文献引用格式： 甘迎辉，程永新，王梓，等．多聚类融合算法在频繁非法访问检测中的应用[J]. 通信技术，2023,56(4):515-520.
doi:10.3969/j.issn.1002-0802.2023.04.017

多聚类融合算法在频繁非法访问检测中的应用 *

甘迎辉[1]，程永新[1]，王　梓[2]，彭　凯[1]
（1. 中国电子科技集团公司第三十研究所，四川 成都 610041；
2. 电子科技大学，四川 成都 610054）

摘　要： 针对频繁非法访问的检查问题，提供了一种机器学习方法来检测登录场景中的频繁非法访问活动。通过特征工程的方法分析登录日志数据，筛选提取有效特征，再使用聚类方法对登录特征数据进行检测，分类出正常用户和异常用户。为了提高无监督识别算法的精度，提出了多聚类融合的检测算法，从多个聚类算法的角度，精确识别出登录日志中的频繁非法访问用户。实验结果证明，该方法可以更准确地提取登录场景中各项指标异常的用户，并可以扩展适应其他频繁非法访问场景。

关键词： 非法访问；机器学习；多聚类融合；检测算法
中图分类号： TP309　　**文献标识码：** A　　**文章编号：** 1002-0802(2023)-04-0515-06

Application of Multi-Clustering Ensembles Algorithm in Frequent Illegal Access Detection

GAN Yinghui[1], CHENG Yongxin[1], WANG Zi[2], PENG Kai[1]
（1.No.30 Institute of CETC, Chengdu Sichuan 610041, China;
2.University of Electronic Science and Technology of China, Chengdu Sichuan 610054, China）

Abstract: For the inspection problem of frequent illegal accesses, a machine learning method is proposed to detect frequent illegal access in the login scenario. Relevant features are extracted and filtered from the login log data by feature engineering methods, after that, the clustering method is used to detect the login data and classify normal users and abnormal users. In order to improve the accuracy of unsupervised recognition algorithm, the detection algorithm of multi-clustering ensembles is proposed, which can more accurately identify frequent illegal access users from the perspective of multiple clustering algorithms. Experimental results indicate that the method can more accurately extract users with abnormal indicators in the login scenario and can be expanded to adapt to other frequent illegal access scenarios.

Keywords: illegal access; machine learning; multi-clustering ensemble; detection algorithm

0　引　言

近年来，随着互联网的巨大发展，人们对网络的依赖程度越来越高，网络信息安全无论对于个人还是企业都逐渐成为一个重要的命题。计算机网络被入侵者攻击，可能导致整个网络服务崩溃、私人信息泄漏和经济损失，甚至扰乱整个社会的秩序。

当前，攻击者针对目标系统的攻击手段和方法越来越丰富，攻击的隐蔽性也越来越高，这种变化为入侵检测和异常检测带来了很大挑战。因此，通过高效机制提升网络安全，及时准确发现网络异常

* **收稿日期：** 2022-12-12；**修回日期：** 2023-03-14　Received date:2022-12-12; Revised date:2023-03-14

攻击行为，构建强大的异常检测系统变得极其重要。

异常检测系统通过监控和分析网络相关数据来发现恶意活动[1-2]。入侵检测可以发生在主机级别或网络级别[3]。在基于主机的入侵检测中，分析进出主机的网络流量，还可以分析主机级别的其他类型的数据（例如日志信息）以检测攻击。在基于网络的入侵检测中，对进出计算机网络的流量进行监控和分析，也可以通过网络防火墙和监测系统生成的日志进行检测。本文主要研究基于网络的频繁非法访问攻击活动。

频繁非法访问是网络安全攻防场景中最常见的一种情况，是指用户在短时间内产生大量接口访问记录，或者访问行为有一定规律，疑似用代码请求接口，区别于正常用户的访问行为模式，这类用户可能存在暴力破解密码或拒绝服务式攻击的行为。

1 相关工作

目前，基于不同的场景，有不同的频繁访问检测方法，比较典型的场景包括超文本传输协议（Hyper Text Transfer Protocol，HTTP）服务的频繁非法访问和文件传输协议（File Transfer Protocol，FTP）服务的频繁非法访问。

对于 HTTP 应用来说，面对频繁非法访问攻击，可以通过使用图像验证码或者二次认证码，如手机短信或动态口令（One-time Password，OTP）验证码，来保护用户的账户[4-5]。或是通过限制用户账号的登录次数，规定系统的检测时间间隔，在时间间隔内记录用户请求接口或刷新页面的次数，如果用户的刷新或请求次数超过了一定的阈值，则将用户账号加入黑名单，或限制用户请求和刷新的速度[6]。但是此类频繁检测方法使用的场景较为单一，图像验证码和二次认证码只能在登录页面中在一定程度上防范登录场景下的暴力破解攻击[7]。而且在 HTTP 场景下，大多数频繁非法检测方法是基于用户在一段时间的登录次数或用户登录次数的排名来设定阈值，依赖的特征比较单一，检测结果不够精确。

对于 FTP 登录中的频繁非法访问，研究者提出了一种基于流量特征进行检测的方法[8]。原型系统对访问端口的流量包持续捕获，并分析 5 min 间隔内端口的连接数量、平均发包数和包的平均大小[9]，发现异常流量特征波动较小，而正常流量特征波动较大。因此对流量特征的方差设定阈值，如果一段时间内流量的波动小于阈值，则判定流量为异常流量，存在频繁非法访问的可能[10]。随着机器学习的发展，研究者也开始使用深度学习的方法来进行相关的检测。Narayan[11] 提出了一种混合分类方法，基于 FTP 登录流量特征，利用朴素贝叶斯和 C4.5 决策树进行入侵检测。建立这种方法的目的是充分挖掘登录特征，并能在一定程度上减少实时的计算特征。Haque 等人[12] 实现了一种方法，可以增强异常检测系统对即将到来的数据的预测性能，该方法使用朴素贝叶斯和随机森林对知识发现（Knowledge Discovery in Database，KDD）数据集和各种特征减少模式进行分类。他们的研究结果表明，通过使用上述分类器，整体执行时间得到了显著改善。但是 FTP 端口的频繁非法访问检测只能针对特定协议或特定端口，不能针对某一种服务场景，不能从应用层的等级筛选特定的流量，使用场景较为局限。

无论 HTTP 服务还是 FTP 服务的频繁非法访问检测，都不能适应动态场景，不支持在动态场景下捕捉数据异常。目前大多数频繁非法访问算法仅能根据目标系统的登录次数和访问流量等信息设置相应的检测标准，不能做到在不同场景下通过学习系统历史数据自动提取异常用户行为，不能灵活地设置特征标准对访问数据进行筛选。检测的实时性也存在一定问题，尽管加入了频繁非法访问检测系统中用户产生的访问行为数据，大多数算法的性能和准确度目前还不能满足实际应用要求，也无法做到通过一段时间的数据特征调整检测算法。而且，当前技术通过有监督的方式进行检测，需要大量的攻击标注数据，费时费力，且扩展性不高。

因此，如何构建一个频繁非法访问检测算法，更好地适应各种异常检测场景，根据定义的特征精确地提取相应的异常用户行为，并且省去大量数据标注的步骤，使算法面对不同场景有更好的扩展性和适应性，是目前研究所面临的难点[13]。

本文主要提出了基于融合聚类的频繁非法访问检测方法，解决现有异常检测方法存在的检测异常不准确、不能更好地利用数据特征的问题，改善检测手段单一，不能适应动态数据等不足，搭建了原型系统，并基于登录日志数据验证了检测算法的准确性。本文的贡献主要有以下两点：（1）结合日志数据，使用无监督的算法实现服务器登录等场景中相关特征的提取和频繁非法访问行为的检测。（2）在异常检测的过程中，为了使检测结果更加精确并结合监督算法的融合思想，提出了无监督算法的融合方法，更全面地计算用户相应的异常值，使检测结果更加精确。

2　频繁非法访问检测系统设计

首先，非法访问检测系统对日志数据进行预处理，包括用户访问特征的生成，访问特征的筛选，数据清洗等几个步骤。

其次，分别使用 K 均值聚类（K means clustering，K-means）、局部异常因子（Local Outlier Factor，LOF）和高斯混合模型（Gaussian Mixture Model，GMM）这 3 种算法对日志数据进行异常分析，得出每个模型下的异常用户。

最后，计算每种检测模型的轮廓系数，与用户的登录次数共同作为评价指标计算用户的异常分数，设定阈值提取异常用户，向系统发出警告。频繁非法访问检测算法的流程如图 1 所示。

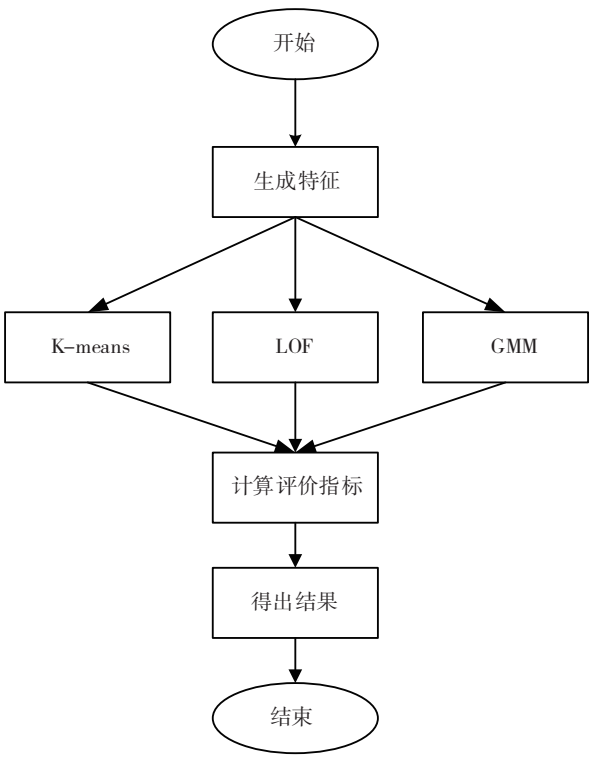

图 1　频繁非法访问检测流程

2.1　数据集预处理

本系统采用的网站登录数据集包括用户登录 ID、上线时间、下线时间、IP 地址、MAC 地址、访问流量等信息，为了分析用户是否存在频繁非法访问的行为，本系统先通过预处理方法从日志数据中提取频繁访问相关特征，数据预处理分为以下 5 个步骤：

（1）计算用户登录次数。利用滑动窗口算法提取每个用户在定义时间间隔内的登录次数，对于每条记录，统计上线时间及定义间隔内的记录条数，得出用户登录次数。

（2）编制每个用户的登录情况列表。统计每个用户的相关登录特征，包括用户最大登录次数，两次上线最大间隔，上下线间隔，5 min 内产生的最大流量，登录转换 IP 地址次数等特征。

（3）数据清洗。对计算的用户数据进行数据清洗，对某些数据的重复内容和空白内容进行删除。

（4）数据截取。初步筛选用户数据和特征，日志数据中在定义时间间隔内只登录一次的用户不存在频繁访问的可能，因此将这部分用户剔除。

（5）特征归一化。将计算出的特征使用线性函数归一化，将原始数据转换为 [0,1] 的范围，方便后续输入模型进行计算。

2.2　无监督频繁非法访问检测

一般情况下，可以把异常检测看成数据不平衡下的分类问题。在现实情况中，异常检测问题往往是没有标签的，训练数据中并未标出哪些是异常点，因此必须使用无监督学习。一般的无监督异常检测模型大致可分为统计与概率模型、线性模型和基于相似度衡量的模型。

统计与概率模型主要是对数据的分布做出假设，找出假设下所定义的“异常”，往往会使用极值分析或者假设检验。比如对最简单的一维数据假设高斯分布，然后将距离均值特定范围以外的数据当作异常点。线性模型是假设数据在低维空间上有嵌入，无法嵌入或者在低维空间投射后表现不好的数据可以认为是离群点。基于相似度衡量的模型中，异常点和正常点的分布不同，相似度较低，由此衍生了一系列算法通过相似度来识别异常点 [14]。比如最简单的 K 近邻算法就可以做异常检测，一个样本和它第 k 个近邻的距离可以被当作异常值，显然异常点的 k 近邻距离更大 [15]。同理，基于密度分析如 LOF 和 LoOP 主要是通过局部的数据密度来检测异常。在本系统中，主要采用 K-means、GMM 和 LOF 分别提取频繁非法访问的用户。

K-means[16] 聚类算法是一种以平均值作为初始聚类中心的基于划分的聚类方法，简单而且快速。算法将数据对象划分为聚类，使所获得的聚类满足同一聚类中的对象相似度较高，而不同聚类中的对象相似度较小，数据多的簇当作正常簇，数据少的簇当作异常簇。本系统在进行聚类前先对中心点进行初始化，使初始的聚类中心之间的距离为数据集中相距最远的点。该算法能够较准确地发现异常用户。

GMM[17] 是一种高斯混合聚类模型，高斯模型就是用高斯概率密度函数（正态分布曲线）精确地

量化事物，将一个事物分解为若干个基于高斯概率密度函数形成的模型。由于正常用户和异常用户的登录数据存在不同的高斯分布规律，GMM 可以通过数据的分布发现频繁登录的异常用户 [18]。

LOF[19] 算法，全称为局部异常因子算法，是一种基于距离的异常点检测算法。该算法会给数据集中的每个点计算一个离群因子 LOF，通过判断 LOF 是否接近 1 来判定是否为离群因子。若 LOF 远大于 1，则认为是离群因子；若 LOF 接近 1，则认为是正常点。在用户登录日志数据中，大多数数据为一次登录流量较小的正常数据，使用离群点检测算法可以发现距离正常点较远的异常点，从而对用户进行分类。

K-means 和 LOF 算法在原理上都是基于相似性度量对正常用户和异常用户进行区分，K-means 善于发现异常点中不同形状的簇，而 LOF 通过离群因子计算某些较为边缘的异常点，可以和 K-means 的异常提取结果相互补充。GMM 则是一种基于分布的模型，与基于密度的模型从特征点对异常数据的分析不同，可以从整体的分布规律的角度提取日志数据中的异常用户。实验结果表明，使用这 3 种算法进行日志数据的异常分析时结果可以达到最优。

2.3 无监督聚类融合算法

上述 3 种聚类算法从不同角度提取出了不同的异常用户集合，为了结合几种检测算法分类的优势，更精确地提取频繁非法访问的用户，本文提出了一种融合聚类的算法，将几种不同聚类模型的结果进行结合，计算用户的异常值。

在监督模型中，投票法是一种遵循少数服从多数原则的集成学习模型，通过多个模型的集成降低方差，从而提高模型的鲁棒性和泛化能力 [20]。投票法对各类检测结果的概率进行求和，最终选取概率之和最大的类标签。在无监督学习中，无法通过标签的预测结果决定模型的权重，因此将聚类算法的轮廓系数 P 值和用户间隔时间内的登录次数作为评判标准。

在无监督融合算法中，设每个聚类算法的轮廓系数为 P，用户在规定间隔内的登录次数为 x，则每个用户频繁非法访问的异常值得分为：

$$Value=\sum_{i=1}^{n}P_i+\frac{x-m}{\delta}$$

式中：m 为潜在异常用户登录的平均次数；δ 为潜在异常用户的方差。

P 值代表模型的聚类效果，P 值越接近 1，数据的分类程度越高，则模型的权重占比越高。规定间隔内的登录次数为频繁非法访问的主要检测特征，作为检验数据异常度的标准。

将所有用户按照异常得分进行排名，提取排名较前的用户为异常用户，从各个特征生成正常用户和异常用户的统计直方图，得到异常用户的访问特征。对数据集中所有异常用户的个人数据使用融合模型进行分析，提取用户个人数据中排名较前且时间间隔较大的登录数据，统计用户频繁非法访问次数，如果用户多次发生频繁非法访问的行为，则将用户从白名单中移除。

3 实 验

3.1 实验细节

实验中的模型采用 Python 的机器学习框架 sklearn 库进行训练。K-means 算法的初值 k 设置为 3，LOF 算法的 contamination 设置为 0.1，n_neighbors 设置为 20，GMM 算法的 n_components 设置为 2。

3.2 实验结果分析

实验过程中，监测时间的间隔为 5 min，运行算法，日志文件中共有 3 511 个用户，使用 3 种算法得出异常用户的集合后，实验发现当阈值设定为 1 时，最终分类的 DI 系数最大，于是筛选出异常分数大于 1 的用户，判断为异常用户，筛选后的部分用户如表 1 所示。

表 1 异常用户结果

MASKID	5 min 访问次数	下次上线最大间隔 /min	IP	平均流量 /MB	分数
D235508	19	0.001 2	10.20.24.140	397.74	6.87
T87602	11	0.016 6	10.20.25.169	85.72	4.03
T190799	6	0.001 4	121.48.161.20	947.71	1.65
T322906	4	0.483 3	10.20.26.54	445.13	1.57
D108779	4	0.083 3	10.20.26.55	53.69	1.56
S336135	4	0.316 7	10.20.13.128	59.23	1.53
S278286	4	0.783 3	113.54.158.42	2.88	1.44
T108347	5	0.166 7	121.48.161.174	497.07	1.28

实验使用融合聚类算法和单一异常检测算法进行对比实验，从实验结果可以看出，融合聚类算法的性能明显优于单一异常检测算法，可以结合单一异常检测算法的优势，更精确地发现日志数据中的异常用户，具体实验结果如表 2 所示。

表 2　算法对比

算法名称	轮廓系数	异常用户数
K-MEANS	0.67	25
LOF	0.63	22
GMM	0.58	24
本文方法	0.74	19

提取出最终的异常用户后，可以发现异常用户在定义间隔内的登录次数较高，为了进一步得出异常用户和正常用户其他特征之间的区别，使用直方图对异常用户和正常用户的特征进行比较。如图 2 所示，从正常用户和异常用户的直方图可知，正常用户和异常用户的平均上下线间隔区分度较小，而异常用户两次登录间隔的时间较小，且发送的访问流量在短时间内较大，算法根据定义间隔内的登录次数、两次登录的间隔、平均流量的大小可判定用户是否为异常。

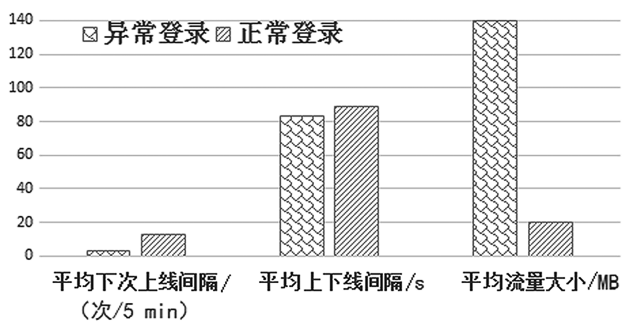

图 2　用户特征直方图

4　结　语

随着互联网规模的不断扩大，使用异常检测系统防御频繁非法访问攻击变得越来越重要。与现有频繁非法访问算法相比，本文有效利用用户登录日志特征，通过机器学习算法实现对用户数据的学习，可以动态地获取数据中的异常值，并依据监督学习投票算法的思想，构造出无监督学习融合算法，利用了各个无监督学习算法的特点和效果，提高了模型分类的准确性，为无监督异常检测提供了新的思路。结合特征工程对现有的数据特征进行筛选和处理，可以更好地适应各种场景，解决了相关算法只能针对部分协议的问题。

参考文献：

[1] JONES A K,SIELKEN R S.Computer system intrusion detection: A survey[J].Computer Science Technical Report,2000:1-25.

[2] SCARFONE K A,MELL P M.1Sp 800-94, guide to intrusion detection and prevention systems (IDPS)[EB/OL].(2007-02-20)[2021-12-13].https://www.nist.gov/publications/guide-intrusion-detection-and-prevention-systems-idps.

[3] DAS N,SARKAR T.Survey on host and network based intrusion detection system[J]. International Journal of Advanced Networking and Applications, 2014,6(2):2266-2269.

[4] VARNE R B,MANE R V.CAPTCHA:A robust approach to resist online password guessing attacks[C]//2014 International Conference on Advances in Communication and Computing Technologies (ICACACT 2014),2015:1-6.

[5] ROUTH C,DECRESCENZO B,ROY S.Attacks and vulnerability analysis of e-mail as a password reset point[C]//2018 Fourth International Conference on Mobile and Secure Services (MobiSecServ),2018:1-5.

[6] POWELL B M,KUMAR A,THAPAR J,et al.A multibiometrics-based CAPTCHA for improved online security[C]//2016 IEEE 8th International Conference on Biometrics Theory,Applications and Systems (BTAS),2016:1-8.

[7] KIRUSHNAAMONI R.Defenses to curb online password guessing attacks[C]//2013 International Conference on Information Communication and Embedded Systems (ICICES),2013:317-322.

[8] 魏琴芳，杨子明，胡向东，等．基于流量特征的登录账号密码暴力破解攻击检测方法 [J]. 西南大学学报（自然科学版），2017,39(7):149-154.

[9] SATOH A,NAKAMURA Y,IKENAGA T.A flow-based detection method for stealthy dictionary attacks against Secure Shell[J].Journal of Information Security and Applications,2015,21:31-41.

[10] ALSALEH M,MANNAN M,VAN OORSCHOT P C.Revisiting defenses against large-scale online password guessing attacks[J].IEEE Transactions on Dependable and Secure Computing,2012,9(1):128-141.

[11] ARJUNWADKAR N M,PARVAT T J.An intrusion

detection system with machine learning model combining hybrid classifiers[J].Journal of Multidisciplinary Engineering Science and Technology,2015,2(4):647–651.

[12] HAQUE M E,ALKHAROBI T M.Adaptive hybrid model for network intrusion detection and comparison among machine learning algorithms[J].International Journal of Machine Learning and Computing,2015,5(1):17–23.

[13] 杨加 , 李笑难 , 张扬 , 等 . 基于大数据分析的校园电子邮件异常行为检测技术研究 [J]. 通信学报 ,2018,39(增刊 1):116–123.

[14] 王彬彬 . 基于 K–means 聚类的软件定义网络异常流量分类研究 [J]. 齐齐哈尔大学学报 (自然科学版),2022,38(2):50–55.

[15] 史小艳 , 陈松灿 . 基于单簇聚类的非对齐多视图异常检测算法 [J]. 中国科学 : 信息科学 ,2021,51(12):2037–2052.

[16] 向继 , 高能 , 荆继武 . 聚类算法在网络入侵检测中的应用 [J]. 计算机工程 ,2003,29(16):48–49.

[17] ALBERTI K G M M,ZIMMET P Z,CONSULTATION W.Definition,diagnosis and classification of diabetes mellitus and its complications.Part 1:Diagnosis and classification of diabetes mellitus.Provisional report of a WHO consultation[J].Diabetic Medicine,1998,15(7):539–553.

[18] 张少锋 , 王建元 . 基于线性判别分析和密度峰值聚类的异常用电模式检测 [J/OL]. 电力系统自动化 :1–12[2021–12–26]. http://kns.cnki.net/kcms/detail/32.1180.TP.20211207.1014.006.html.

[19] BREUNIG M M,KRIEGEL H P,NG R T,et al.Lof[J].ACM SIGMOD Record,2000,29(2):93–104.

[20] 董师师 , 黄哲学 . 随机森林理论浅析 [J]. 集成技术 ,2013,2(1):1–7.

作者简介：

甘迎辉（1971—），男，硕士，高级工程师，主要研究方向为信息安全；

程永新（1978—），男，硕士，正高级工程师，主要研究方向为信息安全；

王　梓（1999—），男，学士，工程师，主要研究方向为网络安全技术；

彭　凯（1984—），男，硕士，工程师，主要研究方向为信息安全。

声　明

近期，我编辑部发现有不法分子冒充本刊编辑部进行非法采稿、虚假宣传，以刊发稿件为由收取所谓“审稿费”及“版面费”，还有人声称《通信技术》已暂停收稿，对我刊声誉和正常工作造成了严重的不良影响，扰乱了正常学术秩序，极大地损害了作者、读者的利益。

为此，编辑部郑重声明如下：

一、本刊从未授权任何单位和机构代理《通信技术》的征稿业务，而且本刊一直正常对外收稿，切勿上当受骗。

二、冒充本刊进行违法活动者，请立即停止一切侵权行为和非法活动。

三、为提升服务质量、保障作者权益，《通信技术》已停用原投稿邮箱，启用新投稿邮箱：txjstgyx@163.com。

四、请广大作者提高警惕，保护个人的合法权益，投稿时请务必核实投稿邮箱及地址。本刊官方投稿渠道如下：

唯一官方投稿网址：www.txjszz.com

唯一官方投稿邮箱：txjstgyx@163.com

联系人：李老师

联系电话：028–85169918

如有其他形式，均为假冒，本社将保留诉讼法律的权利。

特此声明。

《通信技术》编辑部

2023 年 4 月 15 日