基于 K-means 聚类算法的异常流量识别能力分析

叶帅辰[1] 卜 哲[1]
YE Shuaichen BU Zhe

摘 要 通过特征聚类进行异常网络流量识别是实现网络风险防御的基础，其中基于无监督学习的 K-means 聚类算法由于其具有高效、易实现等特点而在业界得到了广泛应用。目前针对 K-means 算法的研究大多聚焦于提升其对流量整体特征的聚类准确性，而较少涉及单一特征对于异常网络流量的快速鉴别能力研究。针对这一问题，依托 K-means 算法逐一分析 KDD Cup99 数据集中各项特征对于拒绝服务攻击、非法访问、扫描渗透等威胁的识别能力，有效实现不完整网络安全监测数据场景下的风险快速判断，为网络安全监测设备能力优化提供参考。

关键词 网络安全；流量特征；风险识别；K-means 聚类算法

doi：10.3969/j.issn.1672-9528.2024.06.027

0 引言

目前，随着互联网规模不断扩大，其面临的安全问题也愈发严峻。在日益频繁的网络攻防博弈中，攻击者采用较为隐蔽的攻击手法，将大量带有恶意攻击载荷的异常流量混杂在庞大的信息通信网络中，使互联网用户及重要数据资源时刻面临着潜在未知风险。当前，针对网络风险较为有效的监测发现手段是在关键网络节点处部署全流量监测分析设备（network traffic analysis，NTA），用以区分正常和异常流量。

目前 NTA 中常用的流量分类方法根据其所使用技术不同分为 4 种：基于端口识别的分类方法、深度报文数据包检测分类方法 [1-2]、基于流量传输行为的分类方法 [3] 和基于机器学习的流量分类方法 [4-5]。上述 4 种方法在流量分类准确性、及时性等方面具有各自的优缺点，适用场景也不尽相同。不过，随着网络流量特征的复杂性不断增加，需要大规模标识数据训练来提高分类准确性的机器学习方法逐渐占据了业内主导地位，其中 K-means 聚类算法由于具有自学习特性，在主流 NTA 系统中得到了广泛应用。

近年来，已有学者针对 K-means 聚类算法的实现及优化进行了相关研究。文献 [6] 详细介绍了基础 K-means 算法的实现流程，并引入近邻密度和最远欧氏距离概念来选择不同聚类中心对传统算法进行优化，使其无论在平均值还是最优值上都表现出了较高的聚类准确率。文献 [7] 提出对采集的各网络安全数据利用 hash 函数进行预处理后再使用 K-means 聚类，可使各异常攻击流量的误检率最多降低 10%。此外，也有相关研究将遗传算法、模糊理论等人工智能算法引入聚类过程，用于对所采集网络信息进行初步分类，以提升 K-means 算法的聚类效率和准确性 [8-9]。

目前已有研究大多将所监测采集的多维网络安全数据进行整体聚类，各流量特征对于最终聚类结果的影响权重存在相互干涉，难以确定单一特征对于异常网络流量的鉴别能力，进而导致在部分网络数据获取不完整的场景下，无法快速识别网络风险 [10]。本研究针对以上问题，使用 K-means 算法对 KDD Cup99 数据集中 41 项流量子特征进行逐一聚类，分析各单项特征对于异常流量的识别能力，为网络风险快速判断提供思路，相关研究结论可为网络安全流量监测设备能力优化过程中的参数权重调节提供参考。

1 K-means 聚类算法的实现原理

聚类算法是一种常见的对象分类方式，在图像处理、数据挖掘、流量识别等领域发挥了重要作用，其主要目的是将数据集中的多维数据按照某种相似性或距离指标分成不同的类别，不同的类别也称为簇，同一簇中的数据相似程度越高则说明聚类效果越好。其中相似程度是通过簇内各数据间的距离来衡量的，距离计算方法包括欧几里得聚类（欧式距离）、曼哈顿距离（曼氏距离）和明可夫斯基距离（明氏距离），其中欧式距离最为常见。假设 $x=(x_1, x_2, \cdots, x_p)$ 和 $y=(y_1, y_2, \cdots, y_p)$ 是两个 p 维数据，它们之间的欧氏距离可计算为：

$$\delta=\sqrt{\sum_{i=1}^{p}(x_i-y_i)^2} \quad (1)$$

1. 中国信息通信研究院安全研究所 北京 100191

那么聚类的目的就是使同一簇内的所有数据和簇中心（质心）的距离之和最小。

K-means 聚类作为一种典型的无监督聚类算法，具有收敛快速、易实现等优势，其主要特点是需要选定 k 个质心（质心可以不包含于原始数据集中），并按下述步骤逐步实现数据集划分。

步骤 1：遍历数据集，依据各数据与质心相似性，将数据分配至不同簇。

步骤 2：根据各簇新增、删除数据情况重新计算质心。

步骤 3：计算各簇内数据与新质心的距离和。

步骤 4：重复上述步骤，直至步骤 2 中质心不发生变化为止。

使用 K-means 聚类后数据集划分效果如图 1 所示。

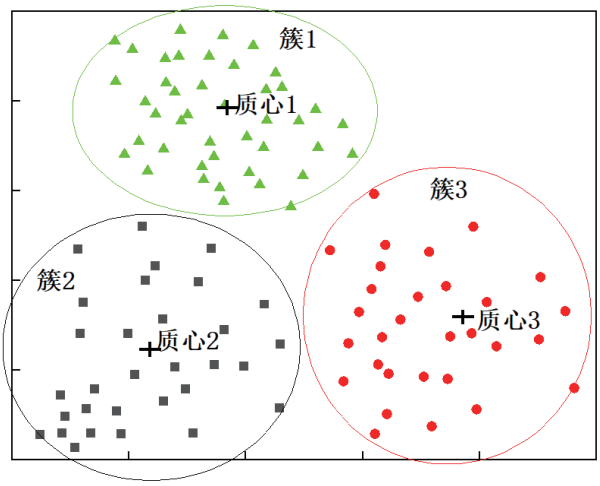

图 1 K-means 聚类效果图

2 算例分析

为了深入比对在 K-means 算法聚类下，不同网络流量特征对于各类网络攻击的鉴别能力，本文选取 KDD Cup99 数据集作为测试算例。该数据集是 1999 年由美国国防部高级规划署在实验环境中监测收集，全量数据共计 500 余万条，是目前公认涵盖流量特征较为准确、类型较全的数据集 [11]，常被用在网络入侵检测算法开发及优化、威胁知识挖掘、流量模式预测等研究领域 [12]。

2.1 KDD 99 数据集预处理

由于 KDD 99 数据集是从实网流量中获取，部分数据的格式、类型、度量等存在差异性，导致无法直接进行聚类，则需对其中每一条数据进行数值化、标准化等预处理步骤。数据集中每一单条数据包括 42 列，除第 42 列为网络攻击标识外，其余 1 ～ 41 列每列表示一项流量子特征，这 41 个子特征可归并为 4 类。

（1）TCP 连接特征（1 ～ 9 列）：包括连接持续时间（duration）、协议类型（protocol_type）、网络服务类型（service）、连接状态（flag）、源至目的发送字节数（src_bytes）等。

（2）TCP 内容特征（10 ～ 22 列）：包括敏感内容访问次数（hot）、登录失败次数（num_failed_logins）、登录状态（logged_in）、受损次数（num_compromised）、是否获取 root shell（root_shell）、是否获取 su root（su_attempted）、根用户访问次数（num_root）、文件操作次数（num_file_creations）等。

（3）基于时间的流量统计特征（23 ～ 31 列）：包括过去两秒内的同目标连接数（count）、同服务连接数（srv_count）、同目标连接中 SYN 错误率（serror_rate）、同服务连接中 SYN 错误率（srv_serror_rate）、同目标连接中 REJ 错误率（rerror_rate）等。

（4）基于连接的流量统计特征（23 ～ 31 列）：包括前 100 个连接中的同目标连接数（dst_host_count）、同目标同服务连接数（dst_host_srv_count）、同目标同服务连接率（dst_host_same_srv_rate）、同目标不同服务连接率（dst_host_diff_srv_rate）、同目标同端口连接率（dst_host_same_src_port_rate）等。

在这 41 个流量特征中，协议类型、服务类型、连接状态 3 个子特征为枚举型变量，无法按照前文方法对其进行欧式距离计算，因此需要在聚类前对各特征进行数值化替换。本文选取的数值化方法为计算各子特征属性在数据集中出现的频次：

$$\delta_i = \frac{P_i}{\sum_{i=1}^{N} P_i} \tag{2}$$

式中：P_i 表示该子特征第 i 个属性在数据集中的枚举次数，N 为该子特征的总属性个数。例如协议类型特征的 TCP、UDP、ICMP 三个属性，在具有 100 条流量数据的子集中三属性出现次数分别为 20、30、50，那么它们的数值化替换分别为 0.2、0.3、0.5，这种数值化方式能够有效提升同一特征不同枚举值间的距离均衡性。

此外，在对数据集聚类的过程中，不同子特征由于其度量单位不同，导致对整体距离计算结果的贡献权重有所差异，因此在枚举型特征数值化后，需要进一步对各子特征进行标准化，消除各子特征对于最终距离计算结果的权重差异，具体标准化计算方法为：

$$\bar{x}_i = \frac{x_i - \frac{1}{n}\sum_{i=1}^{n} x_i}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right)^2}} \tag{3}$$

式中：x_i 表示某一特征第 i 个属性的数值，n 为所选取测试子集中总数据条数。

2.2 异常流量识别准确性分析

在完成数据集的数值化、标准化后，即可开展基于 K-means 聚类的流量识别准确性分析。通过对 KDD 99 数据集的第 42 特征列分析，其共囊括了 39 种攻击标识，具体分为 4 个大攻击类别：（1）拒绝服务攻击（DOS），典型标识包括 back、land、smurf 等；（2）远程主机非法访问（R2L），典型标识包括 imap、spy、ftp_write 等；（3）用户越权访问（U2R），典型的标识包括 rootkit、buffer_overflow 等；（4）扫描渗透攻击（Probing），典型标识包括 nmap、ipsweep 等。本文针对这 4 大攻击类别，从 KDD 99 原始的 500 万条数据集中筛选出 4 个测试数据集作为特征聚类研究对象，除记录总量较少的 U2R 攻击外，针对其余 3 种攻击的总数据量均选取约 15 000 条网络连接，同时为了能够快速有效地完成异常流量聚类，所选取测试集中正常实例数量远大于异常实例数。各测试数据集的记录分布如表 1 所示。

表 1 针对不同攻击类别的测试数据集

	正常连接数	异常连接数	对应攻击类别
数据集 1	14 434	1937	DOS
数据集 2	11 935	1019	R2L
数据集 3	486	26	U2R
数据集 4	12 025	1309	Probing

以数据集 1 中的 9 个 TCP 连接子特征为例，在 Python 3.8.8 环境下逐个对其进行 K-means 聚类，结果如表 2 所示。

表 2 DOS 攻击测试数据集 TCP 连接子特征聚类结果

	duration	protocol_type	service	flag	src_bytes	dst_bytes	land	wrong_fragment	urgent
正常	16 361	15 897	15 360	16 287	14 441	15 714	16 371	16 371	16 370
异常	10	474	1011	84	1930	657	0	0	1

进一步可通过兰德系数计算聚类结果的准确率，从而衡量不同流量特征对于异常流量的识别能力。

$$RI=\frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

式中：TP（true positive）表示被正确分类为正常流量的样本数；TN（true negative）表示被正确分类为攻击流量的样本数；FP（false positive）表示实际为攻击流量，但被分类为正常流量的样本数；FN（false negative）表示实际为正常流量，但被分类为攻击流量的样本数。

从子特征角度对所选 4 个测试数据集各列逐一聚类后，得到对于不同攻击类型识别准确性，结果如图 2 ～图 5 所示，图中 4 个特征区域内的数据点从左至右依次表示 KDD 99 数据集各流量子特征。从图中可以看出基于连接的流量特征下“同目标连接数”子特征对各类型的攻击识别准确率最低，均在 70% 以下，这说明单从流量的访问目标无法识别攻击类型，甚至无法鉴别正常和异常流量。与之相反，基于时间的流量特征中“同服务不同主机率”子特征对各类型攻击的识别准确率均较高，依次为 98.68%、96.84%、99.59% 和 97.51%，因此可以推断，若在流量监测中发现单位时间内访问多个目标相同服务的行为，则此类流量极有可能存在攻击意图。

另外，值得注意的是，TCP 连接基本特征中的“源至目的发送字节数”子特征对于 DOS 攻击的识别准确率为 100%，这是因为基于海量无效请求的 DOS 攻击，其报文特征具有同质化的特点，且字节长度、内容等与正常访问流量存在明显差异，虽然考虑到所筛选测试集存在偶然性因素，但仍可采用此子特征对 DOS 攻击进行初步快速鉴别。与之类似，对于 R2L、U2R、Probing 三类攻击鉴别准确性最高的三个单一流量特征分别为“同目标同服务连接数”“同服务不同访问目标比率”“同访问目标不同源比率”，可依此进行攻击类型的快速判断。

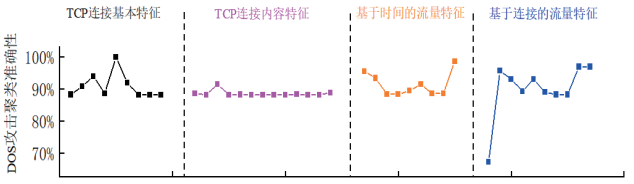

图 2 各子特征对于 DOS 攻击识别准确率

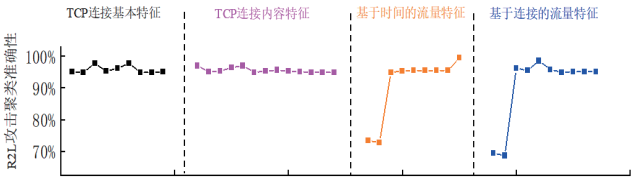

图 3 各子特征对于 R2L 攻击识别准确率

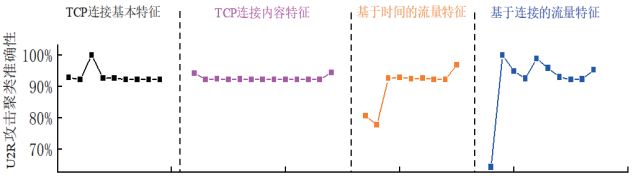

图 4 各子特征对于 U2R 攻击识别准确率

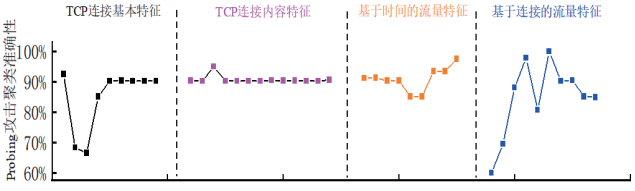

图 5 各子特征对于 Probing 攻击识别准确率

进一步地，从大类特征角度对于不同类型攻击的识别情况如图 6 所示，可看出各特征对 U2R 类攻击整体上具有较好的鉴别能力，而对 Probing 类威胁的识别准确性较低。这是因为越权访问类的攻击流量具有低频次、针对性、构造复杂等特点，无论从内容上还是形式上都迥异于正常业务流量，而扫描类威胁因为并不带有真实攻击载荷，其内容上与正常流量差异并不显著，甚至其数据包常被误检测为 ping 包、握手包等，导致对其识别效果一般。

另外，除 Probing 攻击外，TCP 连接基本特征对于其余 3 类攻击都具有 90% 以上的识别准确率，这说明仅从流量的五元组、持续时间、字节数等基本信息即可对正常和异常流量进行初步筛选，若需进一步细化鉴别攻击类型，才应考虑流内容特征、流统计特征等要素。

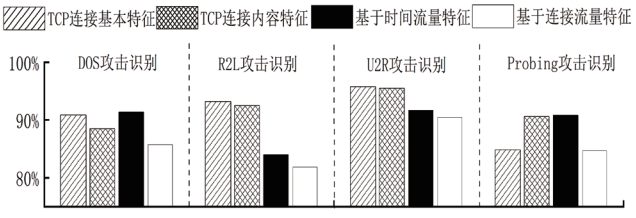

图 6 各大类特征对于不同类型攻击的识别准确率

3 结论

本文聚焦单一流量特征对各类网络风险的识别能力，采用 K-means 算法对 KDD 99 数据集中 41 项流量子特征进行逐一聚类，分析其对于 DOS 攻击、非法访问等威胁的识别准确性，针对分析结果可初步归纳如下结论。

（1）从单一特征对流量识别能力来看，对 DOS 攻击、主机远控攻击、非法访问攻击、扫描渗透攻击判断效果最佳的流量子特征分别为“发送字节数（src_bytes）”“同目标同服务连接数（dst_host_srv_count）”“同服务不同访问目标比率（srv_diff_host_rate）”和“同访问目标不同源比率（dst_host_srv_diff_host_rate）”，识别准确性分别为 99.15%、100%、99.59% 和 99.04%。

（2）从大类特征对流量识别能力来看，各类特征对于非法访问攻击的识别能力较强，甚至高于大多数单一子特征对于该类型攻击的识别结果，而对于扫描渗透类攻击的识别能力却不如多数单一子特征。另外，仅从威胁快速鉴别角度来看，在无法获取完整网络流量监测数据的场景下，可优先考虑使用连接基本特征来筛选异常流量。

综合相关分析结果，在对网络安全监测设备进行开发及优化时，针对不同类型威胁，可考虑增加相应优势特征的判定权重，从而提升异常流量识别的准确性，缩减防御响应时间。

参考文献：

[1] 刘畅 . 面向特定网络流的深度报文检测技术研究 [D]. 哈尔滨 : 哈尔滨工程大学 , 2018.

[2] 郭婷 . 深度数据包和深度数据流检测技术研究 [D]. 长春 : 长春理工大学 , 2014.

[3] 张晓航 . 基于机器学习的加密流量行为分析技术研究 [D]. 北京 : 中国电子科技集团公司电子科学研究院 , 2022.

[4] 顾玥 , 李丹 , 高凯辉 . 基于机器学习和深度学习的网络流量分类研究 [J]. 电信科学 , 2021, 37(3): 105-113.

[5] 孔晓晨 . 基于半监督学习的网络流量分类技术研究 [D]. 北京 : 北京邮电大学 , 2018.

[6] 唐欣 . 三支 K-means 聚类算法及其应用研究 [D]. 银川 : 北方民族大学 , 2024.

[7] 刘福刚 . K-means 聚类算法在网络安全检测中的应用研究 [J]. 绥化学院学报 , 2023, 43(11):157-160.

[8] 马通 . 基于遗传算法的并行化 K-means 聚类算法研究 [D]. 杭州 : 浙江理工大学 , 2018.

[9] 刘仲驰 . 基于改进模糊聚类的网络信息安全风险识别研究 [J]. 信息记录材料 , 2023, 24(3):189-191.

[10] NITESH-SINGH B, MANJU K, VICENTE G D. A review on intrusion detection systems and techniques[J]. International journal of uncertainty fuzziness and knowledge-based systems, 2020, 28(S1):65-91.

[11] WELLER-FANY D J, BORGHETTI B J, SODEMANN A A.A survey of distance and similarity measures used within network intrusion anomaly detection[J]. IEEE communications survey&tutorials, 2014, 17(1):70-91.

[12] PIERPAOLO D, ABDUSSALAM E, ANDREA B. Overview on intrusion detection systems design exploiting machine learning for networking cybersecurity[J]. Applied sciences-basel, 2023,13:7507.

【作者简介】

叶帅辰（1994—），男，辽宁沈阳人，博士，工程师，研究方向：网络安全。

卜哲（1980—），男，安徽淮北人，硕士，高级工程师，研究方向：网络安全、信息通信。

（收稿日期：2024-05-08）