数字化教学软件安全漏洞集成学习检测方法

尚 进[1]
SHANG Jin

摘 要 在数字化教学软件中，学生行为埋点数据具有高维度、多模态和强时序关联的特性，且隐私泄漏模式常表现为低频长尾分布与正常行为的语义模糊边界，导致安全漏洞检测效果较差。因此，文章提出一种新的数字化教学软件安全漏洞集成学习检测方法。移动终端采集的代码、日志和用户反馈数据经主成分分析降维处理后，利用模糊分类矩阵实现漏洞数据精准归类。基于抽象树状图和端口开放率提取静态与动态特征，形成动态更新的特征库。采用随机森林算法集成弱分类器构建强检测模型，通过基尼指数和信息熵评估实现节点分裂优化。特征随机性与样本随机性相结合生成强分类器，最终输出漏洞类型判定结果。实验结果表明，所研究方法在 24 个测试样本中仅出现 1 次误判，整体检测准确率达到 92.6%，漏检率稳定控制在 20% 以内。

关键词 数字化教学软件；安全漏洞；集成学习；随机森林；漏洞检测

doi：10.3969/j.issn.1672-9528.2026.03.029

0 引言

在线上教学和混合式教学等新型教学模式中，数字化教学软件是关键工具之一，具备覆盖课程管理和互动直播以及学习分析等全流程优势，能够显著提高教师的教学质量和学生的学习效果。然而，由于软件复杂度和网络依赖性的日渐提升，数字化教学软件的安全漏洞风险也逐渐提高，容易出现数据泄漏和服务中断的情况，进而威胁教学秩序和用户的隐私安全。

为此，亟须研究一种针对数字化教学软件的安全漏洞检测技术。程广振[1]提出基于多源数据融合的漏洞检测方法，通过整合无线传感器网络、网络流量日志、设备状态日志等异构数据源，提取漏洞检测所需的静态特征，进而通过特征融合与建模确定漏洞风险。然而，其静态特征提取机制无法有效识别动态交互场景中产生的时序敏感型漏洞，导致对需要分析行为序列的复杂漏洞检测效果不佳。曹子亨等[2]提出基于双向数据流分析与图抽象嵌入的漏洞检测方法，采用控制流图与数据依赖图捕获变量间的数据流与控制流关系，结合双向门控图神经网络，对漏洞代码进行分类与定位。但图神经网络模型对动态运行时的多模态交互行为缺乏建模能力，导致无法有效检测由异步操作引发的竞态条件类漏洞。

由于数字化教学软件的特殊应用场景，学生行为数据的高维特性和多模态表征增加了特征提取难度，隐私泄漏行为呈现低频长尾分布且与正常操作存在语义重叠，而复杂的时序关联特性又使得动态交互场景下的异常模式识别更为困难。因此，以数字化教学软件为研究对象，基于集成学习算法并结合实际情况对软件安全漏洞进行检测。本文创新性地提出了一种融合多维度数据的漏洞检测框架，通过主成分分析与模糊分类实现高效数据降维，结合树状图特征提取与端口风险评估构建动态特征库，采用随机森林集成学习算法优化分类器性能，实现了对教学软件漏洞的高精度智能检测。

1 数字化教学软件安全漏洞检测

1.1 教学软件漏洞数据分类

数字化教学软件中的安全漏洞因形成机制、应用位置和威胁模式存在多种类型，因此在进行安全漏洞检测之前，应先对教学软件中的漏洞数据进行分类，为后续识别检测提供优先级依据。

利用移动终端从教学软件的代码、日志、用户反馈等渠道采集教学软件的漏洞数据，包含数据上传时间、连接基站和信号强度以及数据传输频率等多种属性的数据信息[3]。同时为加快漏洞数据的分类速度和精度，采用无监督的主成分分析法将初始运行数据映射到数据分布的最优低维空间，并利用绝对变化率描述数据的模糊属性，以此实现数据降维优化运算[4]。公式为：

$$L=\frac{e_i}{N}\sum_{i=1}^{N}\ln\frac{s_i/\beta}{A} \tag{1}$$

式中：e_i 表示第 i 个数据的绝对变化率；N 表示采集的数据点数量；s_i 表示第 i 个数据的一阶差分项；β 表示有效样本数；A 表示数据分布的最优子空间。

正交分解降维后的数据，去除数据特征之间的关联性，

1. 川北幼儿师范高等专科学校 四川广元 628017

加强数据降维处理效果，在此基础上，利用数据属性协方差矩阵构造模糊分类矩阵 [5]。用公式表示为：

$$\boldsymbol{U}=\sum_{c=1}^{C}\left(\frac{D^2\left(z_b,\varepsilon_i\right)}{V^2\left(\boldsymbol{\iota}_p,L\right)}\right) \tag{2}$$

式中：C 表示时序总数；D 表示所有数据区间的直觉模糊数；z_b 表示时间序列高斯函数的方差；ε_i 表示时间序列的簇中心；V 表示隶属度函数；$\boldsymbol{\iota}_p$ 表示转置的正交矩阵；$\boldsymbol{U}$ 表示模糊分类矩阵。

将概率知识作为模糊分类的理论支撑，利用模糊分类矩阵对教学软件的漏洞数据进行分类处理。公式为：

$$d\left(i,\chi\right)=\frac{p\left(i\right)}{q\bullet X\left(\boldsymbol{U}|v\right)} \tag{3}$$

式中：$p(i)$ 表示第 i 个数据在子空间中出现的概率；q 表示最小分类数量；$X(\boldsymbol{U}|v)$ 表示模糊数据属性 v 与变化状态之间的函数关系；$d(i,\chi)$ 表示数据 i 与类别 χ 之间的兰氏间距测度，取测度最小的类别作为当前漏洞数据的隶属类别，完成漏洞数据分类，得到分类后的漏洞数据集合 $Z(z_1,z_2,\cdots,z_n)$。

1.2 教学软件漏洞特征提取

从分类后的漏洞数据中挖掘出高区分度、可解释性的特征，并构建动态特征库，实时响应新型攻击手段，并以此作为集成学习强分类器的输入信息，进而判定安全漏洞的具体模式。

从代码文件中获取代码结构、函数调用等信息；从日志文件中获取软件运行时的状态信息和错误提示，将这些数据解析为抽象树状图，以总结不同节点之间的语义信息，并根据不同的连接强度将分类的漏洞数据和语义信息投影到向量空间中，得到安全漏洞节点的有序集合 [6]。公式为：

$$\Omega=\frac{j}{um}\prod_{w=1}^{M}S_wZ \tag{4}$$

式中：j 表示树状图中边的数量；u 表示代码中不符合命名规范的变量数；m 表示软件测试用例总数；M 表示文件代码的行数；S_w 表示第 w 行代码对应的链路报文数；Ω 表示安全漏洞节点的有序集合。

根据教学软件的恶意信息流，利用端口开放率评估攻击面，对于其中的高危端口，基于存在异常的软件节点信息提取安全漏洞的基本属性信息。用公式表示为：

$$Y=\frac{\Omega\times E_r}{Z\phi\left(o\right)} \tag{5}$$

式中：E_r 表示开放端口数；Z 表示软件安全漏洞的序列长度；$\phi(o)$ 表示软件系统程序；Y 表示提取出的漏洞属性信息。

引入一种新的激励方法提高软件中安全漏洞特征的最佳搜索途径，则挖掘出的安全漏洞特征可表示为：

$$\varphi=\sigma\cdot\kappa+g\frac{|Y|}{|R|} \tag{6}$$

式中：σ 表示软件运行程序的响应频率；κ 表示安全漏洞攻击的关联分析阈值；g 表示网络向软件传输的数据包总数；R 表示网络与软件交互的特征数量；φ 表示安全漏洞特征，包括静态特征和动态特征。

1.3 集成学习分类器的安全漏洞检测

本研究采用集成学习中的随机森林算法构造多个弱分类器，并将弱分类器集成为一个强分类器，以挖掘出的安全漏洞特征为输入数据，输出对应的漏洞类别。集成学习能够有效融合多模态数据特征，通过组合基模型的互补优势提升检测精度。其自适应权重机制可应对高维数据的稀疏性问题，集成框架对低频异常模式具有更强的敏感性，同时能有效处理时序数据中的长期依赖关系，最终实现更全面的漏洞覆盖和更准确的威胁识别。

在安全漏洞检测过程中，构建 n 个子森林潜在漏洞攻击检测块，并通过抽样策略将子森林划分为若干个组别，引入基尼指数下降量评估每个子森林的重要性 [7]。计算公式为：

$$I_\theta=\frac{1}{Q}\sum_{\theta=1}^{Q}\sum_{\lambda\in n}\Delta\mathrm{Gini}\left(\varphi,h\right) \tag{7}$$

式中：Q 表示森林中决策树的总数；λ 表示划分的组别个数；$\Delta\mathrm{Gini}(\varphi,h)$ 表示特征 φ 在节点 h 分裂时的基尼指数；I_θ 表示第 θ 个子森林的基尼指数下降量。

对所有子森林进行递归运算，进而评价节点数据样本的信息熵 [8-10]。即：

$$T_\psi=I_\theta-\frac{\left|\nu'\cdot F_o\right|}{X}P\left(k'\right) \tag{8}$$

式中：ν'表示落入该节点的所有样本的集合；F_o 表示有放回的随机采样次数；X 表示特征的可能分裂点；$P(k')$ 表示第 k' 棵树的训练数据集；T_ψ 表示节点 ψ 的信息熵。

将信息熵较高的节点看成一个弱分类器，通过样本随机性和特征随机性生成强分类器，并通过输入软件安全漏洞静态特征和动态特征输出漏洞类型。公式为：

$$c'=\arg\max p'\left(T_\psi|\Gamma\right) \tag{9}$$

$$\varsigma^{c'}=\mathrm{SoftMax}\left(\varphi\delta+g_r\right) \tag{10}$$

式中：$p'\left(T_\psi|\Gamma\right)$表示节点信息熵对类别 Γ 的最终决策；c' 表示集成化强分类器；δ 表示分类器分类层的权重；g_r 表示分类器当前训练批次中样本的数量；$\varsigma^{c'}$表示最终的安全漏洞分类结果。

2 测试实验

2.1 实验准备

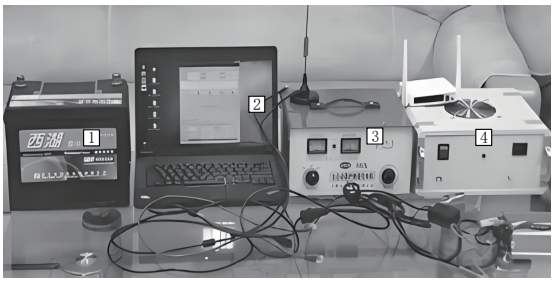
本文实验选取的测试对象为基于微服务架构的在线教学平台，以其中的数字化教学软件为研究背景，该平台包括前端框架和后端服务，应用分片式数据库，采用无线通信协议实现与网络的实时交互。

采用移动终端从该教学软件的交互日志和数据传输数据库中采集安全漏洞数据，并整理后存储于本地化存储模块中，数据集描述如表 1 所示。

表 1 实验数据集样本描述

属性	字段表示	序列长度
漏洞 ID	VARCHAR	20
漏洞名称	VARCHAR	100
所属软件	VARCHAR	50
漏洞类型	VARCHAR	30
CVSS 评分	DECIMAL(3,1)	
漏洞位置	VARCHAR	200
攻击场景描述	TEXT	无限制
漏洞状态	ENUM	—
修复建议	TEXT	无限制
修复版本	VARCHAR	50
发现时间	DATETIME	—
更新时间	DATETIME	—

该教学软件的漏洞模型采用 Windows 系统应用自定义脚本进行开发，使用 VMware ESXi 7.0 虚拟化平台部署教学软件的测试环境，实验现场如图 1 所示。



1. 漏洞扫描设备 2. 网络交换机 3. 采集终端 4. 直流电源

图 1 实验环境

采用 Scikit-learn 1.3.0 集成学习框架实现随机森林模型，算法参数设置如表 2 所示。实验过程为：对采集的安全漏洞数据进行降维处理，并标注漏洞类型，具体包括 SQL 注入漏洞（a）、跨站脚本攻击（b）、跨站请求伪造（c）、逻辑漏洞（d）和敏感数据泄漏（e）五种；通过兰氏间距测度确定安全漏洞数据类别并提取漏洞特征；采用配置完成的随机森林模型对软件安全漏洞类型进行投票决策，得到漏洞类型。

表 2 集成学习模型参数设置

参数	取值
森林中决策树的数量	50
单棵树的最大深度	5
节点分裂所需的最小样本数	2
叶节点所需的最小样本数	1
每棵树分裂时考虑的最大特征数	2
随机种子	60
最小不纯度减少量	0.02
剪枝参数	0.5
并行运行的线程数	20

2.2 教学软件安全漏洞检测结果与分析

将所设计方法应用于该数字化教学软件安全漏洞检测中，将分类器输出的漏洞类型与预设的漏洞代码进行比对，以评估本文方法的检测性能，实验结果如图 2 所示。

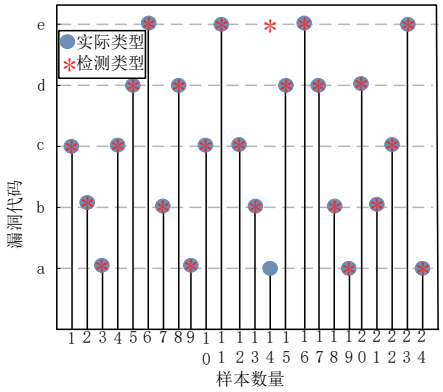


图 2 教学软件安全漏洞检测结果

如图 2 所示，在 24 个测试样本的检测结果中，本文方法的分类输出与真实标注保持了高度一致性。其中仅在第 14 个样本出现误判，将敏感数据泄漏（e）识别为 SQL 注入（a），经分析因样本中存在相似的代码特征干扰了部分决策树的判断。实验结果表明，该方法在数字化教学软件安全漏洞检测任务中展现出可靠的识别准确率。

为客观体现本文方法的综合检测效果，采用对比实验的形式对其进行验证，利用文献 [1] 基于多源数据融合的方法和文献 [2] 基于双向数据流分析与图抽象嵌入的方法作为对照组方法，选取 200 组漏洞数据，分别应用上述方法进行安全漏洞检测，采用漏检率作为衡量指标，漏检率越小。代表对应方法的检测准确度越高，对比结果如图 3 所示。

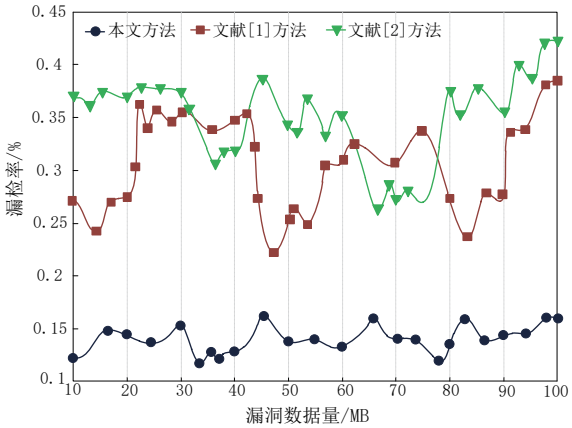


图 3 安全漏洞漏检率对比结果

从图 3 的实验结果对比可以清晰地观察到，文献 [1] 和文献 [2] 提出的方法在安全漏洞检测中存在明显局限性。这两种方法未能有效处理漏洞数据的异构特性，导致特征融合过程中产生时间维度上的匹配偏差，同时对于跨程序边界的复合型漏洞识别能力不足。测试数据表明，两种方法的平均

（下转第 124 页）

[11] 陶琳 . ChatGPT/ 生成式人工智能对高校网络舆情的影响及应对策略 [J]. 情报科学 ,2023,29(8):15-21.
[12]CHELIGEER K, YANG L, NANDI T, et al.Natural language processing (NLP) aided qualitative method in health research[J].Journal of integrated design and process science, 2022, 27(4-5): 1-18.
[13] 胡亮 , 夏磊 , 李伟 . 基于改进 TF-IDF 算法的关键词抽取系统 [J]. 厦门理工学院学报 ,2017,25(5):67-72.
[14] 曹娟 , 张勇东 , 李锦涛 , 等 . 一种基于密度的自适应最优 LDA 模型选择方法 [J]. 计算机学报 ,2008(10):1780-1787.
[15] 张丽娟 . 基于 GBDT+LR 模型在个人信用风险评估中的研究 [J]. 网络安全技术与应用 ,2021(7):40-42.
[16] 刘佳梅 , 丁楷 . 基于 Stacking 集成学习的中文问句分类算法 [J]. 智能计算机与应用 ,2023,13(9):85-88.
[17] 俞立群 . 基于 TextCNN 的试飞运营问题分类 [J]. 民用飞机设计与研究 ,2023(4):1-5.
[18] 刘春磊 , 武佳琪 , 檀亚宁 . 基于 TextCNN 的用户评论情感极性判别 [J]. 电子世界 ,2019(3):48-50.
[19]LIU J X, ZHANG Y L. An attribute-weighted bayes classifier based on asymmetric correlation coefficient[J]. International journal of pattern recognition and artificial intelligence, 2020, 34(10): 2050025.
[20] DEVLIN J, CHANG M W, LEE K,et al.BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.Brussels: ACL, 2019: 4171-4186.

【作者简介】

李潇萌（2000—），女，山东济宁人，硕士研究生，研究方向：大数据挖掘。

赵庆峰（1974—），男，黑龙江齐齐哈尔人，博士，讲师，研究方向：情报管理。

张玉林（1979—），通信作者（email:zhangyulin@sdust.edu.cn），男，山东济宁人，博士，副教授，研究方向：大数据挖掘。

（收稿日期：2025-07-22 修回日期：2026-02-24）

（上接第 119 页）

漏检率达到 35% 以上。相比之下，本研究通过引入集成学习框架，构建了多层次的漏洞特征分析模型，最终将漏检率控制在 20% 以内，整体检测准确率提升至 92.6%，显著优于现有方法。这一改进源于集成模型对各类漏洞特征的自适应学习能力，以及对复杂漏洞模式的有效捕捉。

3 结语

本研究以数字化教学软件安全漏洞检测为切入点，探索了基于集成学习算法的安全漏洞检测方法，该方法引入无监督主成分分析方法对安全漏洞数据进行降维与分类，并结合静态漏洞特征和动态漏洞特征，能够发现潜在的安全漏洞类型，可实现实时、高精度的识别和响应。实验表明，该方法在真实教学软件数据集上的检测准确率较高，满足应用要求，本文研究可为教育软件安全领域提供一种低误报、高可解释性的智能检测手段，有助于构建一个安全可靠的数字教学环境，对推动智能教学生态健康发展具有重要实践价值。

参考文献：

[1] 程广振 . 基于多源数据融合的无线网络安全漏洞信息检测方法 [J]. 桂林航天工业学院学报 ,2025,30(3):433-440.
[2] 曹子亨 , 何立风 , 贾鸥 , 等 . 基于双向数据流分析与图抽象嵌入的漏洞检测方法 [J]. 计算机应用研究 ,2025,42(7):2176-2183.
[3] 辛瑞雯 , 高云 . 基于知识图谱技术的计算机网络链路漏洞检测研究 [J]. 现代电子技术 ,2025,48(1):97-101.
[4] 沈东舟 . 基于深度学习的港口信息系统网络安全漏洞检测与防御技术 [J]. 中国信息界 ,2024(9):228-230.
[5] 费圣翔 , 陈子龙 , 王冲 , 等 . 基于图神经网络的异构网络信息安全漏洞深度检测方法 [J]. 华南地震 ,2024,44(3):180-186.
[6] 刘悦 . 基于孤立森林算法的计算机网络信息安全漏洞检测方法 [J]. 信息与电脑 (理论版),2024,36(4):68-70.
[7] 张和伟 , 王奉章 . 基于被动分簇算法的即时通信网络安全漏洞检测方法 [J]. 智能计算机与应用 ,2023,13(7):119-122.
[8] 杨一健 , 李明 , 方赛银 . 基于信息熵的改进蚁群算法求解 TSP 问题 [J]. 计算机工程与设计 ,2024,45(9):2874-2881.
[9] 李刚 , 吴麒 , 王翔 , 等 . 基于样本信息熵辅助的深度强化学习抗干扰策略 [J]. 通信学报 ,2024,45(9):115-128.
[10] 黄鹤 , 黄佳慧 , 刘国权 , 等 . 采用混合策略联合优化的模糊 C- 均值聚类信息熵点云简化算法 [J]. 西安交通大学学报 , 2024,58(7):214-226.

【作者简介】

尚进（1992—），女，甘肃陇南人，本科，讲师，研究方向：教育信息化。

（收稿日期：2025-08-21 修回日期：2026-03-17）