doi: 10.7690/bgzdh.2025.05.011

基于 *K*-means 算法的通信系统安全防御方法

闫卫刚
（陕西警官职业学院信息技术系，西安 710021）

摘要：为提升通信系统入侵检测性能，在 *K*-means 算法基础上进行算法优化。针对网络数据特征聚类数量无法提前估计问题，提出 *K* 值有效性指标来确定聚类数量和评测聚类质量，同时考虑各类簇特征对聚类的影响，利用特征加权距离考虑类内紧密型和类间的分离性，依此作为聚类中心点。实验结果表明：改进 *K*-means 入侵检测算法具有更优的检测率和误报率，能有效提升系统安全防御质量。

关键词：*K*-means 算法；通信系统；网络攻击；检测率
中图分类号：TN914　　**文献标志码：**A

Communication System Security Defense Method Based on *K*-means Algorithm

Yan Weigang
(*Department of Information Technology*, *Shaanxi Police Officer Vocational College*, *Xi'an* 710021, *china*)

Abstract: In order to improve the intrusion detection performance of communication system, the *K*-means algorithm is optimized. Aiming at the problem that the number of clusters of network data features can not be estimated in advance, the validity index of *K* value is proposed to determine the number of clusters and evaluate the quality of clustering. At the same time, the influence of various cluster features on clustering is considered, and the feature weighted distance is used to consider the closeness within the cluster and the separation between the clusters, which is used as the clustering center. The experimental results show that the improved *K*-means intrusion detection algorithm has better detection rate and false alarm rate, and can effectively improve the quality of system security defense.

Keywords: *K*-means algorithm; communication system; network attack; detection rate

0　引言

互联网、云计算技术的普及和发展，推动了通信系统的建设和开发。当前，通信系统在各大电子商务、互联网、金融企业中承担着数据存储、传输的重要功能。由于各项信息具有较高的价值和机密性的特点，一些不法分子为读取或破坏机密信息，不断对通信系统进行木马或病毒攻击[1-2]；因此，通信系统的安全防御一直是各大企业的关注重点，各大企业都积极的开展通信安全防御技术的研究[3]。市场上的各类杀毒软件，不同通信系统采用的非对称加密技术[4]、免疫网络安全防御技术等[5-6]，一定程度提高了系统的安全防御水平；但是，大部分通信系统采用的是被动模式，只有在遭受攻击时才会启动防御工作，因而系统防御行为落后于攻击发起时间[7]。*K*-means 算法作为人工智能的机器学习算法，能够根据网络病毒或木马特征，从海量数据中发现潜在木马病毒，实现对病毒、木马的实时查杀[8]，但 *K*-means 算法存在初始簇，评价标准单一，导致数据分析不准确，检测率低，误检率高的问题[9]。为提高 *K*-means 算法的准确度，在 *K*-means 算法基础上进行算法优化，提高 *K*-means 算法在通信系统安全防御工作中的有效性。

1　通信系统安全防御现状分析

通信系统朝着智能化的发展，光纤设备、智能手机、传感器等不同类型的设备接入到通信系统中，通信系统结构越来越复杂，各类设备开发架构和实现技术上的差异，使得通信系统存在的漏洞也越大，面临的木马和病毒攻击也很大。为提高通信系统的安全防御水平，各类通信系统安全防御技术也得到推广应用。

1) 非对称加密技术。

相较于传统的数据保密技术，非对称加密技术采用 2 个秘钥进行加密和解密操作，即首先采用哈希算法生成私钥，通过私钥进行数据加密；然后采

收稿日期：2024-08-23；**修回日期：**2024-09-21
基金项目：陕西省职业技术教育学会厅局级研究项目(2022SZX237)
第一作者：闫卫刚(1980—)，男，陕西人，硕士。

用 Base58 算法生成公钥对私钥加密的数据信息进行解密操作[10]。目前，非对称加密算法已经成为互联网主动防御的关键技术，这是由于采用哈希算法生成的私钥数量众多，且每个私钥都是遗传固定的字符串，都可以很好的保护数据，防止黑客攻击。

2) 免疫网络。

免疫网络作为一类自我学习的技术，集成了由软硬件资源、网络安全协议、安全策略，可通过路由器等形成通信系统防御机制来建立深层的、多层次防御规则。免疫网络利用授权认证的方式来接入网络，提升通信系统病毒接入的可信计算能力，防止恶意代码的攻击。采用免疫网络一方面具备严苛的通信系统设备终端接入管控能力，构建一个终端设备双向控制功能，能抵御内外部的攻击，避免了终端设备不兼容发生的攻击，在提高病毒防御能力的同时，提升通信系统抵抗能力[11]。

3) 深度包过滤。

深度包过滤部署在通信传输网关接口，为用户提供一种开放的数据包分析工具，深度包过滤技术针对数据包进行分析、挖掘、识别威胁攻击，并结合 TCP 协议、IP 协议等各类互联网数据传输协议来实现不同层次的数据包分析，利用穿透式准确判定网络威胁是否存在。此外深度包过滤技术利用嵌入式软件提升数据的处理能力，能够快速分析数据包的分发地址，完成通信系统信息过滤来确保深度包过滤的准确度。

上述分析可以看出：传统的通信系统安全防御技术均能在一定程度有效抵御来自外界对系统的攻击，但相关技术只有在遭受攻击时才会启动防御工作，而当前的通信系统安全防御更多的是主动的进行入侵识别和安全防御。*K*−means 技术作为一类机器学习算法，能够对各类数据信息进行主动学习和聚类；因而，能有效满足当前通信系统安全防御的需求。

2 优化的 *K*−means 聚类算法

2.1 *K*−means 算法概述

K−means 聚类算法通过不断的迭代计算进行数据集类别划分，算法简单易操作，且扩展性强[12]。在计算过程中，*K*−means 算法从样本集 S 中随机选择 K 个样本作为初始聚类中心，由制定的规则算法来计算不同聚类中心与对象的间距离，通过迭代计算直到中心距离保持不变后，获得 K 个聚类结果。算法实现流程为：

1) 设定包括 n 个数据对象和 K 个聚类中心的输入样本集；

2) 根据聚类规则，计算样本与中心间距离，选择最小计算距离作为聚类中心重新划分对象。即 2 个 p 维数据点 $x_i=(x_{i1}, x_{i2}, \cdots, x_{ip})$、$x_j=(x_{j1}, x_{j2}, \cdots, x_{jp})$ 间的欧氏距离为：

$$d(x_i, x_j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\cdots+(x_{ip}-x_{jp})^2}\ 。\tag{1}$$

确定样本集的平均距离为：

$$\text{Meandist}(S)=\frac{2}{n(n-1)}\times\sum_{i\neq j,\,i,\,j=1}^{n}d(x_i, x_j)\ 。\tag{2}$$

3) 计算各样本的均值，返回步骤 2)，直到目标函数值满足设定阈值或恒定不变。目标函数误差平方准则函数为：

$$\sigma_i=\sqrt{\sum_{i=1}^{n_i}(x_i-c_i)^2\Big/\left(|C_i|-1\right)}\ 。\tag{3}$$

式中 c_i 为作同类样本数据集的质心点，定义 c_i 计算公式为：

$$c_i=\frac{1}{|C_i|}\sum_{x_i\in T_i}x_j\ 。\tag{4}$$

式中$|C_i|$是类 C_i 数据对象数量。

4) 结束，获得 K 个聚类。

2.2 *K*−means 算法优化

K−means 算法具有算法简单、效率高的优势，但在实际计算过程中，由于聚类数 K 和中心点的选取都是随机给定，缺乏选择标准，往往使算法结果误差较大[13]。提出一种解决聚类数 K 的选择方法。

在选择 K 值时，首先根据实际情况对聚类范围进行有效限定，即假设聚类数 K 范围为(m, n)，则进行 $n\sim m$ 次 *K*−means 传统算法，由多次聚类结果中确定最优聚类数作为最佳聚类数，各聚类节点欧氏距离为：

$$V=\frac{d_{\text{内聚}}}{d_{\text{外聚}}},\ \left\langle d_{\text{内聚}}=\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij}-c_i)^2,\right.$$
$$\left. d_{\text{外聚}}=\frac{1}{k(k-1)}\sum_{i=1}^{K}\sum_{j=1}^{K}(c_i-c_j)^2\right\rangle 。\tag{5}$$

从式(5)中可以看出：外距为各中心距离和的 $1/k(k-1)$。内聚和外聚比值最小时则表示获得聚类数为最优解，此时聚类内高度内聚，聚类间低度耦合。

在选取初始点时，要求初始中心周边点必须密集，而不同中心点相互距离最大。此时定义各中心点互相距离为：

$$d=\frac{1}{k(k-1)}\sum_{i=1}^{K}\sum_{j=1}^{K}(c_i-c_j)^2\text{。}\tag{6}$$

式中 d 为中心点距离均值，用来表现聚类中心相互距离情况。对于聚类中心周边各点的密度，采用式(7)计算：

$$p_i=z_i/\sum_1^n z_k\text{。}\tag{7}$$

式中：p_i 为点 x_i 聚类中心周边点的密集程度，p_i 值越大，则密度越大；z_i 为不同聚类样本点间距离，z_i 的计算公式为：

$$z_i=\sum_{j=1}^{n}(x_i-x_j)^2\text{。}\tag{8}$$

传统 K-means 算法采用误差平方作为目标函数，仅考虑了使类内的距离更小来满足对象紧密性效果，并未考虑到类与类间的距离最大性来保障类间最大分离的效果[14]；因此，引入 K 值的有效性指标利用特征加权距离综合考虑类内紧密型和类间的分离性。

设在数据集 U 中，C_i 为其中的一个类，定义类间距 Intra(C_i)为类 C_i 中任意对象 x, y 的距离平方和，根据式(9)，类距离越小越好；定义类 C_i 到 C_j 距离为 Inter(C_i, C_j)，则根据式(10)，类间距越大越好。

$$\text{Intra}(C_i)=\sum_{x,y\in C_i}\sum_{d=1}^{D}\omega_d^*|x-y|^2\text{；}\tag{9}$$

$$\text{Intra}(C_i,C_j)=\sum_{j=1,j\neq i}^{D}\frac{1}{qp}\sum_{x\in C_i,y\in C_j}\sum_{d=1}^{D}\omega_d^*|x-y|^2\text{。}\tag{10}$$

式中 q，p 为类 C_i 和类 C_j 的样本格式，通过两者的比值来平衡类间和类内距离，定义 K 值的有效性指标为：

$$S=\sum_{i=1}^{k}\text{Intra}(C_i)/\sum_{i=1}^{k}\sum_{j=1,j\neq i}^{k}\text{Inter}(C_i,C_j)\text{。}\tag{11}$$

当 S 最小时，获得类间距离和类内距离的平衡值，即得到最优聚类数据，计算最优聚类数量 K_{best} 为：

$$K_{\text{best}}=\arg\min_{2\leqslant k\leqslant\sqrt{n}}S\text{。}\tag{12}$$

寻找最优 K 值的流程如图 1 所示。首先初始化聚类模型，设定初始聚类数 k=2，随机选取一个初始聚类中心点来计算一个类中的各对象到选定聚类中心点距离，根据每一维聚类数据特征对样本数据的不同聚类来赋予不一样权重。权重考虑到类内和类间对象的特征加权距离，由经验区间 $[2,\text{int}\sqrt{n}]$ 进行划分聚类，寻找有效性指标最小情况下的聚类中心点和聚类数目。

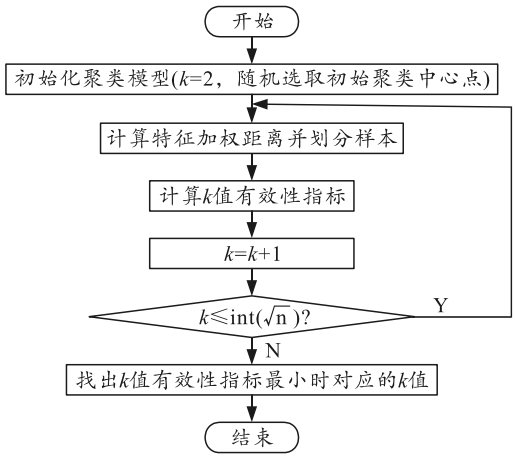

图 1　聚类数目的确定流程

3　算法验证

3.1　实验平台

将改进 K-means 算法运用于网络检测模型构建，设计基于改进 K-means 入侵检测模型，并通过 KDD Cup99 数据集验证模型中的异常检测模块的有效性。图 2 为建立的一种混合式入侵检测模型，分为异常检测模块和误用检测模块，2 种检测模块以级联方式结合在一起。将网络数据首先通过异常检测模块来区分正常行为和入侵行为，使用误用检测模块对入侵数据进行检测，进行具体的入侵行为分类[15]。

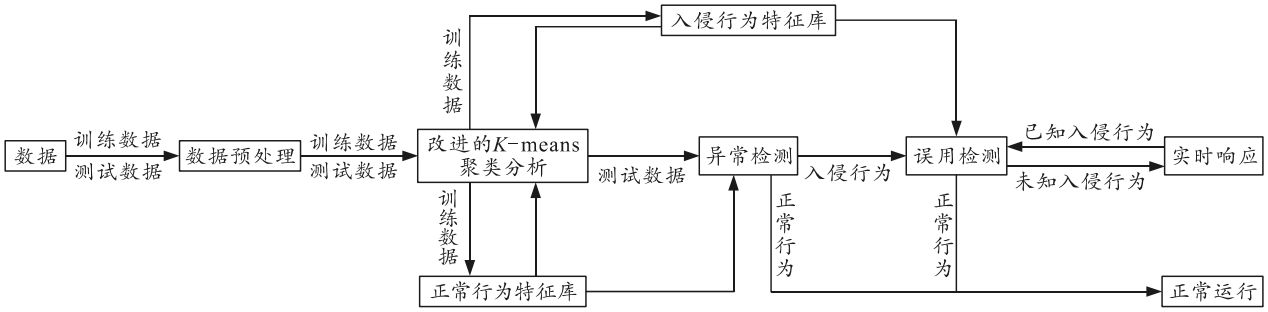

图 2　改进的 K-means 入侵检测模型

模型分为训练和测试阶段。训练阶段首先对训练数据进行标准化预处理，转化为合适的数据格式，

将预处理数据输入改进的 K-means 算法进行处理，获得聚类分析结果。将入侵行为和正常行为划分到不同类簇中，构建正常行为和入侵行为特征库。在测试阶段，首先将测试数据预处理后，与训练得到的正常行为特征库记录匹配，若匹配成功，则该数据记录认定为正常，否则认定为入侵行为，进入误用检测模块。误用检测模块将测试数据与入侵行为特征库进行匹配，若匹配失败，则该数据记录被认为正常行为，若匹配成功，则认定为入侵行为，作用实时响应，并同时将入侵信息反馈至系统管理人员。

3.2 数据的处理

实验基于 Windows 7 系统上采用 Python 编译器进行程序设计，硬件平台为 Interl Core i5 CPU，4GBRAM。入侵检测实验数据集选择 KDD Cup99。KDD Cup99 数据集中提供了 10% 的子集和 Corrected 子集，二者间的攻击类型不一样，且每个数据记录中包含 41 个特征属性和 1 个类别属性，其中正常属性值标定为 normal，而异常数据类别属性分为 DOS 攻击、U2R 攻击、R2L 攻击和 Probe 攻击 4 大类别。

由于 KDD Cup99 数据集中包含符号形的数据属性，而 K-means 算法仅处理数值型属性；因此，首先对 KDD Cup99 数据集中的符号形属性进行赋值处理，如对 protocol_type 属性的 3 种属性值 tcp、icmp、udp 分别转化为 1, 2, 3。同时，为避免不同数据集中特征属性量纲不一致造成变量的值域偏差大，而影响入侵检测结果的判定，本实验选择对数据集进行归一化处理，将数据集数据转换到[0, 1]区间，具体公式为：

$$x=(x-x_{\min})/(x_{\max}-x_{\min})。\tag{13}$$

式中：$x_{\max}$ 为特征属性的最大值；$x_{\min}$ 为特征属性的最小值。将数据集进行归一化处理后，以.csv 文件格式导入 Python 平台进行实验验证。选择检测率(detection rate，DR)和误报率(false positive rate，FPR)2 个评估指标进行入侵检测结果评估，公式定义如下：

$$\mathrm{DR}=\frac{\text{检测到入侵行为个数}}{\text{所有入侵行为个数}}\times 100\%；\tag{14}$$

$$\mathrm{FPR}=\frac{\text{正常行为被误判个数}}{\text{正常行为个数}}\times 100\%。\tag{15}$$

3.3 实验结果

采用改进的 K-means 入侵检测算法确定聚类数据 k 为 25，将 25 作为聚类数目输入传统 K-means 算法模型中，得到不同算法下的检测率和误报率见表 1 所示。

表 1 算法的性能比较 %

算法	DR	FPR
改进 K-means	83.6	2.1
传统 K-means	72.8	8.7

从表 1 中可以看出：改进的 K-means 算法相较于传统的 K-means 算法在检测率和误报率方面都得到了显著优化。

对 2 种算法在已知攻击类型和未知攻击类型分别进行实验，已知攻击是在测试集和训练集中均存在的入侵类型，而未知攻击在训练集中存在，测试集中不存在。从 4 种攻击类型(DOS, U2R, R2L, Probe)中选择某一攻击类型实验，训练集数据从数据样本中选择，测试集数据从 corrected 选取，聚类数 k 取 25，测试结果如表 2 和 3 所示。

表 2 已知攻击类型实验结果 %

类别属性	攻击类型	检测率		误报率	
		改进的 K-means	传统 K-means	改进的 K-means	传统 K-means
R2L	ftp_write	78.5	50.3	3.64	48.60
DOS	smurf	99.0	96.0	2.10	5.73
Probe	ipsweep	95.3	82.6	3.20	19.60
U2R	loadmodule	80.5	53.5	4.10	52.50

表 3 未知攻击类型实验结果 %

类别属性	攻击类型	检测率		误报率	
		改进的 K-means	传统 K-means	改进的 K-means	传统 K-means
R2L	ftp_write	46.7	42.7	5.6	56.4
DOS	smurf	60.7	48.7	4.6	19.7
Probe	ipsweep	65.8	54.7	2.6	5.8
U2R	loadmodule	32.6	30.4	6.8	52.8

可以看出：改进的 K-means 检测算法，在针对未知攻击类型和已知攻击类型数据上的检测效果均优于传统的 K-means 算法。其中针对 U2R 和 R2L 攻击类型数据，表现出高误报率和低检测率的特点，

这是由于该类攻击类型数据通常是以冒充合法身份的方式进行如期攻击的，攻击数据特征相较于正常的数据访问特征非常相似。结果表明：对于单一攻击或者混合攻击类型是，改进的 *K*-means 算法在保持较低误报率基础上，有效提高了入侵检测的检测效率。

4　结论

1) 针对网络数据特征聚类数量无法提前估计问题，提出 *k* 值有效性指标来确定聚类数量和评测聚类质量，依此作为初始聚类中心点，同时考虑各聚类对象紧密型效果，利用特征加权距离考虑类内紧密型和类间的分离性。

2) 将改进 *K*-means 算法融合混合入侵检测模型进行异常入侵检测。实验结果表明，改进 *K*-means 入侵检测算法具有更优的检测率和误报率，能有效提升系统安全防御质量；但对在进行冒充合法身份类型攻击时效果一般，需进一步研究。

参考文献：

[1] 白鹭. *K*-means 算法在通信系统安全防御中的应用与设计[J]. 网络安全技术与应用, 2023(2): 21−22.

[2] 司春波, 赵志强, 高春超, 等. 基于大数据的无线网络优化模式研究[J]. 长江信息通信, 2022, 35(11): 187−189.

[3] 王睿. 基于 *K*-means 算法的网络主动安全防御系统研究与设计[J]. 网络安全技术与应用, 2022(11): 31−33.

[4] 吴南旭, 鲁小琴. 基于K-means算法与蜜罐技术的网络攻击主动防御方法[J]. 电信快报, 2022(10): 31−34.

[5] 黄为. 一种基于大数据的网络安全主动防御系统研究与设计[J]. 网络安全技术与应用, 2022(9): 59−61.

[6] 吕广旭, 卢加奇, 魏先燕, 等. 基于随机森林-聚类混合方法的多分类入侵检测研究[J]. 现代信息科技, 2022, 6(16): 165−167.

[7] 符杨, 顾吉平, 田书欣, 等. 基于地震灾害场景的主动配电网多维韧性评估方法[J]. 电力自动化设备, 2023, 43(3): 1−11.

[8] 李泽龙. 基于机器学习算法的网络安全防御策略探讨[J]. 信息记录材料, 2022, 23(6): 78−80.

[9] 张峰源. 基于分类器联合的铁路时间同步网异常流量自动检测方案[J]. 自动化与仪器仪表, 2022(2): 186−189.

[10] 刘跃鸿. 一种基于人工智能的多层次网络安全体系研究与设计[J]. 网络安全技术与应用, 2021(12): 30−31.

[11] 季赛花, 黄树成. 基于改进的 K-means 入侵检测算法[J]. 计算机与数字工程, 2021, 49(11): 2184−2188.

[12] 赵小林, 赵斌, 赵晶晶, 等. 基于攻击识别的网络安全度量方法研究[J]. 信息网络安全, 2021, 21(11): 17−27.

[13] 刘向举, 路小宝, 方贤进, 等. 软件定义网络环境下的低速率拒绝服务攻击检测方法[J]. 计算机应用, 2022, 42(4): 1301−1307.

[14] 李英. 基于大数据的网络安全防御系统研究与设计[J]. 网络安全技术与应用, 2021(3): 53−54.

[15] 张友鹏, 李响, 兰丽, 等. 基于大数据的铁路时间同步网异常流量检测系统的研究[J]. 铁道科学与工程学报, 2020, 17(2): 306−313.

（上接第 41 页）

[6] 周莽, 高僮, 李晨光, 等. 电网主干数据网络体系结构的设计与实现[J]. 现代信息科技, 2018, 2(12): 56−58.

[7] 汤铭华, 陈健萍, 彭志荣, 等. 基于信息综合判断的移动机器人通信系统故障诊断方法[J]. 自动化与仪器仪表, 2020(11): 51−55.

[8] 施汐, 周鹏, 富思. 基于智慧物联体系及边缘计算的全链路监测系统的开发与应用[J]. 自动化与仪表, 2022(6): 1−5.

[9] 邓智广. 基于贝叶斯分类的电网系统短期负荷预测方法[J]. 电气传动自动化, 2021, 43(5): 28−31.

[10] 郑轶, 王路路, 胡志锋, 等. 泛在物联背景下智慧电力物联网网络安全技术探索[J]. 网络空间安全, 2020, 11(12): 65−72.

[11] 幸茂仁. 无人机电力巡检航迹布设优化方法[J]. 地理空间信息, 2022, 20(10): 117−119, 137.