doi: 10.7690/bgzdh.2025.06.011

基于分布式机器学习算法的科研审计系统安全漏洞识别方法

李俊奕[1]，肖亚纳[2]
（1. 广东省计算技术应用研究所，广州 510033；2. 广东省科技基础条件平台中心，广州 510033）

摘要：为解决科研审计系统存在安全性较差、精确率和召回率较低等问题，设计一种基于分布式机器学习算法的科研审计系统安全漏洞识别方法。采集科研审计系统用户数据，对用户节点数据进行分簇，并引入 k 近邻算法（k-nearest neighbor，KNN）概念建立科研审计系统网络分布式结构模型，将具有代表性和多样性的安全漏洞特征进行组合并分类，基于分布式机器学习算法在实际应用中进行安全漏洞识别。通过 2 种传统的安全漏洞识别方法进行对比。结果表明：该方法可以识别不同类型的安全漏洞，且准确率、精确率和召回率都有提高。

关键词：分布式机器学习算法；科研审计系统；安全漏洞识别；分布式结构模型；安全漏洞特征
中图分类号：TP393.08　　**文献标志码：**A

Security Vulnerability Identification Method of Scientific Research Audit System Based On Distributed Machine Learning Algorithm

Li Junyi[1], Xiao Ya′na[2]
(1. *Guangdong Institute of Computing Technology Application*, *Guangzhou* 510033, *China*;
2. *Guangdong Science and Technology Basic Conditions Platform Center*, *Guangzhou* 510033, *China*)

Abstract: In order to solve the problems of poor security, low precision and recall rate in scientific research audit system, a security vulnerability identification method of scientific research audit system based on distributed machine learning algorithm is designed. Collecting the user data of the scientific research audit system, clustering the user node data, introducing the concept of the *k*-nearest neighbor (KNN) algorithm to establish a network distributed structure model of the scientific research audit system, and combining and classifying the security vulnerability characteristics with representativeness and diversity. Based on distributed machine learning algorithm, security vulnerability identification is carried out in practical application. Two traditional security vulnerability identification methods are compared. The results show that the method can identify different types of security vulnerabilities, and the accuracy, precision and recall are improved.

Keywords: distributed machine learning algorithm; scientific research audit system; security vulnerability identification; distributed structure model; security vulnerability characteristics

0　引言

随着信息化建设的快速发展，我国科研项目管理的信息化程度不断提高。为加强科研项目和经费管理，规范科研管理秩序，提高科研审计效率，一般使用科研审计系统对项目数据进行管理。随着信息化建设的快速发展，我国高校科研审计系统在科研管理工作中发挥了越来越重要的作用。科研审计系统作为一种智能化、自动化的项目管理系统，在高校科研管理工作中起到了重要作用。由于高校科研审计系统具有使用范围广、用户数量大等特点，其内部存在着大量的安全漏洞[1-3]。例如，兵工领域的科研审计涉及军事科研项目的资金使用、管理制度、保密措施、知识产权保护和研究成果等，涉及国家机密和商业机密，其安全性问题引起了社会的高度关注。安全漏洞是指黑客利用技术漏洞或软件缺陷，以物理方式访问机器和硬件分析系统的控制权漏洞，窃取安全访问身份[4]。这些安全漏洞一旦被利用，就可能造成严重后果；因此，对高校科研审计系统中的安全漏洞进行识别和定位，具有十分重要的意义。

传统的安全漏洞识别方法主要有手工查询、静态分析和动态分析 3 种方式。其中，手工查询需要大量的人力进行手动操作，效率较低；静态分析是通过分析程序文件中的数据和关键字来识别程序中存在的安全漏洞，但这种方法需要耗费大量时间来进行人工操作，难以满足高校科研审计系统实时性强、实时性高的要求；动态分析则是通过程序执行过程中产生的数据来进行识别，但这种方法由于程序执行过程无法观察和控制而具有较大的局限性，

收稿日期：2024-08-21；**修回日期：**2024-09-25
基金项目：广东省科技计划项目(2020B1010010005)；广东省科技专项资金项目(210901164532767)
第一作者：李俊奕(1986—)，男，广东人。

识别性能较差。目前，机器学习技术已经被广泛应用于各个领域，其中针对安全漏洞识别问题，采用机器学习算法进行识别具有较高的精度和效率。笔者提出一种基于分布式机器学习算法的科研审计系统安全漏洞识别方法，在对科研审计系统进行需求分析和功能模块设计的基础上，设计系统功能模块和信息安全架构，并对相关技术进行研究。

1　科研审计系统安全漏洞识别方法研究

1.1　建立科研审计系统网络分布式结构模型

将高校科研审计系统的用户作为节点，每个节点都与其他节点相连接。为了解决数据规模大、数据处理复杂等问题，科研审计系统采用了分布式计算技术，通过将数据分布到不同的节点上进行处理。将以上节点进行分簇，能够为后续的分布式机器学习算法提供良好的环境。由于科研审计系统存在大量用户，而且这些用户之间是相互独立的[5]，因此其结构相对简单，系统主要由几个用户组成[6]。这些用户之间通过通信协议相互联系，在一定程度上提高了系统的效率。笔者对以上用户节点进行分簇，并引入 k 近邻算法(KNN)概念，k 最近邻分簇法中的核心思想，就是通过 k 个聚类中心与最近聚类中心距离进行比较。分簇后的科研审计系统网络分布式结构模型如图 1 所示。

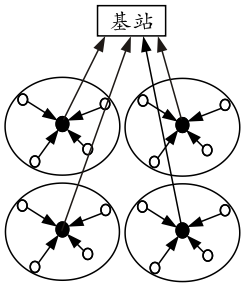


图 1　科研审计系统网络分布式分簇结构模型

科研审计系统是一个由多个客户端和一个服务器构成的分布式系统。通过对科研审计系统中的用户进行分析，可以得出该系统中存在着大量的安全漏洞。为了解决这一问题，将科研审计系统中的数据分布到不同的节点上进行处理，采用分布式机器学习算法进行识别。根据科研审计系统中的数据特点，可以将数据划分为若干个簇。这些簇中包含了大量有价值的信息，也是处理分析的重点对象。

在对科研审计系统中存在的安全漏洞进行识别时，可以将每个节点作为一个簇进行处理。同时根据不同簇内节点之间存在的关联关系，可以对单个用户或者多个用户进行关联分析，检测出本地安全漏洞。在进行安全漏洞识别时，可以根据不同节点间存在的关联关系，将其划分到不同的簇中进行处理。其中在对单个用户或者多个用户进行关联分析时，可以将其划分到同一个簇内进行处理。

1.2　安全漏洞特征选择

安全漏洞特征是指能够反映安全漏洞相关信息的特征集合，用于对安全漏洞进行识别和定位[7-9]。在安全漏洞特征选择的过程中，首先需要选择具有代表性和多样性的安全漏洞特征，其次需要将这些特征进行组合，最后将这些组合后的特征进行分类。入侵监测如图 2 所示。

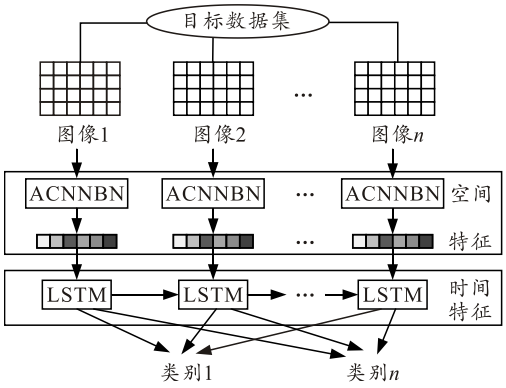


图 2　入侵检测

在安全漏洞特征选择时，首先需要确定选择的安全漏洞特征，然后再选择合适的分类算法[10]。笔者采用了一种基于特征重要性排序和 Bagging 算法相结合的方法。所谓 Bagging 算法，是一种基于层次聚类的机器学习算法，其核心思想是将数据集划分为多个簇，再对每个簇进行训练学习。在 Bagging 算法中，每个样本都会被赋予一个权重矩阵。然后通过迭代过程寻找最大权重矩阵。在寻找最大权重矩阵时，为保证不同样本对训练集的贡献程度一致，采用一种动态权重方法。每个样本都会得到一个动态权重矩阵。如果某个样本出现在了较高的权重区域内，则该样本被认为是较重要的样本；如果某个样本出现在了较低的权重区域内，则该样本被认为是较不重要的样本。

根据每个样本，设置一个新的权值矩阵。最后，在所有测试数据中抽取一定数量的样本作为训练集，将上述过程重复若干次。在迭代过程中，每次都会在新产生的权值矩阵中更新测试集中所有样本对测试集贡献程度的权重，并利用训练集对新产生权值矩阵进行训练；再从新产生权值矩阵中随机抽取一定数量的样本作为测试集。如此重复迭代，最终通过这种方式得到一组最优权重矩阵。

通常可以根据安全漏洞特征的重要性来选择特征[11-13]。对于安全漏洞特征重要程度的判断可以从以下 2 方面进行：1) 安全漏洞特征对于安全漏洞检测分类能力是否重要；2) 安全漏洞特征对于分类效果是否重要。

1.3 基于分布式机器学习算法的安全漏洞识别

随着分布式计算平台的发展，采用分布式机器学习算法对高校科研审计系统中存在的安全漏洞进行识别已经成为可能[14-15]。分布式机器学习算法是一种新兴的机器学习方法，利用分布式计算平台多个计算节点之间进行并行计算，从而实现多个节点同时对同一个训练集进行训练和学习。该方法不但可以提高识别效率，而且还可以保证识别结果的准确性和可靠性。采集各高校科研审计系统的安全漏洞信息，并对数据进行预处理；其工作过程是利用 n 维的特征向量来表示上节通过若干次迭代训练选择的安全漏洞的数据样本特征最优权重矩阵，即：

$$X=\{x_1, x_2, \cdots, x_n\}。 \tag{1}$$

上述特征集中，每个特征样本都在集合中通过各个元素表示出来，能够描述出特征样本不同属性的度量。假设以上样本可以进行划分成不同的类别，表示为：$C_1, C_2, \cdots, C_m$，对于一个给定的特征样本，将其分类到其中的一个漏洞类别 C_i 中，当存在：

$$P(C_i|X)>P(C_j|X),\ 1\leqslant j\leqslant m,\ j\neq i。 \tag{2}$$

式中 $P(C_i|X)$中的最大分类 C_i 为最大后验假设。根据分布式机器学习算法可知：

$$P(C_i|X)=P(X|C_i)P(C_i)/P(X)。 \tag{3}$$

数据中的所有类都是常数，因此在识别过程中，保证 $P(X|C_i)P(C_i)$的值最大。当类别的先验概率未知情况下，通过分布式机器学习算法的假设，将 $P(X|C_i)$的值设为最大，即可使用先验概率计算漏洞概率。在识别过程中，其流程如图 3 所示。

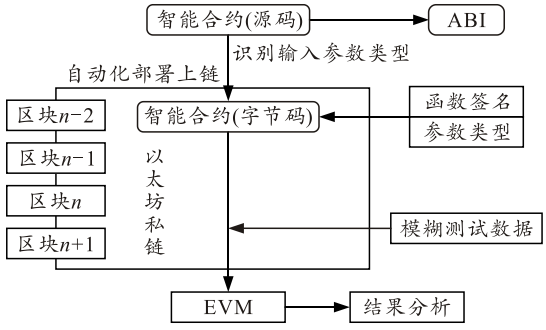


图 3 安全漏洞检测流程

图 3 中，源码智能合约是一种在区块链网络上执行的智能合约，通过编译器转换为机器可执行的代码，可以被区块链上的参与者查看。识别输入参数类型，输入应用程序二进制接口(application binary interface，ABI)，定义合约方法和参数的规范。通过多个区块将智能合约自动化部署到以太坊链形成包含函数签名和参数类型的字节码。对其进行模糊测试，输入到以太坊虚拟机(Ethereum virtual machine，EVM)中，分析并执行以太坊上的智能合约代码，对数据进行分类，根据分类结果确定哪些数据是安全漏洞，哪些数据是正常的；然后，采用分布式机器学习算法进行安全漏洞识别。至此完成基于分布式机器学习算法的科研审计系统安全漏洞识别方法的研究与设计。

2 实验

2.1 实验对象与过程

为了验证笔者设计的基于分布式机器学习算法的科研审计系统安全漏洞识别方法的有效性，通过不同识别方法的准确率、精确率、召回率和检测效率 4 方面进行判定。选择某军工科研项目使用资金的科研审计系统作为安全漏洞识别的测试对象，考虑该军工系统的安全隐私性，在该系统的仿真模型中设置本地安全漏洞，嵌入包含缺陷追踪的数据集为测试数据集，如表 1 所示。

表 1 测试数据集

数据来源	安全漏洞报告数量	非安全漏洞报告数量
谷歌浏览器	3 159	8 947
猎人	1 248	7 593
流加密	1 504	6 236
常见、公开漏洞数据库	3 581	6 215
岩浆-M	2 911	7 776

收集到 49 170 个缺陷报告，包含 12 403 个安全漏洞报告和 36 767 个非安全漏洞报告，随机选取 70%的数据做测试训练，30%为实验数据。在以上的实验数据集下，分别使用不同的安全漏洞识别方法进行测试。笔者利用式(3)计算漏洞类别的先验概率，识别出安全漏洞类别，作为测试结果。为了将式(3)识别结果的测试结果进行量化，选择准确率、精确率、召回率和拟合平均值作为评价指标。准确率计算公式如下：

$$Ac=(TP+TN)/(TP+TN+FP+FN)。 \tag{4}$$

式中：TP 为识别正确报告数量；TN 为误判为安全漏洞报告数量；FP 为安全漏洞报告误判为非安全漏洞报告数量；FN 为正确识别非安全漏洞报告数量。

测试中精确率计算公式如下：

$$P_r=TP/(TP+FP)。 \tag{5}$$

测试中召回率计算公式如下：

$$R_e=\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})。\tag{6}$$

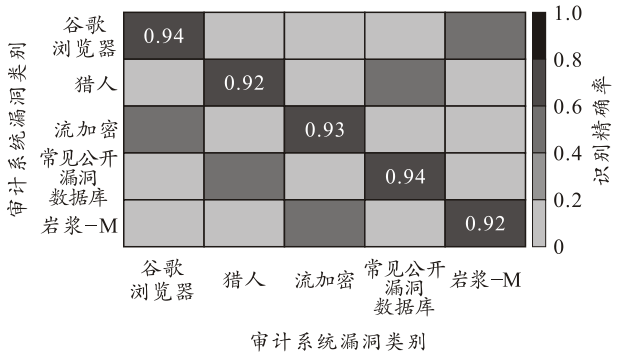
拟合平均值的计算公式如下：

$$F_1=2P_rR_e/(P_r+R_e)。\tag{7}$$

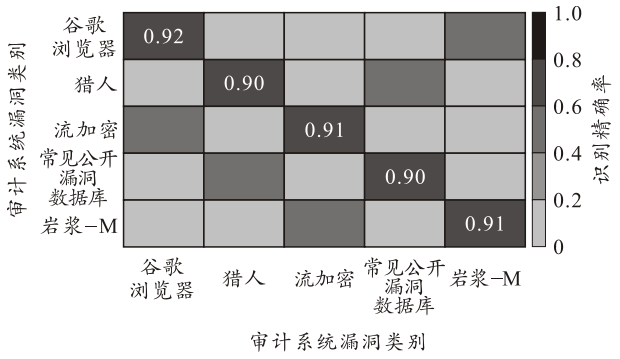
在不同的安全漏洞识别方法中，F_1 值越高，表示稳定性越高。在以上的测试环境下，分别使用本文中设计的基于分布式机器学习算法的科研审计系统安全漏洞识别方法、基于源代码扫描的识别方法、基于反汇编扫描的识别方法、基于环境错误录入的识别方法共同进行测试，并将测试结果进行对比与分析。

2.2　实验结果对比与分析

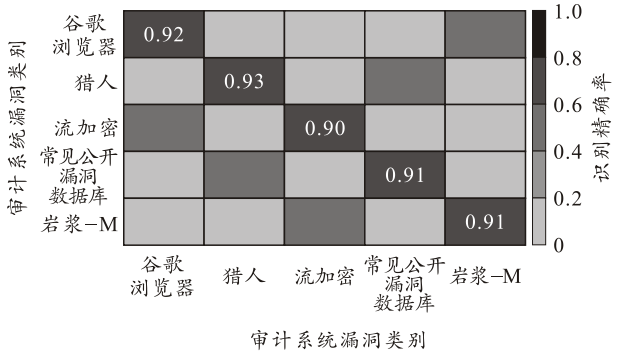
在测试过程中，本文中方法得到的识别统计结果如图 4 所示。



(a) 本文中方法识别准确率



(b) 本文中方法识别精确率



(c) 本文中方法识别召回率

图 4　本文中安全漏洞识别方法检测结果

从图 4 可知：本文中方法具有识别准确率、精确率和召回率较高的特点。

由于程序运行是动态变化的，常规科研审计系统安全漏洞识别一般分为静态检测和动态检测 2 种。代码检测更适用于静态检测，动态检测效果受限，无法实现精准检测，对此，设置源代码扫描和反汇编扫描 2 种静态检测方法，环境错误录入为动态识别方法，通过实验结果对比选出最优的检测方法。分析不同方法的检测效率对比结果，如图 5 所示。

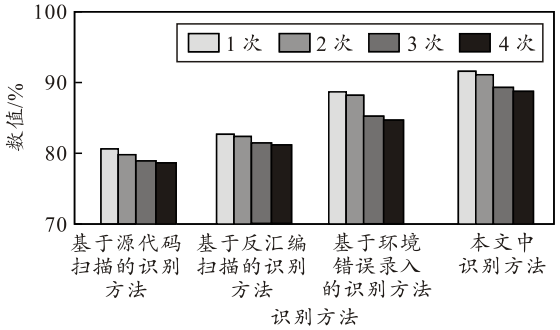


图 5　不同方法的检测效率对比结果

从图 5 可知：其他方法对检测环境要求较高，检测速度较慢，且容易受到恶意攻击，不具有普遍性；但是本文中方法在静态、动态方面均表现良好，兼容性好，综合性能优于其他检测方法。

3　结束语

笔者针对科研审计系统可能存在的安全漏洞，分类融合后进行识别。从检测识别结果可以看出：融合之后安全漏洞特征更具有代表性，检测效率及性能包括准确率、精确率和召回率都有提高，较单一安全漏洞识别效果更好。由综合实验结果可知，笔者设计的基于分布式机器学习算法的科研审计系统安全漏洞识别方法，在检测效率及准确率等方面均表现优异，具有在军工科研审计系统应用的高可行性及稳定性。

参考文献：

[1] 徐晓君, 常会丽. 多线程交互学习软件系统安全漏洞自动化检测[J]. 计算机仿真, 2022, 39(4): 335-340.

[2] 程若曦, 王凯, 焦健, 等. 基于知识库的配网网络化下系统全过程安全态势识别方法[J]. 科技通报, 2021, 37(9): 46-51.

[3] 王剑, 匡洪宇, 李瑞林, 等. 基于 CNN-GAP 可解释性模型的软件源码安全漏洞检测方法[J]. 电子与信息学报, 2022, 44(7): 2568-2575.

（下转第 81 页）