Hans Journal of Data Mining 数据挖掘, 2024, 14(3), 172-178
Published Online July 2024 in Hans. https://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2024.143016

Hans汉斯

面向企业数据安全场景基于孤立森林算法的数据访问异常检测与告警收敛方法

高玉涛，张若愚，赵旭栋，田晓伟，黄成祁，姚雪茜，李　刚

中国联合网络通信有限公司软件研究院，北京

收稿日期：2024年6月8日；录用日期：2024年7月8日；发布日期：2024年7月17日

摘　要

随着数字经济的快速发展，数据安全成为企业关注的焦点。本文提出了一种基于孤立森林算法的异常行为告警收敛方法，旨在提升企业数据安全合规及业务安全。通过用户实体行为分析(UEBA)技术建立用户访问行为基线，并监测实时行为与基线的偏差，以识别异常行为。然而，UEBA技术在异常告警收敛方面存在不足，导致告警数量庞大，难以维护。为此，本文引入孤立森林算法，对UEBA检出的异常样本进行进一步分析和收敛，以提高告警质量，减少告警数量，实现可维护的异常告警。文章详细介绍了基于UEBA的数据访问异常行为检测技术，存在的问题及困难，以及孤立森林算法的原理和优势。此外，还探讨了孤立森林算法在数据访问异常行为检测场景的告警收敛应用，包括数据准备、模型训练、异常分析、样本收敛和结果解释等步骤。最终，本文总结了利用孤立森林算法进行异常分析和告警收敛的有效性，并对未来的发展方向进行了展望。

关键词

用户实体行为分析，异常检测，孤立森林算法

Data Access Anomaly Detection and Alarm Aggregation Method Based on Isolation Forest Algorithm for Enterprise Data Security Scenarios

Yutao Gao, Ruoyu Zhang, Xudong Zhao, Xiaowei Tian, Chengqi Huang, Xueqian Yao, Gang Li

China Unicom Software Institute, Beijing

Received: Jun. 8th, 2024; accepted: Jul. 8th, 2024; published: Jul. 17th, 2024

文章引用: 高玉涛, 张若愚, 赵旭栋, 田晓伟, 黄成祁, 姚雪茜, 李刚. 面向企业数据安全场景基于孤立森林算法的数据访问异常检测与告警收敛方法[J]. 数据挖掘, 2024, 14(3): 172-178. DOI: 10.12677/hjdm.2024.143016

Abstract

With the rapid development of the digital economy, data security has become a focal point for enterprises. This paper proposes an anomaly alert convergence method based on the Isolation Forest algorithm to enhance enterprise data security compliance and business security. By establishing a baseline of user access behavior through User and Entity Behavior Analytics (UEBA) technology and monitoring real-time behavior deviations from the baseline, abnormal behaviors are identified. However, UEBA technology has shortcomings in alert convergence, leading to a large number of alerts that are difficult to maintain. To address this, this paper introduces the Isolation Forest algorithm to further analyze and converge the abnormal samples detected by UEBA, thereby improving alert quality, reducing the number of alerts, and achieving maintainable anomaly alerts. The article provides a detailed introduction to the data access anomaly behavior detection technology based on UEBA, the existing problems and difficulties, as well as the principles and advantages of the Isolation Forest algorithm. Additionally, it discusses the application of the Isolation Forest algorithm in alert convergence for data access anomaly behavior detection scenarios, including steps such as data preparation, model training, anomaly analysis, sample convergence, and result interpretation. Ultimately, the paper summarizes the effectiveness of using the Isolation Forest algorithm for anomaly analysis and alert convergence and provides an outlook on future development directions.

Keywords

UEBA, Anomaly Detection, Isolation Forest Algorithm

1. 引言

随着数字经济的飞速发展，我国不断加强和完善数据安全立法，并将数据提升到生产要素地位。企业数据安全工作面临合法合规及业务安全双重挑战，数据泄露风险检测成为保障企业数据安全合规及业务安全的重要的技术手段，面向数据访问场景的异常行为检测技术成为最重要的数据泄露检测手段。技术上通常基于用户实体行为分析(User and Entity Behavior Analytics, UEBA)技术来落地。UEBA 通过收集和分析用户访问行为数据，建立用户访问行为基线，然后监测实时行为与基线的偏差。这些偏差可能表明异常行为的发生。这种基于行为基线的检测方法在异常告警收敛方面表现并不理想，带来的直接后果就是异常告警数量大，难于维护。本文通过引入孤立森林算法(Isolation Forest)对 UEBA 检出的异常样本作进一步分析和收敛，从而提升告警质量，压缩告警数量，达到异常告警可维护的效果。

在数字化转型的浪潮中，企业数据安全的重要性日益凸显，数据泄露事件频发，给企业带来了巨大的经济损失和信誉损害。因此，研究和开发有效的数据安全防护技术，特别是针对异常行为的检测与告警收敛技术，具有重要的现实意义和迫切性主要体现在以下几个方面：

数据泄露风险的增加：随着企业数据量的爆炸性增长，数据泄露的风险也随之增加。传统的安全防护措施往往关注于外部攻击，而忽视了内部威胁，如员工误操作、内部人员滥用权限等。因此，研究面向内部威胁的异常行为检测技术显得尤为重要。

现有技术局限性：尽管 UEBA 技术在行为基线建立和实时监测方面取得了一定的成效，但其在异常告警收敛方面存在明显的不足，如误报率高、告警数量庞大等，导致安全团队难以有效应对。因此，研究新的告警收敛方法，以提高异常检测的准确性和效率，是当前数据安全领域的一个迫切需求。

动态适应性的需求：随着企业业务的不断变化，用户行为模式也在不断演变。研究一种能够动态适应用户行为变化的告警收敛方法，对于企业持续的数据安全防护至关重要。

法规遵从的压力：在数据安全立法日益严格的背景下，企业面临着合法合规的压力。亟需寻找能够帮助企业更好地遵守相关法律法规的技术抓手，减少因数据泄露导致的法律风险。

朱丽琴(2020)在研究中探索了基于孤立森林的入侵检测方法，以及如何通过检测网络流量中的异常行为来识别潜在的入侵行为[1]；王诚和狄萱(2021)侧重研究了孤立森林算法实现其并行化，在提高算法的效率和可扩展性，使其能够处理大规模数据集创造了条件[2]；祝诚勇、黄鹏翔和李理敏(2024)提出了一种基于专家反馈的广义孤立森林异常检测算法涉及到算法的改进，通过集成专家知识来提高异常检测的准确性[3]；周杨、王春林和郭锐(2023)探讨了基于随机森林算法的数据中心运维异常告警方法，虽然主要关注的是随机森林，但考虑到随机森林与孤立森林在决策树领域的相似性，该研究可能间接为孤立森林的应用提供了参考或对比[4]。这些学者在不同领域以不同角度对孤立森林算法进行了研究和应用，包括入侵检测、算法优化、专家系统集成以及数据中心运维等。这些研究成果不仅推动了孤立森林算法本身的发展，也为相关领域的异常检测提供了新的视角和方法。

面向企业数据安全场景的基于孤立森林算法的数据访问异常行为告警收敛方法的研究，不仅能够提升企业的技术能力，符合技术发展趋势，而且具有重要的行业推广价值和实践意义。通过本研究，企业可以更有效地应对数据安全挑战，实现数据安全防护的持续优化和迭代，从而在企业数字化转型的浪潮中保持优势，实现可持续发展。

2. 基于 UEBA 的数据访问异常行为检测技术

2.1. EBA 技术简介

UEBA 是一种先进的安全分析技术，它通过收集、分析和评估用户及实体的行为数据，以识别潜在的安全威胁和内部风险。在数据访问异常行为检测场景中，UEBA 技术发挥着至关重要的作用，它能够帮助组织在海量的用户访问行为中自动学习正常的行为基线，并依此检测出偏离正常行为基线的异常行为。以下是 UEBA 在数据访问异常行为检测中的应用介绍：

2.2. UEBA 技术在数据访问异常行为检测场景的应用

数据收集：UEBA 系统首先需要从不同的数据源收集用户和实体的行为数据。这些数据源可能包括身份认证系统、网络日志、访问控制系统、企业资源规划(ERP)系统等。UEBA 平台将这些分散的数据整合到一个中央数据库中，为后续的分析提供统一的数据视图。

建立行为基线：通过对数据访问历史行为数据分析，UEBA 能够建立正常行为基线模型。这些模型可以基于统计分析、机器学习算法等方法，对用户和实体的正常访问模式、操作习惯等进行建模。基线模型是识别异常行为的关键，它为后续的异常检测(Anomaly Detection)提供了参照标准。

实时行为监测与分析：UEBA 会实时监测和分析用户及实体的数据访问行为。当检测到与基线模型不符的行为时，系统会标记为潜在的异常行为。这些异常行为可能包括非正常工作时间的数据访问、对敏感数据的频繁查询、不寻常的文件下载等。

异常行为评估与告警：一旦识别出潜在的异常行为，UEBA 系统会对其进行进一步的评估。评估过程可能涉及行为的严重性、频率、影响范围等多个维度。根据评估结果，系统会生成告警信息，并通过

合适的渠道通知给安全分析师或管理人员。

行为基线动态更新：UEBA 系统会不断学习和适应用户及实体的行为变化。通过持续收集和分析数据，系统能够更新基线模型，提高异常检测的准确性。

2.3. 存在的问题及困难

UEBA 系统依托行为基线作为异常判别的最重要依据，基线模型的宽松或收紧不可避免导致其检出异常样本不可避免存在大的误报或漏报，从而给异常告警维护带来极大压力。

3. 孤立森林算法介绍

3.1. 孤立森林算法介绍

孤立森林(Isolation Forest)算法是一种用于异常检测的机器学习算法。该算法基于随机森林原理，通过随机选择特征和随机选择切分点的方式，利用二叉树将数据点逐一孤立开来，从而识别出异常数据点。该算法特别适用于高维数据集，并且在处理大规模数据时具有较好的效率。

3.2. 算法原理

孤立森林算法通过构建多棵孤立树(Isolation Tree)，每棵树都试图将数据点分隔开来，使得每个数据点尽可能地孤立。在构建每棵树的过程中，算法会随机选择一个特征和一个切分值，然后将数据集划分为两个子集，一个包含所有小于切分值的数据集，另一个包含所有大于切分值的数据集。这个过程会递归地进行，直到每个数据点都被孤立到单独的节点中，或者达到预设的树深度。

3.3. 异常检测

在孤立森林中，正常数据点通常需要更多的划分才能被孤立，因为它们在特征空间中是密集分布的。相反，异常数据点由于数量较少且分布稀疏，往往可以更快地被孤立。因此，通过计算数据点在每棵树的孤立路径长度(即从根节点到叶节点的分裂次数)，可以作为判断数据点是否为异常的依据。路径越短，数据点是异常的可能性就越大。

3.4. 算法优势

高维数据处理：孤立森林算法不需要对数据进行任何形式的预处理，如特征缩放或降维，使其非常适合处理高维数据集。

可扩展性：算法能够有效地处理大规模数据集，因为它不需要存储整个数据集，只需要在构建每棵树时访问数据点。

异常检测性能：孤立森林在许多数据集上都显示出了良好的异常检测性能，尤其是在检测子空间异常时[5]。

4. 孤立森林算法在数据访问异常行为检测场景的告警收敛应用

为了利用孤立森林算法对 UEBA 输出的异常样本进行进一步的异常分析和异常样本收敛，我们将从数据准备、模型训练、异常分析、样本收敛和结果解释等细分工作来展开。

4.1. 数据准备

数据收集：从 UEBA 系统中收集已经标记为异常的用户行为数据样本。这些数据可能包括用户登录时间、账号、IP、访问时间、访问频率、请求数据量、权限变更等信息。

数据预处理：由于 UEBA 本身会对用户的数据访问行为数据作预处理，本方案仅对 UEBA 输出的异常样本作分析，因此不涉及数据预处理。

特征选取：选择与异常行为强相关的特征，如访问频率、访问时间、访问数据量等，并对选取特征设置不同权重，不同特征在不同业务场景中的异常关切程度可能不同，特征的权重可根据业务需求灵活调整。

4.2. 模型训练

初始化配置：配置孤立森林算法参数，如孤立树的数量(参考值 80~100 棵)、每棵树的最大深度(参考值 6~8 层)、用于划分节点的随机特征数量(参考值 1~3 个)等。参数的选择应基于数据的特性和先验知识，特征的选取应考虑不同特征的权重保证高权重的特征被随机选取的概率更大。

模型训练：随机选取 UEBA 输出的异常样本数据训练孤立森林模型。在训练过程中，算法将构建指定数量的孤立树，每棵树都尝试孤立特征值异常的样本数据。

4.3. 异常分析

异常评分：利用训练好的孤立树，对每个异常样本计算异常分数，该分数反映了样本被孤立的快慢程度[6]。通常越快被孤立的样本(即路径较短的样本)具有越高的异常分数。

异常模式识别：分析异常分数分布，识别异常模式。例如，如果某个用户近期集中存在敏感数据过频、过量访问行为，则通过异常输出可判断是否存在离职人员数据泄露风险。

4.4. 告警收敛

聚类分析：对异常样本进行聚类分析，将相似的异常行为归为一类[7]。这有助于识别和分析重复出现的异常样本，并减少单独处理每个异常样本的工作量。

优先级设置：根据异常分数和聚类结果，为每个异常告警分配优先级。高优先级的告警应首先被安全分析师审查。

4.5. 模型迭代与优化

结果反馈：支持告警确认结果反馈给模型，据此优化孤立森林算法的特征筛查及权重设置机制，并持续提升告警收敛效果。

模型优化：定期更新孤立森林模型，以适应不同业务场景下用户行为模式的变化和新的敏感数据泄露威胁。

5. 研究成果创新特点和价值

在企业数据安全领域，异常行为的检测和告警收敛是保障数据安全的关键环节。传统的用户实体行为分析(UEBA)技术虽然在行为基线建立和实时监测方面发挥了重要作用，但在告警收敛方面却存在效率不高和误报率高的问题。针对这一挑战，本文提出的基于孤立森林算法的告警收敛方法具有以下几个创新特点：

算法的高效性与适应性：孤立森林算法本身具有处理高维数据集的能力，无需对数据进行复杂的预处理，这使得算法能够快速适应企业中多变的数据环境。此外，算法的随机性保证了其在面对不同数据源和行为模式时的适应性和泛化能力。

异常检测的精确性：通过构建多棵孤立树，孤立森林算法能够有效地识别出异常数据点。在数据访问异常行为检测中，该算法能够快速区分正常行为与异常行为，从而提高异常检测的精确度，减少误报

和漏报。

告警收敛的智能化：本文提出的方法不仅关注异常行为的检测，更重视告警的智能收敛。通过聚类分析和异常评分机制，可以将相似的异常行为归类，并根据异常的严重程度分配不同的优先级，使得安全分析师能够更高效地处理告警。

特征权重的动态调整：在特征选取阶段，本文的方法允许根据不同业务场景的需求对特征进行权重设置和调整。这种灵活性使得算法能够根据不同的安全需求和威胁模型，动态调整关注的重点，提高检测的针对性和有效性。

模型的持续优化：通过结果反馈机制，安全团队可以根据告警确认的结果对模型进行迭代优化。这种持续的学习过程保证了模型能够适应新的威胁模式和用户行为的变化，从而不断提升告警收敛的效果。

业务场景的广泛适用性：孤立森林算法的引入不仅限于某一特定的数据访问场景，而是可以广泛应用于各种企业数据安全场景中。无论是内部威胁检测、数据泄露预防还是网络安全态势感知，该算法都能够提供有效的支持。

风险评估的全面性：本文的方法不仅关注单一的异常行为，还通过异常模式识别，对潜在的安全风险进行全面评估。这有助于企业从宏观角度理解和应对数据安全威胁，而不仅仅是解决眼前的告警问题。

综上所述，本文提出的基于孤立森林算法的数据访问异常行为告警收敛方法，通过其高效性、精确性、智能化、动态调整、持续优化、广泛适用性和全面性等特点，为企业数据安全提供了一种创新的解决方案。这种方法有望在实际应用中显著提升企业对异常行为的响应速度和处理效率，从而更好地保护企业的数据资产。

6. 总结及展望

在当前企业数据安全领域，随着网络攻击手段的不断演变和内部威胁的日益增加，传统的安全防护措施已经难以满足企业的需求。用户实体行为分析(UEBA)技术虽然在行为基线建立和实时监测方面发挥了重要作用，但在异常告警收敛方面存在明显的不足，如误报率高、告警数量庞大等，导致安全团队难以有效应对。因此，探索一种新的异常行为告警收敛方法，对于提升企业数据安全防护能力具有重要的现实意义。

本文提出的基于孤立森林算法的异常行为告警收敛方法，通过以下几个方面对企业数据安全防护进行了创新和提升：

算法优势的充分利用：孤立森林算法作为一种高效的异常检测算法，其在处理高维数据集、无需预处理、快速隔离异常点等方面具有明显优势。本文将其应用于企业数据安全领域，充分发挥了算法的优势，提高了异常检测的效率和准确性。

告警收敛的智能化：通过聚类分析、异常评分、优先级设置等方法，本文提出的方案实现了告警的智能收敛，将相似的异常行为归类，根据异常的严重程度分配不同的优先级，使得安全团队能够更加高效地处理告警。

特征权重的动态调整：本文的方法允许根据不同业务场景的需求对特征进行权重设置和调整，这种灵活性使得算法能够根据不同的安全需求和威胁模型，动态调整关注的重点，提高检测的针对性和有效性。

模型的持续优化：通过结果反馈机制，安全团队可以根据告警确认的结果对模型进行迭代优化。这种持续的学习过程保证了模型能够适应新的威胁模式和用户行为的变化，从而不断提升告警收敛的效果。

业务场景的广泛适用性：孤立森林算法的引入不仅限于某一特定的数据访问场景，而是可以广泛应用于各种企业数据安全场景中，如内部威胁检测、数据泄露预防、网络安全态势感知等，提供了一种通用的解决方案。

展望未来，随着技术的发展和威胁环境的变化，基于孤立森林算法的告警收敛方法仍有很大的优化和提升空间：

算法性能的进一步提升：通过算法优化、参数调优等手段，可以进一步提升孤立森林算法在异常检测和告警收敛方面的性能，如提高异常检测的准确率、减少计算资源的消耗等。

多源数据融合与分析：未来的研究可以探索如何将孤立森林算法应用于多源数据的融合与分析，如结合网络流量、系统日志、用户行为等多种数据源，以获得更全面和准确的异常检测结果。

主动防御机制的构建：除了异常检测和告警收敛，未来的研究可以探索如何利用孤立森林算法构建主动防御机制，如基于异常预测的动态访问控制、基于异常模式识别的威胁狩猎等。

人工智能技术的融合应用：结合机器学习、深度学习等人工智能技术，可以进一步提升孤立森林算法的智能化水平，如利用深度学习自动提取特征、利用强化学习优化模型参数等。

安全运营的自动化与智能化：未来的研究可以探索如何将孤立森林算法应用于安全运营的自动化与智能化，如自动生成安全报告、自动响应安全事件等，以进一步提升安全团队的工作效率。

法规遵从与伦理问题的考量：随着数据安全立法的日益严格，未来的研究需要考虑如何使基于孤立森林算法的告警收敛方法符合相关法律法规的要求，同时处理好数据隐私保护、算法透明度等伦理问题。

跨学科研究的推进：数据安全是一个涉及计算机科学、网络安全、管理学等多个学科的综合性问题。未来的研究可以加强跨学科的合作与交流，如结合管理学的理论优化安全策略、结合心理学的理论分析用户行为等。

国际合作与交流：数据安全是全球性的挑战，需要各国的共同努力。未来的研究可以加强国际合作与交流，共享数据安全的最佳实践，共同应对跨国网络攻击等安全威胁。

综上所述，基于孤立森林算法的异常行为告警收敛方法为企业数据安全防护提供了一种新的解决方案。通过持续的研究和优化，该方法有望在提升企业数据安全防护能力、应对复杂安全威胁等方面发挥更大的作用。同时，未来的研究需要从多个角度出发，不断探索和创新，以适应数据安全不断变化的技术环境和安全需求。

参考文献

[1] 朱丽琴. 基于孤立森林的入侵检测方法研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工程大学, 2020.

[2] 王诚, 狄萱. 孤立森林算法研究及并行化实现[J]. 计算机技术与发展, 2021, 31(6): 13-18.

[3] 祝诚勇, 黄鹏翔, 李理敏. 基于专家反馈的广义孤立森林异常检测算法[J]. 计算机应用研究, 2024, 41(1): 88-93.

[4] 周杨, 王春林, 郭锐. 基于随机森林算法的数据中心运维异常告警方法[J]. 现代电子技术, 2023, 46(8): 143-148.

[5] 钱瑞祥. 基于孤立森林的感知数据异常检测方法研究[D]: [硕士学位论文]. 杭州: 杭州电子科技大学, 2022.

[6] 林国顺, 王野. 基于自编码器-孤立森林的网购消费者异常行为检测[J]. 计算机应用与软件, 2022, 39(2): 253-258.

[7] 吴欣然, 张凌, 顾淼. 基于孤立森林算法的统计报表异常数据检测[J]. 信息技术与信息化, 2023(12): 208-211.