时序数据异常检测 [J]. 计算机工程与设计，2024, 45(3): 762-768.

[2] 傅世元，高欣，张浩，等 . 基于元学习动态选择集成的电力调度数据异常检测方法 [J]. 电网技术，2022, 46(8): 3248-3261.

[3] 钟杰，罗冲，张恒，等 . 基于相关性参数选择的飞行数据异常检测 [J]. 北京航空航天大学学报，2024, 50(5): 1738-1745.

[4] 吴花平，黄尹薇，刘自豪 . 基于 K-means 聚类算法的碳排放审计预警研究 [J]. 中国注册会计师，2022(12):14-20.

[5] 陈艾荣，李梓巍，潘玥，等 . 基于本福特定律的桥梁健康监测数据审计方法 [J]. 同济大学学报 (自然科学版), 2023, 51(4): 534-541.

[6] 何家辉，程志君，郭波 . 联合字典学习与 OCSVM 的遥测数据异常检测方法 [J]. 航空学报，2023, 44(13): 207-219.

[7] 张建，胡小锋，张亚辉 . 基于自步学习的刀具加工过程监测数据异常检测方法 [J]. 上海交通大学学报，2023, 57(10): 1346-1354.

[8] 凌继红，邢金城，李昂，等 . 基于孤立森林算法的集中供热系统异常数据识别研究 [J]. 暖通空调，2023, 53(2): 97-102.

[9] 唐立，郝鹏，任沛阁，等 . 基于改进孤立森林算法的无人机异常行为检测 [J]. 航空学报，2022, 43(8): 584-593.

[10] 魏泰，贺少雄，胡子武，等 . 基于改进孤立森林算法的风电机组异常数据清洗 [J]. 科学技术与工程，2024, 24(9): 3691-3699.

[11] 范斌，宁德军，卢俊哲，等 . 基于加权 KNN 与代价敏感多分支深度神经网络的审计数据异常检测 [J]. 计算机应用与软件，2024, 41 (2): 100-108.

[12] 李金花 . 基于关联规则的医院内部审计异常信息挖掘方法 [J]. 信息与电脑 (理论版), 2023, 35 (21): 211-213.

【作者简介】

蔡玲嘉（1989—），女，广东汕头人，硕士，高级工程师，研究方向：数字化、审计。

李雄（1987—），男，陕西渭南人，硕士，高级会计师，研究方向：审计、信息。

王泽涌（1988—），男，湖南湘潭人，本科，高级工程师，研究方向：电气工程、软件工程。

陈哲瀚（1984—），男，广东陆丰人，本科，工程师，研究方向：软件工程。

（收稿日期：2024-10-22）