基于改进关联聚类算法的网络异常信息挖掘研究

卜 浏[1]
BU Liu

摘 要 由于网络信息量庞大且复杂多变，网络异常行为往往隐藏在大量正常数据中，且其表现形式多样，边界模糊。传统的聚类算法无法有效区分异常点与正常点，导致误判和漏判，文章提出了基于改进关联聚类算法的网络异常信息挖掘方法。首先通过一系列数据预处理步骤，包括数据清洗、归一化及数值化转换，确保网络信息数据的高质量与可用性，随后利用预处理后的数据构建了网络异常信息库，引入改进的关联聚类算法对该库进行深度挖掘。实验结果表明，这一方法显著提升了网络异常信息挖掘的精准度和效率，有效减少了数据缺失问题，为网络安全防护提供了有力的技术支持。

关键词 改进关联聚类算法；网络异常信息挖掘；关联规则；关联强度

doi：10.3969/j.issn.1672-9528.2024.11.037

0 引言

随着网络技术的进步，网络空间的安全性与稳定性正面临着日益严峻的挑战。网络异常信息挖掘作为一道坚实的防线，对于及时发现并应对潜在的网络威胁、保障网络安全具有不可替代的作用。近年来，网络异常信息挖掘这一领域因其重要性和挑战性，吸引了众多研究者的关注与投入。在现有研究中，文献 [1] 提出了基于支持向量机的通信网络异常流量数据挖掘方法，SVM 虽然能够学习到一个分类超平面来区分不同类别的数据，但在面对高度重叠或边界模糊的流量数据时，其分类效果可能会受到限制，导致挖掘质量下降。文献 [2] 提出了基于深度集成学习的社交网络异常数据挖掘算法。虽然深度集成学习通过集成多个模型来提升对复杂数据的处理能力，但是异常行为与正常行为之间的界限不明显时，模型可能难以准确地区分异常点与正常点。文献 [3] 提出了基于关联规则的网络信息数据挖掘方法。关联规则挖掘算法更倾向于捕捉出现频率较高的正常行为模式而忽略异常行为，导致算法无法有效地识别出隐藏在大量正常数据中的

1. 江苏联合职业技术学院 江苏南京 210000

[8] 白伊史，翟海霞，刘园．一种混合的动态社区发现算法 [J]. 小型微型计算机系统，2023, 44(4) :773-778.

[9] 汪焱，黄发良，元昌安．基于标签影响力的半同步社区发现算法 [J]. 计算机应用，2016,36(6):1573-1587.

[10] 王高飞．内容相似度的微博兴趣社区发现方法研究 [D]. 太原：太原理工大学，2018.

[11] LI P Z, HUANG L, WANG C D, et al. EdMot: an edge enhancement approach for motif-aware community detection[C]//KDD'19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, New York: ACM, 2019: 479-487.

[12] 郑文萍，张浩杰，王杰．基于稠密子图的社区发现算法 [J]. 智能系统学报，2016,11(3):426-432.

[13] LUSSEAU D. The emergent properties of a dolphin social network[J]. Proceedings of the royal society of london. series b: biological sciences, 2003, 270(S2): 186-188.

[14] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. Proceedings of the national academy of sciences of the United States of America, 2002, 99(12):7821-7826.

【作者简介】

彭慧豪（1995—），通信作者（email: penghhedu@163.com），男，河南许昌人，硕士研究生，助教，研究方向：机器学习、复杂网络及社区检测。

张怡欣（1996—），女，江西抚州人，硕士研究生，助教，研究方向：智能信息处理及应用。

李文豪（1996—），男，江西抚州人，硕士研究生，助教，研究方向：管理科学与工程。

刘天宇（1994—），男，江西抚州人，硕士研究生，助教，研究方向：地理信息系统开发。

（收稿日期：2024-07-29）

异常行为。文献 [4] 提出了网络安全中计算机文本信息挖掘方法。但是在预处理过程中未能有效去除噪声或错误处理了某些关键信息，这导致分类挖掘效果不佳。

鉴于网络信息异常行为的表现形式多样且边界模糊的特征，本文提出了一种基于改进关联聚类算法的数据挖掘方法，旨在更准确地识别网络中的异常行为。

1 网络信息预处理

数据清洗作为网络信息预处理流程的开端，其核心在于识别和修正网络数据集中的无效或错误记录 [5]。首先，本文定义一系列布尔函数 IsValid(x_i)，其中，i 表示清洗规则的索引，x 表示网络日志记录。这些函数基于时间戳的合理性、IP 地址的有效性、请求字段的完整性等具体规则，对每条记录进行逐一评估。通过应用这些函数，即 $\mathrm{IsValid}(x_i)=\begin{cases}\mathrm{True}\\ \mathrm{False}\end{cases}$，能够筛选出符合所有规则的有效记录，从而剔除无效或错误的数据，保留纯净且有用的数据集。数据清洗过程可以表示为：

$$Q'=\mathrm{Clear}(Q)=\left\{x_i \mid \mathrm{IsValid}(x_i)=\mathrm{True}, x_i\in Q\right\} \tag{1}$$

式中：Q 表示原始网络信息，Clear 表示原始网络信息的清洗过程，Q' 表示清洗后的数据，当记录 x 的第 i 个特征满足特定条件时，返回 True。

为了消除量纲差异对聚类分析的不利影响，本文采取归一化处理技术，通过按比例缩放数据的特征值 [6]，使得所有特征都转换到同一尺度上，从而在后续聚类过程中能够公平地贡献其信息，避免某些特征因量纲过大而主导聚类结果。具体而言，本文采用最小 - 最大归一化方法来实现网络信息数据的归一化。假设 f_{ij} 表示 Q' 中第 i 条记录中第 j 个特征的值，F_j 表示第 j 个特征在所有记录中的值集合，则归一化后的网络数据特征 f'_{ij} 可以表示为：

$$f'_{ij}=\frac{f_{ij}-\min_j(F_j)}{\max_j(F_j)-\min_j(F_j)} \tag{2}$$

式中：$\min_j(F_j)$ 和 $\max_j(F_j)$ 分别表示特征 j 在所有记录中的最小值和最大值。

采用独热编码技术，将分类数据转换为聚类算法可处理的数值型数据。设分类特征 C 有 k 个类别，则转换后的特征矩阵 F' 可以表示为：

$$F'=\mathrm{OneHot}(C)=\begin{bmatrix} f'_{11} & f'_{12} & \cdots & f'_{1k}\\ f'_{21} & f'_{22} & \cdots & f'_{2k}\\ \vdots & \vdots & \ddots & \vdots\\ f'_{n1} & f'_{n2} & \cdots & f'_{nk}\end{bmatrix} \tag{3}$$

利用式(3)的转换过程不仅能够保留分类信息的完整性，还能使后续聚类过程深入挖掘数据间的内在联系，进一步提升网络异常信息挖掘的精度和深度。通过上述设计的网络信息预处理步骤，提升网络信息数据质量，为后续网络异常信息挖掘奠定了坚实的数据基础。

2 基于改进关联规则算法的网络异常信息库构建

在网络信息预处理的基础上，为了更有效地挖掘网络中的异常信息，利用改进关联规则算法构建一个全面且高效的网络异常信息库，提升关联规则挖掘的效率。网络异常信息库构建流程如图 1 所示。

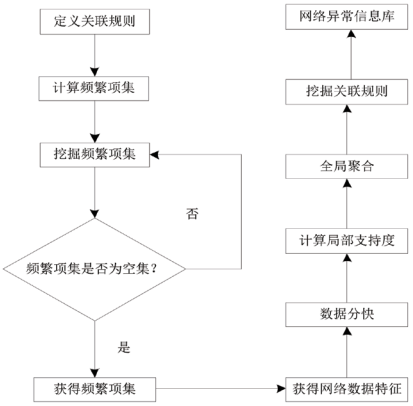

图 1 网络异常信息库构建流程图

首先，本文定义关联规则的基本形式，用于精确描述网络行为特征与异常事件之间的潜在关系。关联规则的形式化表示为：

$$X\rightarrow Y \tag{4}$$

式中：X 表示网络行为特征集合，Y 表示在给定特征组合下可能发生的异常事件。

为了提升关联规则挖掘的效率和准确性，本文采用改进的 Apriori 数据挖掘技术通过迭代搜索数据中的频繁项集，可以减少候选集的大小，从而提高挖掘效率。利用 Apriori 算法处理预处理后的海量网络数据，以识别出频繁出现的特征组合。假设频繁项集为 $I=\{i_1,i_2,\ldots,i_m\}$，其中，i_k 表示网络数据的某一特征。项集的支持度计算过程公式为：

$$p(I)=\frac{\left|T\in D: I\subseteq T\right|}{\left|D\right|} \tag{5}$$

式中：D 表示预处理后的网络信息数据集，T 表示数据集中的一条数据记录，$|D|$ 表示事务总数。

为了处理大规模数据集，本文采用分布式计算策略。将网络信息数据分为多个子集 $D=\{D_1,D_2,\ldots,D_p\}$，每个子集在分布式计算集群的一个节点上独立计算其局部支持度，其公式为：

$$p_j(I_o^j)=\frac{\left|T\in D_j: I\subseteq T\right|}{\left|D_j\right|} \tag{6}$$

式中：$p_j(I_o^j)$ 表示在第 j 个数据子集中项集 I_o^j 的局部支持度；D_j 表示第 j 个网络信息数据块。

随后将局部支持度通过全局聚合函数合并，以得到整个数据集的频繁项集支持度 $p(I_o)$：

$$p(I_o)=\frac{\sum_{j=1}^{P}(p_j(I_o^j)\cdot|D_j|)}{|D|} \tag{7}$$

在识别出频繁项集后，为了进一步挖掘这些项集之间的深层关联，本文通过计算支持度和置信度，筛选出既频繁又可靠的关联规则。置信度的计算过程为：

$$\mathrm{conf}(X\to Y)=\frac{p(X\cup Y)}{p(X)} \tag{8}$$

最后，将从各子集中挖掘得到的关联规则合并成一个全面的网络异常信息库 R：

$$R=\left\{X\to Y\middle|p(X\to Y)\geq F'\theta_{\sup p}\wedge \mathrm{conf}(X\to Y)\geq F'\theta_{\mathrm{conf}}\right\} \tag{9}$$

式中：θ_p 表示支持度阈值，θ_{conf} 表示置信度阈值，用于筛选高质量的关联规则。

通过上述方式构建的网络异常信息库不仅包含了丰富的网络行为特征与异常事件之间的关联模式，还为后续的网络异常检测、安全分析等工作提供了强有力的数据支持。通过这种方法，能够更快速、准确地识别网络中的潜在威胁，提升整体网络安全防护能力。

3 基于改进聚类算法的网络异常信息挖掘实现

由于传统的 K-means 聚类算法在初始聚类中心的选择上表现出高度的敏感性，这种敏感性常常导致聚类结果的不稳定性和对异常数据的不充分识别。为了解决这一问题，本文引入最大最小聚类准则，通过计算网络异常信息库中数据点之间的最大最小距离，智能选取能最大化初始聚类间差异性的点作为初始聚类中心。网络异常信息库 R 中的初始中心集 $Z=\{z_1,z_2,\ldots,z_l\}$ 的选择遵循以下公式：

$$z_m=\arg\max_{d\in D}\min_{z\in Z}\|d-z\|^2 \tag{10}$$

为了提升异常识别的精确性，本文设计了一种融合关联强度和距离信息的异常关联强度度量标准。该标准通过引入调整因子 a，灵活地结合相似度和距离的倒数，构建一个综合性的异常度量模型。其公式为：

$$W(z_l,z_m)=a\cdot s(z_l,z_m)+(1-a)\cdot\frac{1}{d(z_l,z_m)} \tag{11}$$

式中：$s(z_l,z_m)$ 表示相似度，$d(z_l,z_m)$ 表示距离。

该模型不仅考虑了数据点间的空间位置关系，还深入挖掘了它们之间的潜在联系，使得异常群体的划分更加精准，异常行为模式的识别更加透彻。

鉴于网络数据量持续快速增长的现状，本文引入了 Spark 分布式计算框架，利用其强大的并行计算能力，将复杂的计算任务分解为多个可并行执行的小任务，并在集群中高效执行。这种并行化处理方式能够显著减少资源消耗，大幅度提升数据处理速度和挖掘效率。同时，定义一种基于聚类中心距离的离群点得分 d_s，用于量化数据点的异常程度。离群点得分 d_s 可以用公式表示为：

$$Z_{\mathrm{s}}=\left(\sum_{z\in Z_c}\frac{1}{W(z_l,z_m)}\right)^{-1} \tag{12}$$

式中：Z_c 表示当前迭代过程中的聚类中心集合。得分越高，表示网络信息数据点 d 越可能是离群点。

为了全面评估不同网络信息数据集之间的复杂关系并识别潜在的关联异常模式，本文定义一种综合度量标准 $\sigma(D_X,D_Y)$。该标准结合距离与相似度信息，通过计算两个数据集之间的综合距离值来评估它们之间的关联紧密程度，其公式为：

$$\sigma(D_X,D_Y)=\frac{Z_{\mathrm{s}}}{nm}\sum_{D_X\in D}\sum_{D_Y\in D}\mathrm{dist}(D_X,D_Y)\cdot\mathrm{sim}(D_X,D_Y) \tag{13}$$

式中：D_X 和 D_Y 表示待比较的网络异常信息数据集，n 和 m 分别表示 D_X、D_Y 的大小。

通过设定合理的阈值，快速准确地识别出具有显著关联性的数据集，从而实现网络异常信息的深度挖掘，具体计算公式为：

$$S=\sum_{i=1}^{n}\sum_{k=1}^{K}r_{ik}\|x_i-\mu_k\|^2+\lambda\sum_{k=1}^{K}\sum_{j\neq k}\frac{1}{\|\mu_k-\mu_j\|^2}+\gamma\sum_{k=1}^{K}\frac{1}{\sigma(D_X,D_Y)} \tag{14}$$

式中：K 表示簇的数量，x_i 表示第 i 个数据点，μ_k 表示第 k 个簇的中心，r_{ik} 表示指示变量，λ、γ 表示不同得分正则化参数。

4 实验

4.1 实验准备

为了验证本文方法的可行性，在实验准备阶段，本文搭建了一个高效且稳定的实验环境。该环境集成了先进的硬件资源、稳定的操作系统、高效的数据处理与机器学习工具、强大的数据库支持以及全面的监控与日志记录系统，同时配备了专业的网络安全工具，实验环境配置详细说明如表 1 所示。

表 1 实验环境参数配置表

参数类别	参数
硬件环境	CPU：Intel Xeon Gold 6230
	内存：128 GB DDR4 ECC
	存储：2 TB NVMe SSD（系统盘）+ 10 TB SATA HDD（数据盘）
	网络带宽：1 Gbit/s 专用网络接入
操作系统	类型：Ubuntu Server 20.04 LTS
数据处理	库 / 工具：Pandas 1.2.4, NumPy 1.19.5
机器学习	框架：TensorFlow 2.4.1, PyTorch 1.7.1
数据库	类型：PostgreSQL 13
监控与日志	工具：Grafana 7.5.3, Logstash 7.10, Kibana 7.10
网络安全工具	抓包工具：Wireshark 3.4.0

在搭建实验环境的同时，本文还收集了多样化的网络异常数据集，这些数据集具有较大的规模，能够充分测试算法在大规模数据处理方面的能力。具体收集的实验数据集详情如表 2 所示。

表 2 实验数据集规模及来源

数据集名称	数据规模	数据来源
跨站脚本攻击数据集	100 000 条	公开漏洞库
僵尸网络数据集	200 000 条	网络安全研究组织
云服务数据泄露数据集	150 000 条	云服务提供商
物联网非法通信数据集	300 000 条	物联网设备测试
网络钓鱼欺诈链接数据集	80 000 条	网络安全实验室
DDoS 攻击流量数据集	500 000 条	网络流量监测

4.2 实验结果及分析

为了验证本文方法的优越性，将本文方法与文献综述的 4 种流行的数据挖掘方法进行对比。基于支持向量机的通信网络异常流量数据挖掘方法作为对比方法 1；基于深度集成学习的社交网络异常数据挖掘算法作为对比方法 2；基于关联规则的网络信息数据挖掘方法作为对比方法 3；网络安全中计算机文本信息挖掘技术研究作为对比方法 4。分别记录这些方法在不同异常实验场景下的挖掘数据缺失情况，对比实验结果如下图 2 所示。

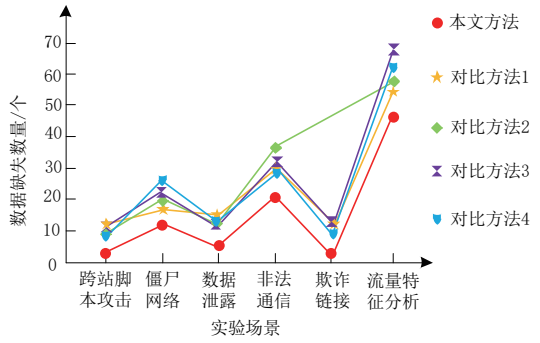

图 2 不同实验场景下 5 种挖掘方法的数据缺失数量对比

据实验数据的图表显示，本文提出的方法在各种异常场景下相较于对比方法具有明显较低的数据缺失个数，这充分验证了其在网络异常信息挖掘中的优异性能。特别是在需要高精度的场景，例如跨站脚本攻击检测和网络钓鱼欺诈链接识别，本文方法的数据缺失个数极低，显示出其在大幅度捕捉细微异常方面的卓越能力。同时，在处理恶意软件传播、僵尸网络识别、云服务数据泄露等更为复杂的场景时，本文方法也展现出了强大的处理能力和高准确性，有效减少了数据缺失，并增强了网络安全的防御能力。

在此基础上，对比了不同方法的网络异常信息挖掘耗时，结果如表 3 所示。

表 3 网络异常信息挖掘耗时

实验场景	挖掘耗时 /s				
	本文方法	对比方法 1	对比方法 2	对比方法 3	对比方法 4
跨站脚本攻击	0.63	1.66	3.47	2.14	3.14
僵尸网络	0.85	1.24	2.63	2.32	3.78
数据泄露	1.21	1.47	2.57	2.58	3.69
非法通信	0.85	1.66	2.78	2.63	3.58
欺诈链接	0.36	2.05	2.96	2.47	3.72
流量特征分析	0.75	1.78	2.74	2.98	3.14
均值	0.78	1.64	2.86	2.52	3.51

分析表 3 中的结果可知，本文方法的平均挖掘耗时仅为 0.78 s，远低于对比方法 1 的 1.64 s、对比方法 2 的 2.86 s、对比方法 3 的 2.52 s，以及对比方法 4 的 3.51 s。这一结果表明，本文方法在处理速度上实现了显著提升，进一步证明了其高效性。

5 结语

本文提出的基于改进关联聚类算法的网络异常信息挖掘方法，通过引入关联强度维度并优化聚类策略，成功在复杂网络环境中实现了异常检测效率和准确性的双重提升。实验验证不仅验证了该算法在精准识别网络异常行为上的优势，还突出了其在减少误判与漏判、提升挖掘速度方面的显著成效。在未来研究中，将积极探索引入更多的先进机器学习技术和数据挖掘方法，以应对日益复杂多变的网络异常行为。

参考文献：

[1] 劳雪松 . 基于支持向量机的通信网络异常流量数据挖掘方法 [J]. 信息与电脑 (理论版), 2023, 35(12):197-200.

[2] 戴礼灿 , 代翔 , 崔莹 , 等 . 基于深度集成学习的社交网络异常数据挖掘算法 [J]. 吉林大学学报 (工学版), 2022, 52(11): 2712-2717.

[3] 王润芳 , 丁晓敏 . 基于关联规则的网络信息数据挖掘方法 [J]. 科学技术创新 ,2021(11):80-81.

[4] 黄细标 . 网络安全中计算机文本信息挖掘技术研究 [J]. 长江信息通信 ,2023,36(9):121-123.

[5] 姜宁 . 以信息熵为基础的通信网络异常流量检测方法 [J]. 延安大学学报 (自然科学版),2024,43(2):101-104.

[6] 童莉 , 刘三民 . 网络安全中文本信息挖掘技术优化策略研究 [J]. 电脑知识与技术 ,2024,20(14):79-82.

【作者简介】

卜浏（1984—），男，江苏泰兴人，硕士，副教授，研究方向：网络工程、物联网。

（收稿日期：2024-08-14）