融合多算法对比的高校数字化校园异常数据检测与智能预警模型研究

黄敏菁
柳州城市职业学院，广西 柳州 545036

DOI:10.61369/EST.2026070032

摘　　要： 随着高校数字化校园建设的推进，多业务系统产生了大量数据，但在整合与治理过程中易出现缺失、重复、冲突及行为异常等问题，影响数据质量与决策效率。本文以某高校脱敏数据为样本，构建异常数据检测指标体系，选取K-Means、LOF、Isolation Forest、Random Forest、XGBoost等算法进行对比分析，并基于模型性能设计融合多算法的异常检测与智能预警模型，实现异常识别、评分与分级预警。研究结果表明，多算法融合方法能够有效提升异常检测的准确性与稳定性，为高校数据治理与智能决策提供参考。

关 键 词： 数字化校园；异常数据检测；机器学习；多算法对比；智能预警

Research on the Intelligent Early Warning Model for Abnormal Data Detection in University Digital Campus Based on the Integration of Multiple Algorithms Comparisons

Huang Minjing
Liuzhou City Vocational College, Liuzhou , Guangxi 545036

Abstract: With the advancement of digital campus construction in universities, multiple business systems have generated a large amount of data. However, during the integration and governance process, problems such as missing data, duplication, conflicts, and abnormal behaviors may occur, which affect data quality and decision-making efficiency. This paper takes the anonymized data of a certain university as a sample, constructs an abnormal data detection index system, selects algorithms such as K-Means, LOF, Isolation Forest, Random Forest, and XGBoost for comparative analysis, and designs an abnormal detection and intelligent early warning model based on model performance, achieving abnormal identification, scoring, and graded early warning. The research results show that the multi-algorithm fusion method can effectively improve the accuracy and stability of abnormal detection, providing a reference for university data governance and intelligent decision-making.

Keywords: digital campus; abnormal data detection; machine learning; multi-algorithm comparison; intelligent early warning

引言

随着高校数字化校园建设持续推进，教务、学工、在线教学、智慧门禁、一卡通、科研、后勤、人事等系统积累了大量结构化和半结构化数据，为教学诊断、学生服务、资源调度和风险预警提供了基础。但多系统分散建设也伴随着字段标准不统一、跨系统信息冲突和异常行为识别滞后等问题，影响数据质量与管理决策效率[1-2]。从应用场景看，学校数据异常问题主要在字段缺失、格式错误、跨系统不一致和业务数据异常波动等方面。如果不能及时发现和处理，容易造成各系统之间数据交换不畅通、信息需重复采集、分析结果失真和风险响应滞后。教育数据挖掘研究表明，机器学习方法能够从复杂教育过程数据中发现潜在规律，为学习状态识别、学生画像和管理决策提供支持[3]。异常检测作为数据挖掘的重要方向，可通过聚类、密度估计、集成学习等方法识别偏离正常模式的样本[4]。同时，不同算法在不同数据类型和异常模式下表现差异明显，单一模型难以适应高校多源业务数据场景，多算法对比与融合成为提升检测稳定性的重要思路[5]。因此，本文以某高校数字化校园脱敏数据为对象，构建异常数据检测指标体系，选取K-Means、LOF、Isolation Forest、Random Forest、XGBoost等模型开展对比实验，并设计加权融合模型，实现异常识别、风险评分与分级预警。

作者简介：黄敏菁（1995.5-），女，壮族，广西柳州人，高级工程师，研究方向：大数据技术与应用，数据挖掘，数学建模。

一、异常数据特征分析

高校数字化校园数据具有来源多样、结构复杂、更新频繁和业务关联性强等特点。结合某高校数据中台运行情况，本文将异常数据划分为三类。

（一）数据质量异常

通过学校数据中台汇集数据后发现数据质量异常主要产生于数据采集环节，存在的主要问题有字段缺失、格式错误、各系统中相同业务表述不一致等问题。例如，学生在学工系统和一卡通系统中的 ID 或专业信息不一致，入学年份信息确实，消费记录不完整等。数据质量异常会影响系统间数据交换和后续模型训练，是数字化校园数据治理的基础问题。

（二）行为模式异常

行为模式异常主要体现在用户使用学校系统过程中的行为偏离，如通过统一身份认证平台数据发现短时间内频繁登录、异常异地登录，或通过在线教学平台发现长期未登录平台、资源访问频次突增或突降等。行为模式异常可能意味着学习投入不足、账号异常使用或系统访问风险等问题。

（三）业务过程异常

业务过程异常与学校的教学管理、资产管理等核心业务运行密切相关，主要表现为考勤异常、资源使用异常、业务流程异常等。例如，某班级课程资源访问量持续偏低，某实训室用电量明显异常，某类报修记录短期集中出现等。业务过程异常具有较强业务解释性，需要结合具体管理场景进行判断，不能单凭数据下结论。

二、融合多算法对比的异常检测模型构建

针对学校管理系统多、数据来源复杂、异常类型多样和业务场景差异明显等特点，本文尝试构建融合多算法对比的异常数据检测与智能预警模型。模型以某高校数据中台采集的脱敏数据为样本，按照“数据预处理—特征构建—模型训练—结果对比—融合预警”的实验步骤开展研究。

（一）数据预处理

首先对教务系统、学工系统、在线教学平台、一卡通和智慧后勤等常用业务系统数据进行清洗与规范化处理。一是统一字段名称并对相同业务含义字段进行映射保证代码统一，如：学工系统中的学生 ID 和一卡通系统中的学生 ID 统一学生代码；二是处理缺失值，可根据实际业务需求采取填充或删除处理；三是筛查明显异常极值，如毕业学生仍产生在校记录等；四是对涉及个人隐私的数据进行脱敏，保证实验合规。

（二）特征指标构建

异常检测模型的效果很大程度上收到特征值的影响，因此选择和构造合适的特征值成为模型成功的关键。结合学校业务特点，从数据质量和业务过程两个个层面构建特征值。数据质量层面包括缺失率、重复率、字段冲突率、格式错误率和数据更新延迟等；业务过程层面包括考勤异常率、成绩波动率、课程资源访问偏离度、消费频次变化率等。通过这些指标，可以将分散的业务数据转换为可计算、可比较的特征矩阵，为后续模型训练提供输入。

（三）模型选择与评价

考虑学校数据以结构化数据为主，且需要兼顾检测效果和可解释性，本文选取五类模型进行对比。K-Means 用于发现群体分布异常；LOF 用于识别局部离群样本；Isolation Forest 适合高维未知异常检测；Random Forest 可用于有标签异常分类并具有较强稳定性；XGBoost 适合复杂特征条件下的高精度异常识别。

模型评价采用 Accuracy、Precision、Recall、F1值、AUC 值和检测时间等指标。其中，召回率用于衡量模型发现异常数据的能力，F1值用于综合评价精确率与召回率。考虑到校园异常预警更强调风险发现能力，本文重点关注 Recall 和 F1值，而不单纯以 Accuracy 作为判断依据。

（四）融合预警策略

在多算法对比基础上，本文进一步设计加权融合策略。首先，根据各模型在验证集上的 Recall、F1值、AUC 值和检测时间确定权重，其中 Recall 和 F1值作为主要依据，AUC 值作为辅助依据，检测时间作为运行效率约束；其次，采用 Min-Max 归一化方法将不同模型输出的异常分数统一映射到 [0,1] 区间；最后计算综合异常得分，并据此划分预警等级。综合异常得分低于0.60时判定为正常，$0.60 \leqslant S<0.75$时生成一级预警，$0.75 \leqslant S<0.90$时生成二级预警，$S \geqslant 0.90$时生成三级预警。通过“模型检测—等级判定—人工核验—结果反馈—模型优化”的闭环机制，可以不断修正模型参数，提高异常检测的准确性和稳定性。

三、实验设计与结果分析

为验证融合多算法异常检测模型在高校数字化校园场景中的适用性，本文以某高校数据中台脱敏数据为实验样本，选取教务管理系统、学生工作系统、在线教学平台、智慧门禁、一卡通系统和智慧后勤中的部分业务数据进行建模分析。经数据清洗、去重、字段统一和隐私脱敏处理后，形成有效样本12000条，提取18个异常检测特征指标。其中，依据业务规则和人工核验结果标注异常样本980条，异常样本占比约8.17%。该比例符合高校数字化校园异常数据“少量存在、分布分散、类型多样”的基本特征。

（一）实验设计

实验首先对原始数据进行预处理，并按照业务规则与人工核验相结合的方式构建异常样本标签。标注规则包括三类，一是数据质量规则，如关键字段缺失、跨系统字段冲突、格式不合法、数据重复；二是业务风险规则，如考勤异常、成绩波动异常、账号异常访问、系统操作异常等。监督模型按照7:2:1划分训练集、验证集和测试集；无监督模型输出异常分数后按阈值转换为正常或异常标签，并与人工核验标签进行对比。K-Means 以样本到聚类中心距离作为异常得分，LOF 和 Isolation Forest 以模型输出的离群分数作为异常得分。使用 python 中的机器学习库 scikit-

learn、xgboost 等完成建模，检测时间取3次运行平均值。

（二）实验结果分析

从实验结果看，五种单一模型表现存在差异。K-Means 运行效率较高，但 Precision、Recall 和 F1值相对偏低，说明其对局部异常和复杂业务异常识别能力有限；LOF 整体优于 K-Means，但在数据密度不均衡时可能误判；Isolation Forest 在检测效率和未知异常识别方面表现较好；Random Forest 和 XGBoost 作为监督学习模型，在异常样本标注明确时效果更优，其中 XGBoost 的 Recall、F1值和 AUC 值均高于其他单一模型，说明其复杂特征建模能力较强。实验结果如表1所示。

表1 多算法异常检测性能对比表

模型	Accuracy	Precision	Recall	F1值	AUC值	检测时间/s
K-Means	0.842	0.781	0.736	0.758	0.811	1.86
LOF	0.861	0.803	0.768	0.785	0.837	2.34
Isolation Forest	0.884	0.829	0.812	0.82	0.872	1.52
Random Forest	0.907	0.866	0.841	0.853	0.901	2.71
XGBoost	0.921	0.884	0.867	0.875	0.924	3.18
融合模型	0.936	0.902	0.891	0.896	0.942	3.86

在融合模型权重设置上，本文遵循“风险发现优先”的原则，将 Recall 和 F1值作为权重分配的主要依据，将 AUC 值作为模型区分能力的辅助依据，将检测时间作为运行效率约束。经过调优根据各模型在验证集上的表现，XGBoost 的 Recall、F1值和 AUC 值均为最高，说明其在异常识别和风险区分方面表现最优，故赋予最高权重0.30；Random Forest 的模型稳定性和可解释性较好，权重设为0.25；Isolation Forest 对未知异常具有较强发现能力，且检测效率较高，权重设为0.20；LOF 和 K-Means 分别用于局部离群检测和群体分布异常识别，但整体 Recall 和 F1值相对较低，权重分别设为0.14和0.11。

在计算综合异常得分前，采用归一化方法将各模型输出的异常得分统一映射到 [0,1] 区间，融合模型综合异常得分计算公式如下：

$$S=0.11*S_{KMeans}+0.14*S_{LOF}+0.2*S_{IForest}+0.25*S_{RF}+0.3*S_{XGBoost}$$

其中，S 表示综合异常得分，S_{KMeans}、S_{LOF}、$S_{IForest}$、S_{RF}、$S_{XGBoost}$ 分别表示各模型归一化后的异常得分。根据综合异常得分划分预警等级：当 S<0.60时，判定为正常；当0.60 ≤ S<0.75时，生成一级预警；当0.75 ≤ S<0.90时，生成二级预警；当 S ≥ 0.90时，生成三级预警。一级预警主要用于提示轻微异常并纳入观察，二级预警推送至相关管理部门核查，三级预警则生成处置工单并要求限时处理。

从表1可以看出，融合模型在 Recall、F1值和 AUC 值等核心指标上取得最优结果，分别达到0.891、0.896和0.942，说明多算法融合能够综合不同模型优势，提升异常发现能力和检测稳定性。虽然融合模型检测时间为3.86s，高于单一模型，但仍处于可接受范围，适合高校周期性数据治理和智能预警应用。

四、结论与展望

本文围绕高校数字化校园数据治理与智能预警需求，构建了面向数据质量异常、行为模式异常、业务过程异常和安全风险异常的检测指标体系，选取 K-Means、LOF、Isolation Forest、Random Forest、XGBoost 等算法进行对比，并提出加权融合模型。实验结果表明，融合模型能够弥补单一模型适应性不足的问题，在 Recall、F1值和 AUC 值上取得较优表现，可为高校异常数据识别、风险评分和分级预警提供技术支撑。本文仍存在不足：实验数据来自单一高校，异常样本标注依赖业务规则和人工经验，对非结构化数据和实时流数据处理不足。后续可引入 AutoEncoder、LSTM、Transformer 等模型，并结合知识图谱和大语言模型提升异常解释与处置建议生成能力，推动预警系统向实时化、智能化和可解释方向发展。

参考文献

[1] 吴瑞芳 . 大数据背景下高校智慧校园建设研究与探索 [J]. 中国信息化 ,2023,(9):94-95+101.

[2] 陈荣荣 . 基于分布式计算的数字化校园云存储网络安全策略研究 [J]. 自动化与仪器仪表 ,2023,(9):31-35.

[3]ROMERO C, VENTURA S. Educational data mining and learning analytics: An updated survey[J/OL]. arXiv, 2024.

[4]PANG G S, SHEN C H, CAO L B, et al. Deep Learning for Anomaly Detection: A Review[J/OL]. arXiv, 2020.

[5]HAN S Q, HU X Y, HUANG H L, et al. ADBench: Anomaly Detection Benchmark[J/OL]. arXiv, 2022.

