

一种轻量级网络入侵检测系统模型与算法

蒋昊东, 张宗飞

台州职业技术学院信息技术工程学院, 浙江 台州

收稿日期: 2026年9月18日; 录用日期: 2026年10月17日; 发布日期: 2026年10月26日

摘 要

为了解决网络入侵检测中检测精度低、检测速度慢的难题, 提出了一种轻量级网络入侵检测模型, 并通过设计特征选择算法、检测模型建立算法、入侵检测算法实现了系统模型。在Kddcup99数据集上的实验表明, 系统的入侵检测性能比较好, 检测包含72,788条连接记录的测试集, 检测时间不到1s, 检测率达到91.03%, 误警率控制在3.47%, 检测精度和检测速度均优于对比模型, 然而实验也发现, 系统存在建模时间较长、对样本占比极少的U2R和R2L两种入侵类型检测率偏低的局限性。

关键词

量子进化算法, 特征选择, 聚类分析, 特征子集, 检测模型

A Lightweight Network Intrusion Detection System Model and Algorithm

Haodong Jiang, Zongfei Zhang

School of Information Technology Engineering, Taizhou Vocational & Technical College, Taizhou Zhejiang

Received: September 18, 2026; accepted: October 17, 2026; published: October 26, 2026

Abstract

In order to solve the problems of low detection precision and slow detection speed in network intrusion detection, this paper puts forward a lightweight network intrusion detection system model, and implements the system model by designing feature selection algorithm, detection model building algorithm and intrusion detection algorithm. The experiment results carried out on Kddcup99 datasets verify that the proposed system model has a good detection performance. The detection of the test set containing 72,788 connection records takes less than 1 second, achieving a detection rate of 91.03% and a false alarm rate of 3.47%. Both detection precision and detection speed are

superior to those of the comparison model. However, the experiments also reveal that the system has limitations such as relatively long modeling time and low detection rates for the two intrusion types U2R and R2L, which account for extremely small proportions of the samples.

Keywords

Quantum Evolutionary Algorithm, Feature Selection, Cluster Analysis, Feature Subset, Detection Model

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

互联网应用的普及,一方面给人们的工作、学习和生活带来无比益处,另一方面网络攻击行为也日益增多,严重影响着人们对网络的使用,因此网络入侵检测技术作为一种重要的网络安全主动防御技术,显得尤为重要[1]。随着网络速度的提升,网络流量随之快速扩大,网络入侵检测技术面临着计算负荷大,来不及处理网络中传输的海量数据而造成检测精度低、检测速度慢的难题,如何开发出检测精度高、检测速度快的网络入侵检测系统,成为当前研究的重点[2]。

为了提高网络入侵检测系统的检测精度和检测速度,许多专家和学者对网络入侵检测问题进行相应研究,提出许多有效的系统模型和检测算法[3]-[7]。在网络入侵检测中,通常将网络连接数据分解为用多个特征属性来描述,然后分析特征属性的状态行为,将其划分为正常或入侵,网络入侵检测实质上是模式识别中的二分类问题。许多研究表明,在描述网络连接的多个特征属性,包含噪声特征和冗余特征,这些特征不但会降低入侵检测分类的效率,而且会影响分类的精度,因此如何从网络连接数据中选择能够反映系统状态行为的有效特征参与入侵检测分析是提高网络入侵检测精度和检测速度的关键[8] [9]。

基于上述研究,本文将量子进化算法(Quantum Evolutionary Algorithm, QEA)应用于网络入侵检测,利用 QEA 独特的寻优性能全局优化特征子集和聚类中心,只选取少量有效特征参与运算,并生成以各簇中心为模式特征的检测模型,从而降低系统计算负荷、精简检测模型结构,构建出一种检测计算量少、检测算法简单的轻量级网络入侵检测系统模型,模型架构如图 1 所示。

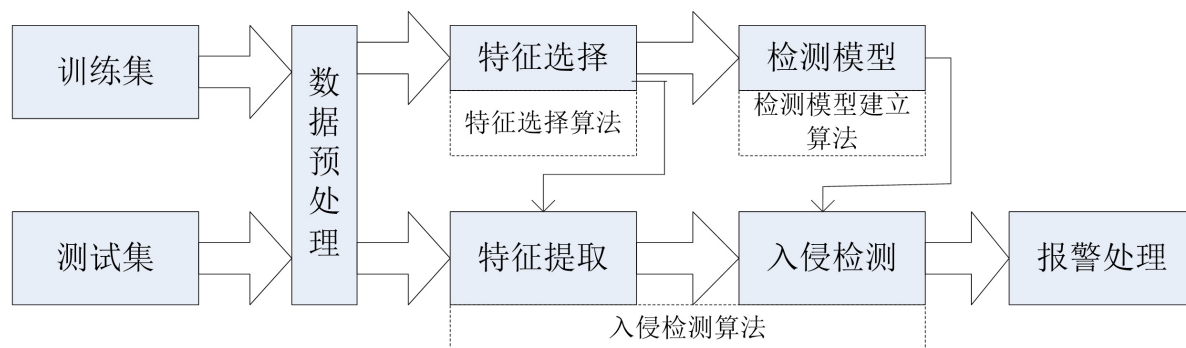


Figure 1. System model architecture

图 1. 系统模型架构

2. 相关技术

2.1. 量子进化算法

QEA 是一种融合量子计算与进化算法的新型智能优化算法[10], QEA 的诱人之处在于采用量子位表示染色体的编码, 通过量子门实现染色体的更新。这种独特的编码与进化机制, 使得 QEA 的优化性能优于进化算法, 但是 QEA 在求解复杂优化问题时, 存在收敛速度慢或收敛于局部最优解的缺陷, 其主要原因在于量子旋转门染色体更新时, 通过查表方式获得的旋转角固定在有限的几个值中, 导致进化后期种群多样性缺失。近年来, 在 QEA 的应用中普遍采用动态旋转角的量子旋转门作为进化算子, 本文提出一种使用角距离表示动态旋转角的策略。

对于组合优化问题, 可以用一对满足归一化条件的实数来表示量子位的概率幅, 此时量子位可以看作位于二维实空间单位圆上的点, 并且第二、三、四象限上的量子位均可以根据其对称性投影到第一象限上。

定义 1 如果量子位 $|\varphi\rangle$ 的概率幅为 $[\alpha, \beta]^T$ (α, β 为实数), 则称 $\arctan \frac{|\beta|}{|\alpha|}$ 为 $|\varphi\rangle$ 的相位角, 记为 $\omega_{|\varphi\rangle}$ ($\omega_{|\varphi\rangle} \in [0, \frac{\pi}{2}]$)。

$$\omega_{|\varphi\rangle} = \begin{cases} \arctan \frac{|\beta|}{|\alpha|} & |\alpha| \neq 0 \\ \frac{\pi}{2} & |\alpha| = 0 \end{cases} \quad (1)$$

定义 2 如果量子位 $|\varphi_1\rangle$ 的相位角为 $\omega_{|\varphi_1\rangle}$, $|\varphi_2\rangle$ 的相位角为 $\omega_{|\varphi_2\rangle}$, 则称 $\omega_{|\varphi_2\rangle} - \omega_{|\varphi_1\rangle}$ 为 $|\varphi_1\rangle$ 到 $|\varphi_2\rangle$ 的角距离, 记为 $\Delta\theta_{|\varphi_1\rangle \rightarrow |\varphi_2\rangle}$ 。

由定义 2 可知, 角距离的大小表示偏转幅度, 角距离的符号表示偏转方向: + 为逆时针方向, - 为顺时针方向。

分析 QEA 中染色体的更新过程可以发现, 其实质就是在量子旋转门的作用下使量子位发生偏转, 改变其取 0 或取 1 的概率, 据此提出一种采用角距离表示的动态调整旋转角 θ_i 的策略。

$$\theta_i = \delta \left(1 - \frac{f_x}{f_b} \right) \cdot \Delta\theta_{|\varphi_i\rangle \rightarrow} \quad (2)$$

其中, f_x 与 f_b 分别表示当前待更新个体与最优个体的适应度值, δ 表示惩罚因子,

$$\Delta\theta_{|\varphi_i\rangle \rightarrow} = \begin{cases} \Delta\theta_{|\varphi_i\rangle \rightarrow |0\rangle} = \omega_{|0\rangle} - \omega_{|\varphi_i\rangle} = -\arctan \frac{|\beta_i|}{|\alpha_i|} & (f_b \geq f_x) \wedge (b_i = 0) \wedge (x_i = 1) \\ \Delta\theta_{|\varphi_i\rangle \rightarrow |1\rangle} = \omega_{|1\rangle} - \omega_{|\varphi_i\rangle} = \frac{\pi}{2} - \arctan \frac{|\beta_i|}{|\alpha_i|} & (f_b \geq f_x) \wedge (b_i = 1) \wedge (x_i = 0) \\ 0 & \text{其它} \end{cases} \quad (3)$$

其中, x_i 与 b_i 分别表示当前待更新个体与最优个体的第 i 个量子位的观测值。

由(2)、(3)式计算旋转角 θ_i 是动态连续的, 避免了查表方式获取旋转角的缺陷, δ 的作用是在进化初期防止 θ_i 的调整幅度过大而陷入局部收敛, 在进化后期进一步缩小 θ_i 的调整幅度有助于实现精细搜索。本文实验中 δ 的值取 0.1。

本文改进了 QEA 的优点是对量子旋转门的旋转角实现了自适应调整, 能够随种群进化程度动态连续调整旋转角的偏转幅度, 较好地取得了快速收敛与全局收敛的平衡, 使其优化性能超过基准 QEA。本文中 QEA 的作用是优化特征子集和聚类中心, 而在优化之前均已进行初步处理而获得了比较规范的数据集, 从而降低了优化难度, 我们的实验表明使用改进的 QEA 进行优化, 不需要交叉和变异操作就可以实现对解空间的全方位搜索。

2.2. Fisher 分

Fisher 分(Fisher Score)是特征选择研究中衡量特征鉴别能力的一种度量方法[11], 其主要思想是鉴别能力强的特征表现为它的类内离散度小、类间离散度大。在网络入侵检测中, 可以将网络连接数据划分为正常数据和入侵数据。据此本文给出单特征及特征子集的 Fisher 分定义。

定义 3 第 i 维特征在网络连接数据集上的类间离散度 $S_b^{(i)}$ 与类内离散度 $S_w^{(i)}$ 的比值称为该特征的 Fisher 分, 记为 $fs^{(i)}$ 。

$$fs^{(i)} = \frac{S_b^{(i)}}{S_w^{(i)}} = \frac{\sqrt{(m_+^{(i)} - m^{(i)})^2} + \sqrt{(m_-^{(i)} - m^{(i)})^2}}{\frac{1}{n_+} \sum_{x \in X_+} \sqrt{(x^{(i)} - m_+^{(i)})^2} + \frac{1}{n_-} \sum_{x \in X_-} \sqrt{(x^{(i)} - m_-^{(i)})^2}} \quad (4)$$

定义 4 由 k 个特征构成的特征子集在网络连接数据集上的类间离散度 S_b 与类内离散度 S_w 的比值称为该特征子集的 Fisher 分, 记为 fs 。

$$fs = \frac{S_b}{S_w} = \frac{\sum_{i=1}^k \left(\sqrt{(m_+^{(i)} - m^{(i)})^2} + \sqrt{(m_-^{(i)} - m^{(i)})^2} \right)}{\sum_{i=1}^k \left(\frac{1}{n_+} \sum_{x \in X_+} \sqrt{(x^{(i)} - m_+^{(i)})^2} + \frac{1}{n_-} \sum_{x \in X_-} \sqrt{(x^{(i)} - m_-^{(i)})^2} \right)} \quad (5)$$

2.3. 流形距离

在聚类分析中, 欧氏距离是最常用的数据对象相似性的度量方法, 但是对于非线性结构数据, 使用欧氏距离度量会出现偏差。随着数据分析技术的发展, 人们已发现网络连接数据中存在非线性结构, 因此使用欧氏距离来度量其相似性会对计算结果产生影响。流形学习作为一种高效的处理非线性结构数据的方法在近年得到广泛研究[12], 借鉴流形学习研究成果, 本文选择基于流形学习方法的流形距离来度量网络数据对象的相似性。

在流形距离度量中, 将数据集中的数据看作无向完全图的顶点, 据此给出流形距离定义如下。

定义 5 连接两个顶点 x_i 、 x_j 之间的边长度称为这两顶点的边距, 记为 $E(x_i, x_j)$ 。

$$E(x_i, x_j) = e^{d(x_i, x_j)} - 1 \quad (6)$$

其中, x_i 、 x_j 表示数据集中的两个数据对象, $d(x_i, x_j)$ 表示 x_i 、 x_j 之间的欧氏距离。

定义 6 连接两个顶点 x_i 、 x_j 之间的最短路径称为这两顶点的流形距离, 记为 $D(x_i, x_j)$ 。

$$D(x_i, x_j) = \min_{p \in P_{ij}} \sum_{k=1}^{|p|-1} E(p_k, p_{k+1}) \quad (7)$$

其中, p 是连接顶点 x_i 、 x_j 之间的某一条路径, P_{ij} 是连接顶点 x_i 、 x_j 之间所有路径的集合, $|p|$ 是路径 p 的顶点个数, p_k 是路径 p 的第 k 个顶点。

3. 特征选择

系统中特征选择的目的是去除网络连接数据中的噪声特征和冗余特征, 降低特征维数, 从而减少检测时的计算量, 提高检测速度和检测精度。

本文采用两级特征选择, 先根据特征的 Fisher 分值过滤噪声特征, 然后对去噪声特征后的特征向量进一步优化, 去除冗余特征。

3.1. 特征过滤

由定义 3 可知, Fisher 分值大的特征, 识别入侵的能力也强, 据此本文提出一种基于 Fisher 分的特征过滤算法 FS-Fisher, 算法基本思想是设置一个阈值 ε , 将 Fisher 分值低于该阈值的特征作为噪声特征被过滤。

算法 1 基于 Fisher 分的特征过滤算法 FS-Fisher

输入: 网络连接数据集 X , 原始特征集 F

输出: 初次特征优化的特征子集 F_Sub , 初次特征优化的数据子集 X_Sub

$F_Fisher = \text{calucateFisherScore}(X, F); /* \text{计算特征集中各特征的 Fisher 分值} */$

$F_Fisher = \text{sort}(F_Fisher, \text{desc}); /* \text{按照 Fisher 分值对特征降序排列} */$

$F_Sub = \text{deleteNoiseFeatures}(F_Fisher, \varepsilon); /* \text{将 Fisher 分值小于 } \varepsilon \text{ 的特征作为噪声特征删除, 得到初次优化的特征子集 } F_Sub */$

$X_Sub = \text{generateDataSubset}(X, F_Sub); /* \text{去除噪声特征后, 得到初次优化的数据子集 } X_Sub */$

3.2. 特征优化

特征过滤后的特征子集还存在冗余特征, 需要进一步优化。本文提出一种基于 QEA 的特征优化算法 FS-QEA, 算法的基本思想是使用本文改进的 QEA 对特征过滤后的特征子集进行优化, 筛选出剔除了冗余特征的近似最优特征子集, 实现过程是对特征子集进行编码并生成初始种群, 然后在适应度函数的指导下逐代进化, 最后输出最优个体。

3.2.1. 算法设计

QEA 优化的流程包括: 编码、种群初始化、种群观测、种群个体适应度计算、种群进化, 当达到最大进化代数后, 输出最优个体。FS-QEA 中的种群初始化和种群观测方式与 QEA 相同, 本文仅对编码方案、适应度函数、种群进化机制的设计过程进行分析。

(1) 编码方案

根据特征选择的特点, 采用一个二进制字符串 $(s_1 s_2 s_3 \cdots s_m)$ 来表示特征集的一个子集, 其中, m 表示特征集的特征维度, $s_i \in \{0, 1\}$, $s_i = 1$ 表示特征集中的第 i 个特征被选择, 反之, $s_i = 0$ 表示不选择第 i 个特征。根据这一思路, 在 FS-QEA 中可以用 m 位量子位概率幅构成的量子染色体来表示特征子集。

(2) 适应度函数

适应度函数是算法中衡量个体优劣的标志, FS-QEA 中的个体就代表特征子集, 因此其优劣可以用 2 个指标来度量: 入侵识别能力和特征维度, 如果一个特征子集的入侵识别能力越强、特征维度越低, 则表示该特征子集越优秀, 其对应个体的适应度就应该越大。特征子集的入侵识别能力可以用定义 4 的 Fisher 分表示, 特征维度可以用对应观测态个体中“1”的个数表示, 据此适应度函数可以设计为:

$$f = fs \cdot e^{-np} \quad (8)$$

其中, fs 表示个体所代表的特征子集的 Fisher 分, np 表示个体观测值中“1”的个数。

(3) 染色体更新

染色体更新是实现种群进化的主要机制, 使用(9)式的量子旋转门按照(10)式更新量子位来完成染色体更新过程。

$$U(\theta) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \quad (9)$$

其中, θ 为旋转角度。

$$\begin{bmatrix} \alpha_i^{t+1} \\ \beta_i^{t+1} \end{bmatrix} = U(\theta_i) \begin{bmatrix} \alpha_i^t \\ \beta_i^t \end{bmatrix} = \begin{bmatrix} \cos(\theta_i) & -\sin(\theta_i) \\ \sin(\theta_i) & \cos(\theta_i) \end{bmatrix} \begin{bmatrix} \alpha_i^t \\ \beta_i^t \end{bmatrix} \quad (10)$$

其中, $[\alpha_i^t, \beta_i^t]^T$ 为更新染色体第 i 个量子位的概率幅, $[\alpha_i^{t+1}, \beta_i^{t+1}]^T$ 为更新后染色体第 i 个量子位的概率幅。 θ_i 为更新第 i 个量子位时的旋转角, 按(2)、(3)式进行自适应动态调整。

3.2.2. 算法描述

算法 2 基于 QEA 的特征优化算法 FS-QEA

输入: 初次特征优化的特征子集 F_Sub, 初次特征优化的数据子集 X_Sub

输出: 近似最优特征子集 F_Final, 近似最优特征的数据子集 X_Final

t=0; B[m]=0; fB=0; /*t:进化代数, B[m]:最优个体, fB:最优个体的适应度*/

m=F_Sub.size(); /*m 为 F_Sub 的特征维度*/

Q[n, (α, β)_m] = initialize(n, $\begin{bmatrix} \alpha \\ \beta \end{bmatrix}$, m); /*初始化种群, n:种群规模, $\begin{bmatrix} \alpha \\ \beta \end{bmatrix}$:量子位概率幅*/

while(t<=500){ /*最大进化代数为 500*/

P[n, {0,1}_m] = measure(Q[n, (α, β)_m); /*观测量子态种群, 获得观测态种群*/

fitness[n] = calculateFitness(P[n, {0,1}_m], f); /*计算各个体的适应度, f:适应度函数*/

fb = getBestFitness(fitness[n]); /*获取当代最大适应度*/

b[m] = getBestMember(P[n, {0,1}_m], fb); /*获取当代最优个体*/

if(fb > fB){

fB = fb; B[m] = b[m]; /*保留当前最优个体及其适应度*/

}

Q[n, (α, β)_m] = updateMembers(Q[n, (α, β)_m); /*更新染色体*/

t=t+1;

}

F_Final = decode(B[m]); /*解码最优个体获得近似最优特征子集*/

X_Final = generateOptimalDataSubset(X_Sub, F_Final); /*去除冗余特征后, 得到近似最优特征的数据子集*/

4. 网络入侵检测

系统以聚类为工具进行入侵检测, 首先根据训练样本建立检测模型, 然后确定检测方法对网络连接数据展开检测。

4.1. 入侵检测模型

以聚类为工具的入侵检测系统, 检测模型库中存储的是聚类中心, 决定系统检测性能的关键因素是

聚类中心优化的质量, 聚类中心优化越彻底, 得到检测模型就越精准, 检测过程也会越简单。在众多聚类算法类型中, K-Means 以内存消耗小、收敛速度快、能有效处理大数据集等优点而被广泛应用, 但是 K-Means 也存在明显的缺点, 主要表现在: (1) 受噪声数据影响大, (2) 对初始聚类中心敏感, (3) 易收敛于局部聚类结果。本文是对特征选择后的数据集进行聚类, 数据已经过去噪声处理, 针对(2)、(3)两项缺点, 我们采取先根据数据密度生成初次优化的聚类中心, 再使用 QEA 的优化性能对该聚类中心进行全局优化, 获得全局近似最优聚类结果, 然后对聚类簇标号生成较为精准的入侵检测模型, 从而降低入侵检测算法的复杂度。

4.1.1. 确定初始聚类中心

根据数据对象的密度对数据集进行划分, 从而确定初始聚类中心, 首先给出数据对象密度权重的定义。

定义 7 将数据对象 x_j 与数据集中其它数据对象之间的距离之和与数据集中各数据对象与其它数据对象之间的距离之和的比值称为 x_j 的密度权重, 记为 $dw^{(x_j)}$

$$dw^{(x_j)} = \frac{\sum_{i=1, i \neq j}^n D(x_j, x_i)}{\sum_{j=1}^n \sum_{i=1, i \neq j}^n D(x_j, x_i)} \quad (11)$$

其中, $D(x_j, x_i)$ 表示数据对象 x_j 与 x_i 之间的流行距离。

由定义 7 可知, 如果 $dw^{(x_j)}$ 值越大, 则数据对象 x_j 周围的数据就越多, 从而对其它数据的影响也就越大; 反之, 如果 $dw^{(x_j)}$ 值越小, 则数据对象 x_j 就越疏远于其它数据, 从而对其它数据的影响也就越小。基于这一思想, 本文提出了一种基于密度权重划分初始聚类的算法 Cluster-DW。

算法 3 基于密度权重确定初始聚类中心算法 Cluster-DW

输入: 特征选择后的数据集 X_Final, 聚类数 K

输出: 初始聚类中心 V^0 , 初始簇 C^0

$X=X_Final$; /*X 的初始值为 X_Final 中的数据集合*/

for($i=1$; $i \leq k$; $i++$) {

$DW=calculated(X)$; /*DW 为数据集 X 中各数据对象的密度权重集合*/

$max_DW=getMax_DW(DW)$; /*max_DW 为 DW 中的最大密度权重值*/

$x_{max_DW}=getX_DW(X, max_DW)$; /* x_{max_DW} 为数据集 X 中最大密度权重的数据对象*/

$m=X.size()$; /*m 为数据集 X 的大小*/

$\bar{d} = \frac{\sum_{i=1, i \neq j}^n D(x_{max_DW}, x)}{m-1}$; /* $\bar{d}$ 为 x_{max_DW} 到数据集 X 中其它数据对象的平均距离*/

$C_i=class(X, x_{max_DW}, \bar{d})$; /*以 x_{max_DW} 为中心, $\bar{d}$ 为半径对数据集 X 划分, 得到簇 C_i */

$X=X-C_i$; /*从数据集 X 中删去簇 C_i 中的数据对象*/

}

$C^0=fusionCluster(C_1, C_2, \dots, C_k, X)$; /*将数据集 X 中余下数据合并到与之最近的簇中, 得到初始聚类 C^0 */

$V^0=getClusterCenter(C^0)$; /* V^0 为初始聚类中心*/

4.1.2. 聚类中心优化

得到初始聚类中心后, 按照 K-Means 算法思想, 使用本文改进的 QEA 对聚类中心进行全局优化,

获得近似最优聚类中心及聚类簇。根据这一思路,提出了一种基于 QEA 的聚类中心优化算法 Cluster-QEA。

(1) 算法设计

模仿 3.2.1 节,我们仅对与 QEA 流程不同的编码方案、观测方法及适应度函数的设计过程进行分析,由于种群进化方法与 FS-QEA 类似,故也不在本节分析。

(a) 染色体编码

根据聚类的特点,我们以聚类结果来编码染色体,假设数据集中 n 个数据对象被划分成 k 类,用基因串 $(a_1 a_2 a_3 \cdots a_n)$ 表示某一聚类结果,其中: $k \geq 2$, $1 \leq a_i \leq k$, $(i=1,2,\cdots,n)$, a_i 的值表示第 i 个数据对象的簇序号,例如 $a_i = 5$ 则表示数据集中第 i 个数据对象属于第 5 簇。 a_i 需用 $(\text{INT}(\log_2^k)+1)$ 位二进制来表示,因此染色体的长度为 $n * (\text{INT}(\log_2^k)+1)$, 每 $(\text{INT}(\log_2^k)+1)$ 位为一个基因段,表示一个数据对象。根据这一编码方案, Cluster-QEA 中可用 $n * (\text{INT}(\log_2^k)+1)$ 位量子位概率幅表示一个量子染色体。

(b) 种群观测

为了保证算法从初次优化的聚类中心开始进化,在种群初始化后的第一次观测中,将各个体都观测成 4.1.1 节中得到的初始聚类中心的结果,然后在个体适应度的指导下进化并按照 QEA 中的方法进行观测。

(c) 适应度函数设计

K-Means 算法是基于目标函数的聚类分析方法,而 Cluster-QEA 的适应度函数可以根据目标函数转化而得到。

首先,我们将 K-Means 的目标函数 J 进行适当转换,用本文提出的流形距离来表示:

$$J = \sum_{k=1}^n \sum_{i=1}^c u_{ik} D(x_k, V_i) \quad (12)$$

其中, $D(x_k, V_i)$ 表示第 k 个数据对象到第 i 个聚类中心的距离, $u_{ik} \in \{0,1\}$, $u_{ik} = 1$ 表示第 k 个数据对象属于第 i 个簇, $u_{ik} = 0$ 表示第 k 个数据对象不属于第 i 个簇。

对(12)式的目标函数进行转换,得到 Cluster-QEA 的适应度函数为:

$$f = \frac{1}{1+J} \quad (13)$$

(2) 算法描述

算法 4 基于 QEA 的聚类中心优化算法 Cluster-QEA

输入: 初始聚类中心 V^0 , 初始簇 C^0

输出: 近似最优聚类中心 V , 近似最优聚簇 C

(算法流程与算法 2 类似,故不做描述)

4.1.3. 聚簇标类

得到最优聚簇 $C = \{C_1, C_2, \cdots, C_k\}$ 后,还需要对这 k 个簇进行类标识,即确定是正常类还是异常类,从而获得入侵检测模型。最优聚簇是由带类标识的训练数据集生成,因此可以根据各簇中的正常连接与异常连接的数据比例来确定聚簇的类标识,标类算法如下。

算法 5 聚簇标类算法 Cluster-Label

输入: 近似最优聚簇 C

输出: 带类标签的聚簇 C^{label}

for($i=1; i \leq k; i++$) {

```

    num+, num- = countNumber(Ci); /* num+ 为 Ci 中正常网络连接的数目, num- 为 Ci 中入侵网络连接的数目 */
    if(num+ > num-){
        Ci+ = Ci; /* Ci+ 为正常连接类聚簇 */
    }
    Else{
        Ci- = Ci; /* Ci- 为入侵连接类聚簇 */
    }
}

```

4.2. 入侵检测

系统经训练数据集训练, 即经过预处理、特征选择、聚类分析后, 建立了入侵检测模型, 模型库中聚类中心的特征维度低、优化程度高, 因此降低了入侵检测的计算量和计算复杂度。检测算法如下。

算法 6 入侵检测算法 Detection-Cluster

输入: 网络连接数据包 x

输出: 检测结果

读取一个网络连接数据包 x ;

对 x 进行预处理;

$X' = \text{selectFeatures}(x, F_Final)$ /*根据最优特征子集剔除 x 中的噪声特征和冗余特征*/

$V' = \text{findNearestClusterCenter}(x', V)$ /*获取与 x' 最近的簇中心*/

if($v' \in C^+$){

x 为正常连接, 扔掉该数据包;

}else{

x 为异常连接, 报警处理

}

5. 实验与结果分析

5.1. 实验准备

5.1.1. 数据集选取

实验使用 Kddcup99 数据集[13], 它是目前网络入侵检测研究中常用的实验数据集之一, 包含训练集(kddcup.data.gz)和测试集(corrected.gz)。实验中, 训练集选择 kddcup.data.gz 的一个 10%子集(kddcup.data-10-percent.gz), 包含 494,020 条连接记录, 去重处理后得到 145,582 条连接记录; 测试集从 corrected.gz 中按 Normal、DoS、Probe、U2R、R2L 所占比例选取 72,788 条不重复连接记录。

5.1.2. 数据预处理

Kddcup99 数据集中的每条数据连接记录包含 41 个特征属性, 其中 7 个符号特征, 34 个数值特征, 符号特征值不能参与文中的距离计算; 而数值特征包含不同的量纲, 如果直接参与计算就会出现大数吃小数问题, 因此需要对训练集和测试集数据进行标准化处理, 包括符号特征数值化和数值特征归一化。

对于符号特征, 采取十进制编码方式将符号特征值转化为数值类型; 对于数值特征(包括数值化后的符号特征), 按下列算式进行归一化。

Step1 计算数据集所有数据第 k 维特征的平均绝对偏差 $\bar{m}^{(k)}$;

$$\bar{m}^{(k)} = \frac{1}{n} \sum_{i=1}^n (x_i^{(k)} - m^{(k)}) \quad (14)$$

其中, $x_i^{(k)}$ 为数据集中第 i 个连接记录的第 k 维特征值, $m^{(k)}$ 为数据集中所有数据第 k 维特征的平均值:

$$m^{(k)} = \frac{1}{n} \sum_{i=1}^n x_i^{(k)} \quad (15)$$

Step2 计算数据集第 i 个连接记录第 k 维特征的标准化值 $\bar{x}_i^{(k)}$:

$$\bar{x}_i^{(k)} = \frac{x_i^{(k)} - m^{(k)}}{\bar{m}^{(k)}} \quad (16)$$

5.1.3. 环境搭建

实验在一台 Intel(R) Core(TM) i5-12400@2.50 GHz、8.00 GB RAM, Windows 10 专业版的台式机上
进行, 使用 Python 语言结合第三方机器学习库完成算法代码。

5.1.4. 参数确定

文中算法需要确定特征过滤阈值 ε 和聚类数 k 2 个参数, 我们采用试探法进行多组试验, 得出参数 ε 取值为 0.1、参数 k 取值为 2 时, 结果最为理想。

5.2. 对比验证

为了验证本文提出的网络入侵检测系统模型的性能, 采用对比实验, 与基于常规 K-Means 构建的入侵检测系统模型(即不经过特征选择和聚类中心优化)进行比较, 选取建模时间、测试时间、检测率、误警率 4 个指标来衡量系统性能。

首先以预处理后的训练集对本文提出的系统模型进行训练, 即经过特征选择、聚类分析、聚簇标类步骤后, 得到入侵检测模型, 记录建模时间; 再使用预处理后的测试集按照文中提出的检测算法进行入侵检测测试, 记录测试时间、检测率、误警率 3 个指标值。进行 10 次验证, 并统计建模时间、测试时间、检测率、误警率的平均值。

然后以同样的训练集对基于常规 K-Means 构建的入侵检测系统模型进行训练, 即使用常规 K-Means 对数据集进行聚类分析(参数 k 也取值为 2), 得到入侵检测模型, 记录建模时间; 再使用同样测试集参照文中提出的检测算法进行入侵检测实验, 记录测试时间、检测率、误警率 3 个指标值。也进行 10 次验证, 并统计建模时间、测试时间、检测率、误警率的平均值。

5.3. 实验结果分析

Table 1. Test results of the system models

表 1. 系统模型测试结果

系统模型	建模时间(s)	测试时间(s)	检测率(%)	误警率(%)
本文模型	74.0039	0.9997	0.9103	0.0347
对比模型	70.6014	1.7799	0.8591	0.0396

实验结果统计如表 1 所示, 本文模型训练时需要训练集进行特征选择和聚类中心优化, 因此模型的训练时间较对比模型要长。相较于入侵检测过程, 模型训练是建模过程, 在系统使用前进行, 不会对系统的检测性能产生影响。从表 1 数据可以发现, 本文模型在测试时间、检测率、误警率 3 个指标都要优于对比模型, 表明本文模型对入侵的检测速度和检测精度都表现更好, 分析其原因可以得出: 本文提

出的特征选择策略能够筛选出反应系统状态行为的有效特征, 从而降低系统计算负荷; 本文提出的聚类中心优化策略能够逼近全局最优聚类结果, 从而获得较为精准的入侵检测模型, 降低了入侵检测算法的复杂度。

6. 结束语

本文针对当前网络入侵检测中的问题, 提出了一种轻量级网络入侵检测系统模型, 并设计算法实现了系统模型。从实用性角度看, 系统结构比较简单, 检测模型较为精准, 检测计算量较小, 适合于在线检测。但是我们在实验中发现系统对 U2R 和 R2L 两种入侵类型的检测率偏低, 经过分析得知是由于这两种入侵数据在训练集中所占比例严重偏少, 影响了聚类质量, 这个问题需要今后继续研究。

基金项目

2025 年浙江省大学生科技创新项目(2025R486A005)。

参考文献

- [1] 杨宇, 陈一丁, 赵荣, 等. 网络安全主动防御研究综述[J]. 科学技术与工程, 2025, 25(7): 2654-2663.
- [2] Liu, J.Y. and Sun, W. (2026) Network Intrusion Detection Techniques Based on Improved Multi-Scale 1D Convolutional Neural Networks. *Engineering Research Express*, **8**, Article ID: 035110. <https://doi.org/10.1088/2631-8695/ae3cdb>
- [3] Das, S., Saha, S., Priyoti, A.T., Roy, E.K., Sheldon, F.T., Haque, A., *et al.* (2022) Network Intrusion Detection and Comparative Analysis Using Ensemble Machine Learning and Feature Selection. *IEEE Transactions on Network and Service Management*, **19**, 4821-4833. <https://doi.org/10.1109/tnsm.2021.3138457>
- [4] 张茜, 王晓菲, 王亚洲, 等. 基于机器学习的网络入侵检测技术综述[J]. 网络安全与数据治理, 2024, 43(12): 1-9, 18.
- [5] Wang, Y.L., Li, J., Han, Z.J., Cheng, P. and Kumar, R. (2025) FedSTGCN: A Novel Federated Spatiotemporal Graph Learning-Based Network Intrusion Detection Method for the Internet of Things. *Frontiers of Information Technology & Electronic Engineering*, **26**, 1164-1179. <https://doi.org/10.1631/fitee.2400932>
- [6] Nayak, V., Pawar, S. and Kumar, B.L.S. (2026) A Review of Machine and Deep Learning Techniques for Network Intrusion Detection. *International Journal of Communication Networks and Distributed Systems*, **32**, 136-182. <https://doi.org/10.1504/ijcnds.2026.152133>
- [7] Vinod, D. and Prasad, M. (2026) Enhancing Network Security: A Novel Intrusion Detection System Utilizing Dual-Optimization Techniques for Feature Selection and Classification. *Computer Networks*, **277**, Article ID: 112021. <https://doi.org/10.1016/j.comnet.2026.112021>
- [8] 任家东, 张亚飞, 张炳, 等. 基于特征选择的工业互联网入侵检测分类方法[J]. 计算机研究与发展, 2022, 59(5): 1148-1159.
- [9] 刘景美, 闫义博. 人工鱼群特征选择的网络入侵检测系统[J]. 西安电子科技大学学报, 2023, 50(4): 132-138.
- [10] Han, K.H. and Kim, J.H. (2002) Quantum-Inspired Evolutionary Algorithm for a Class of Combinatorial Optimization. *IEEE Transactions on Evolutionary Computation*, **6**, 580-593. <https://doi.org/10.1109/tevc.2002.804320>
- [11] 汪正凯. 基于 Fisher Score 的多标记特征选择算法研究[D]: [硕士学位论文]. 漳州: 闽南师范大学, 2021.
- [12] 王博林. 基于流形学习的文本表示与分类研究[D]: [博士学位论文]. 大连: 大连理工大学, 2023.
- [13] KDD Cup 1999 Data. <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>