Computer Science and Application 计算机科学与应用, 2022, 12(9), 2128-2136
Published Online September 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.129216

Hans汉斯

一种智能化网络入侵检测模型设计

姚宇杰，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：2022年8月16日；录用日期：2022年9月14日；发布日期：2022年9月21日

摘 要

本文针对当前网络入侵检测系统在识别新型入侵行为时存在误警率和漏警率偏高的弊端，将新型优化技术和机器学习技术引入网络入侵检测，设计了一种具有智能性的网络入侵检测模型，实验模拟表明，本文设计的模型是有效的，并且对新型网络攻击行为的识别能力比较好，为开发新型的网络入侵检测系统提供了设计思路。

关键词

网络入侵检测，量子进化算法，k-Means，新型入侵

Design of an Intelligent Network Intrusion Detection Model

Yujie Yao, Zongfei Zhang

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

Received: Aug. 16th, 2022; accepted: Sep. 14th, 2022; published: Sep. 21st, 2022

Abstract

Aiming at the disadvantages that current network intrusion detection systems have high false alarm rate and missing alarm rate when identifying new types of intrusion, new optimization technology and machine learning technology are applied to network intrusion detection, and an intelligent network intrusion detection model is designed in this paper. Simulation results show that the model designed in this paper is effective, and has better ability to identify new types of network attacks, provides a design idea for developing new network intrusion detection system.

文章引用: 姚宇杰, 张宗飞. 一种智能化网络入侵检测模型设计[J]. 计算机科学与应用, 2022, 12(9): 2128-2136.
DOI: 10.12677/csa.2022.129216

Keywords

Network Intrusion Detection, Quantum Evolutionary Algorithm, k-Means, New Types of Intrusion

1. 引言

网络安全问题伴随计算机网络而存在，互联网应用的普及，使得网络安全问题越来越突出，世界各国每年都会报道出大量的网络安全事件。针对网络系统存在的安全隐患，出现了加密与数字签名、身份认证与访问控制、防火墙、网络入侵检测等多种网络安全防御技术，其中网络入侵检测采用主动防御技术，弥补了防火墙技术的局限性，成为目前最主流的网络安全防御技术。近年来以机器学习为代表的人工智能技术发展迅速，在多个领域取得突破性进展，与此同时，采用机器学习算法进行网络入侵检测也受到广大研究人员的关注，多种机器学习模型应用于网络入侵检测中，如文献[1]研究了基于支持向量机模型的网络入侵检测，首先采用遗传算法对支持向量机惩罚系数和核参数进行优化，然后设计了用于网络入侵检测的 GA-SVM 算法，将其应用于 KDDCup99 数据集中 4 种网路入侵数据的检测，获得了较好的检测精度；文献[2]研究了基于极限梯度提升决策树模型的网络入侵检测，首先采用人工蜂群算法提取网络连接的特征，然后利用极限梯度提升 XGBoost 算法将特征进行分类，获得最优特征子集，利用这些特征完成网络异常检测，获得了较高的检测效率；文献[3]研究了基于 k 近邻模型的网络入侵检测，首先利用极限学习机算法将低维度线性不可分样本转换为高维特征空间中的线性可分样本，然后用 k 近邻算法对高维特征空间中的样本进行分类，建立了网络入侵检测分类器，提高了入侵检测正确率。然而人们在享受技术成果时，也不可避免面临着新技术带来的威胁，随着人工智能技术的发展和应用，网络攻击行为不断升级，攻击类型不断更新，导致目前的网络入侵检测系统在使用过程中不断出现误警和漏警现象，网络入侵检测技术正面临严峻的挑战[4]。

如何提升网络入侵检测的智能性，识别出新型网络攻击行为，是当前网络入侵检测研究需要解决的难题，为此本文将新型优化技术和人工智能技术引入网络入侵检测中，设计了一种网络入侵检测特征模式生成算法，使用该算法生成的特征模式能够根据检测过程自动更新簇中心，从而具备识别新型入侵的能力。

2. 相关技术

2.1. k 均值聚类算法

k 均值聚类算法(k-means clustering algorithm, k-means)是基于划分的聚类分析方法[5]，是一种经典的机器学习算法模型，具有实现过程简单，时空复杂度低，数据处理效率高等优点，在工业、科学等领域中得到广泛应用。k-means 的算法流程如图 1 所示。

2.2. 量子进化算法

量子进化算法(Quantum Evolutionary Algorithm, QEA)是量子计算与进化算法相融合的一种新型智能

优化算法[6]，QEA 使用量子比特编码染色体来代表问题的可行解，通过量子门更新染色体来进化种群个体，这种独特的编码和进化机制，使得算法具有优异的寻优能力和寻优速度，在求解组合优化问题时表现出显著的全局优化性能。QEA 的算法流程如图 2 所示。

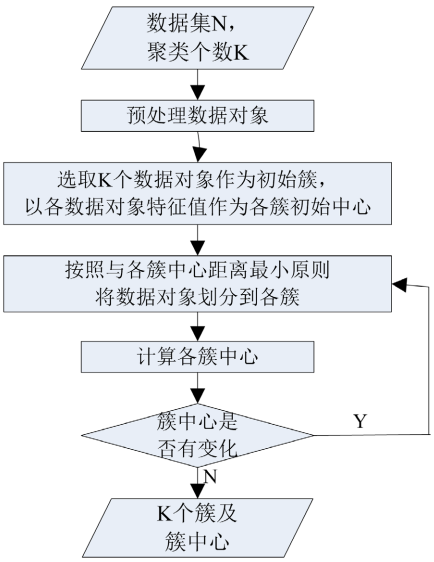

Figure 1. Algorithm procedure of k-means
图 1. k-means 算法流程

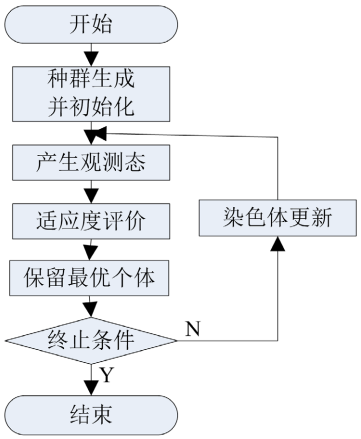

Figure 2. Algorithm procedure of QEA
图 2. QEA 算法流程

2.3. 基于 QEA 的 k 均值聚类算法

虽然 k-means 被广泛应用，但是在应用过程中也发现算法的局限性也相当明显，主要表现在迭代寻找最优聚类中心时容易陷入局部极值，无法获得全局最优解，主要原因是算法的优化能力有限，在对结构复杂的数据集进行聚类时难以获得最佳聚类中心，导致聚类质量不够理想。网络入侵检测研究中对网络连接数据包分析表明，网络连接数据特征结构复杂，数据维度高，为此本文首先对 k-means 进行改进，引入 QEA，利用 QEA 显著的全局优化能力来弥补 k-means 的不足，提出一种基于 QEA 的 K 均值聚类算法，记为 k-means-QEA。

算法：k-means-QEA

输入：数据集 X，聚类个数 k，种群大小 n

输出：k 个簇以及簇中心

流程：

Step 1 种群生成：构造量子染色体，生成包含 n 个个体的初始种群；

Step 2 种群初始化：对初始种群中的个体进行初始化，获得初始化后的种群 $Q(t)$；

Step 3 种群观测：对 $Q(t)$中各个体进行测量，获得观测态种群 $P(t)$；

Step 4 种群评价：使用适应度函数对 $P(t)$中各个体进行评价，记录最优个体；

Step 5 终止条件判断：如果满足终止条件，则算法停止迭代转 Step 8，否则转 Step 6；

Step 6 种群进化：使用量子门对 $Q(t)$中各个体进行更新，获得进化后新种群；

Step 7 数据聚类：以新种群个体为聚类中心，进行 n 次 k-means 划分，获得新一代种群 $Q(t)$，转 Step 3；

Step 8 解码输出：对最优个体进行解码得到 k 个簇中心，并聚类得到 k 个簇。

3. 特征模式生成算法

本文基于 k-means-QEA，设计了一种入侵检测特征模式生成算法，记为 FMGA。

3.1. 算法设计过程

FMGA 设计的基本思想是使用 k-means-QEA 对网络连接数据集进行聚类，获得簇中心，并对各簇中心标记“正常”或“异常”的标签。

3.1.1. 量子染色体构造

量子染色体就是 QEA 中的个体，构造量子染色体需要对所求问题的解进行编码。算法需要对聚类中心进行优化，为此针对 k 个簇中心，按照 QEA 中量子染色体的结构进行编码，编码方案如下：

Step1：将 k 个簇中心向量值按序连接成数字串；

Step2：将数字串中每一位数字转化为二进制，转化时二进制的位数取当前位数字所代表的特征属性值的最大值所需的二进制位。

3.1.2. 种群生成

从数据集 X 中随机选择 k 个数据对象作为初始聚类中心并编码构成量子染色体，重复 n 次生成由 n 个个体构成的初始种群 $Q(t)=\left\{q_1^t,q_2^t,\cdots,q_n^t\right\}$。

3.1.3. 种群初始化

将初始种群 $Q(t)$中量子染色体的各量子比特初始化为等概率状态，即将量子比特概率幅 (α_i,β_i) 初始化为 $\left(1/\sqrt{2},1/\sqrt{2}\right)$。

3.1.4. 种群观测

对量子态种群 $Q(t)$中量子染色体的各量子比特进行测量，使其坍塌到 0、1 基态，从而获得观测态种群 $P(t)=\left\{p_1^t,p_2^t,\cdots,p_n^t\right\}$。

3.1.5. 种群评价

种群评价使用适应度函数对个体的优劣程度进行评估，个体对应的适应度函数值称之为该个体的适应度，其值越大表示该个体越优秀，即表示越接近最优解。适应度函数需要根据求解的问题进行设计，k-means 中的准则函数 J 是评价聚类中心的目标函数，J 值越小表示聚类质量越好，据此本文基于 k-means

准则函数设计算法的适应度函数，如(1)式所示。

$$f = 1/(1+J) \tag{1}$$

使用(1)式的适应度函数计算 $P(t)$中各个体的适应度，记录最优个体。

3.1.6. 种群进化

种群进化通过使用量子门对染色体的更新来完成，QEA 中通常使用(2)式所示的量子旋转门作为更新算子，按照(3)式所示将量子旋转门作用于染色体的量子比特，使量子比特朝着目标方向偏转，从而实现染色体的更新。

$$U(\theta) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \tag{2}$$

其中，θ为旋转角度。

$$\begin{bmatrix} \alpha_i^{t+1} \\ \beta_i^{t+1} \end{bmatrix} = U(\theta_i) \begin{bmatrix} \alpha_i^{t} \\ \beta_i^{t} \end{bmatrix} = \begin{bmatrix} \cos(\theta_i) & -\sin(\theta_i) \\ \sin(\theta_i) & \cos(\theta_i) \end{bmatrix} \begin{bmatrix} \alpha_i^{t} \\ \beta_i^{t} \end{bmatrix} \tag{3}$$

其中，$\left[\alpha_i^t, \beta_i^t\right]^{\mathrm{T}}$为更新前染色体第 i 个量子比特的概率幅，$\left[\alpha_i^{t+1}, \beta_i^{t+1}\right]^{\mathrm{T}}$为更新后染色体第 i 个量子比特的概率幅。$\theta_i = s(\alpha_i\beta_i)\cdot\Delta\theta_i$，$\Delta\theta_i$、$s(\alpha_i\beta_i)$分别表示旋转角的大小和方向，可以通过查表获得。

3.1.7. 数据聚类

完成染色体更新得到新种群后，以新种群的每个个体为聚类中心，分别对数据集 X 中的数据进行 k-means 划分，生成 n 个新聚类中心，以这 n 个新聚类中心作为算法的新一代种群进入下一次迭代。

3.1.8. 循环停止

算法经过不断循环迭代，种群个体的适应度不断增大，朝着最优解不断逼近。为了在合理的时间内获得满意解，本文设置一个最大进化代数 T，作为 FMGA 的循环停止条件。

3.1.9. 解码输出

当算法迭代到达最大进化代数，即算法停止循环后，将当前最优个体按照编码规则进行解码，得到 k 个簇中心，并以此对数据集 X 中的数据进行 k-means 划分得到 k 个簇。

3.1.10. 标记簇标签

标记簇标签是根据给定的判定规则给 FMGA 输出的 k 个簇中心标上“正常”或“异常”的标签。

在实际网络环境中，网络入侵行为具有 1 个明显的特点，就是入侵网络连接远少于正常网络连接，据此本文通过设定一个阈值 δ，将簇中数据个数大于等于该阈值的簇中心标记为“正常”，将簇中数据个数小于该阈值的簇中心标记为“异常”。

完成标记后，将带标签的 k 个簇中心存入特征模式库中。

3.2. 算法流程

算法：FMGA

输入：网络连接数据集 X，聚类个数 k，种群大小 n，最大进化代数 T，簇标记阈值 δ

输出：带标签的 k 个簇中心

流程：

设置当前进化代数 $t = 0$

设置当前最优个体 p_best=null

由网络连接数据集 X 生成 $Q(t)=\{q_1^t,q_2^t,\cdots,q_n^t\}$

初始化 $Q(t)=\left\{\dfrac{1}{\sqrt{2}}\ \dfrac{1}{\sqrt{2}},\cdots,\dfrac{1}{\sqrt{2}}\right\}$
$\left\{\dfrac{1}{\sqrt{2}}\ \dfrac{1}{\sqrt{2}}\right.\qquad\left.\dfrac{1}{\sqrt{2}}\right\}$

while($t \le T$) {

观测 $Q(t)$获得 $P(t)=\{p_1^t,p_2^t,\cdots,p_n^t\}$

记录当前最优个体 p_best = max$\left(f\left(P(t)\right)\right)$ // $f\left(P(t)\right)$为 $P(t)$中各个体的适应度值

$Q(t)=\text{update}\left(Q(t)\right)$ //使用量子旋转门更新 $Q(t)$中各个体

for i=1 to n {

$q_i^t=\text{k\_means}\left(X,p_i^t\right)$ //以 p_i^t 初始聚类中心对数据集 X 中的数据进行 k-means 划分

$Q(t)=Q(t)\cup q_i^t$

}

t=t+1

}

解码 p_best 获得 $M=\{m_1,m_2,\cdots,m_k\}$ //M 为最优聚类中心

$C=\text{k\_means}(X,M)$ // $C=\{C_1,C_2,\cdots,C_k\}$ 为 k 个簇

for i=1 to k {

if ($|c_i|\ge\delta$) then

$m_i=m_i^+$

else

$m_i=m_i^-$

}

4. 入侵检测模型设计

4.1. 模型构建

1998 年美国计算机科学家 Stuart Stanifor-Chen 等人提出了一个通用的入侵检测框架(Common Intrusion Detection Framework, CIDF)，根据这一框架的基本思想，本文基于 FMGA 算法，提出了一种具有智能性的网络入侵检测模型，模型框架如图 3 所示。

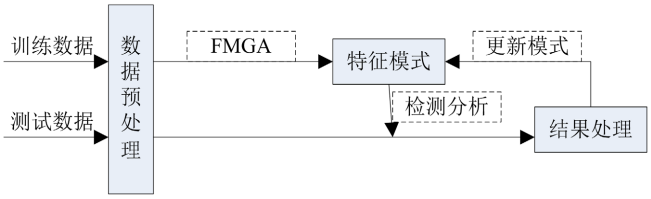

Figure 3. Framework of network intrusion detection model
图 3. 网络入侵检测模型框架

4.2. 模型训练

模型训练是针对事先准备的网络连接数据，即训练数据集，先进行预处理，然后使用 FMGA 算法生

成特征模式。

训练数据预处理包含标准化和特征选择 2 个步骤，标准化是对数据对象的符号特征数值化和数值特征归一化处理，使数据能够适合后续处理；特征选择是从标准化处理后的数据对象的原始特征属性中选出一组能够反映系统入侵状态的重要特征，降低数据维度，从而可以提高计算效率，也就是提高了检测效率，本文实验中使用基于 Fisher Score 的特征评判标准选取有效特征[7]。

训练数据集预处理完成后，作为 FMGA 的输入，运行 FMGA 生成特征模式。

4.3. 入侵检测

入侵检测是针对实际网络环境的连接数据，即测试数据，先进行预处理，然后检测分析。

测试数据预处理包含标准化和特征提取 2 个步骤，标准化是对测试网络连接的符号特征数值化和数值特征归一化处理，使数据能够适合后续处理；特征提取是根据模型训练时的特征选择结果从测试网络连接的原始特征属性中提取出这些特征值参与后续计算。

检测分析是针对从测试网络连接中提取出来特征值，根据特征模式进行分类的过程，分类方法如下：

Step 1 计算距离：计算检测数据与特征模式库中各簇中心的距离；

Step 2 标记标签：将检测数据标记为与其距离最近的簇中心的标签；

Step 3 判断结果：如果标签为正常则直接扔掉该数据，如果标签为异常则进行报警并处理，并将该数据的特征属性反馈至模式更新器。

4.4. 模型的智能性

模型训练后的特征模式库中保存着带标签的簇中心，簇中心将随着检测过程被实时更新，当模式更新器获得异常数据信息时，根据该数据的标签对模式库中对应标签的簇中心进行更新，更新方法是：将该簇中心特征向量与当前异常数据特征向量求平均得到新的簇中心。因此，特征模式库会根据检测过程中的异常信息而不断被调整，使特征模式能够适应当前各种入侵行为，尤其是对于训练时没有出现的新型入侵也能够有效识别，具有智能化检测的能力。

5. 实验验证

5.1. 实验数据

实验使用 KDDCup99 数据集[8]，它是网络入侵检测研究中最经典的数据集，包含训练集(kddcup.data.gz)和测试集(corrected.gz)，本文实验中，从 kddcup.data.gz 的一个 10%子集 kddcup.data-10-percent.gz 中选取训练数据，经去重后获得 60,653 条连接记录，其中正常连接记录占比 90%以上；从 corrected.gz 中选取测试数据 30,333 条连接记录。

5.2. 实验环境

实验在一台 Intel(R) Pentium(R) CPU G3220@4.00 GHZ、8.00 GB 内存、Windows 10 的台式机上进行，使用 Anaconda3 开发工具包完成应用程序和入侵检测。

5.3. 实验参数

实验中各参数设置为：种群大小 $n = 50$、聚类个数 $k = 10$、最大进化代数 $T = 500$、簇标记阈值 $\delta =$ 训练数据总数 * 20%。

5.4. 实验方案与过程

为了验证本文设计的网络入侵检测模型的性能，采用对比实验，对比模型基于传统的 k-means 算法生成特征模式，其他结构与本文模型完全一致。

在 Jupyter Notebook 中完成特征模式生成程序、检测分析程序和簇中心更新程序；对预处理后的训练数据集进行聚类后获得特征模式；对预处理后的测试数据进行入侵检测并实时更新特征模式。分别使用本文模型和对比模型进行 10 次实验，记录各次实验结果。

5.5. 实验结果与分析

选取模型训练时间、入侵检测时间、检测率、误警率、新型入侵识别率 5 个指标来评价模型性能，计算 10 次实验结果的平均值，得到结果如表 1 所示。

训练时间：对 60,653 条连接记录进行聚类生成特征模式库的时间

检测时间：完成 30,333 条连接记录的检测分析时间

检测率：正确识别的测试连接记录/测试连接记录总数

误警率：将入侵测试连接误判为正常连接的记录数/入侵测试连接记录总数

新型入侵识别率：正确识别新型入侵的测试连接记录/新型入侵测试连接记录总数

Table 1. Comparison of experimental results
表 1. 实验结果对比

模型	训练时间(s)	检测时间(s)	检测率(%)	误警率(%)	新型入侵识别率(%)
本文模型	24.91	9.72	87.72	7.63	80.33
对比模型	18.66	10.51	85.12	12.52	72.29

从表 1 数据可以看出，本文模型在检测率、误警率、新型入侵识别率和检测时间 4 个指标都要优于对比模型，而 2 种模型的区别仅仅在于特征模式库不同，因此这一实验结果反映出本文模型的特征模式库质量更好，分析其原因可以得知本文改进了 k-means 后，提升了算法的全局优化能力，使生成特征模式库中的聚类中心更准确，另外在本文模型中 10 次检测过程发现新型入侵识别率一次比一次好，这表明模型随着检测过程能够朝着更优方向调整模式库，体现了模型的智能性。

6. 总结

本文利用量子进化算法显著的优化性能，弥补了 k-means 在复杂结构数据集中聚类结果不够理想的缺陷，保证了模型中特征模式库的质量；利用 k-means 的距离划分特性，使模型的特征模式库能够随检测过程实时更新，提升了模型对新型入侵的识别能力。实验结果表明，本文模型的检测性能比较好，具有智能性，但是生成特征模式库的训练时间比较长，这是因为设计的 FMGZ 算法在迭代过程中都要进行 k-means 划分，时间开销比较大，这是今后研究中需要改进之处。

基金项目

台州职业技术学院 2022 年度大学生科技创新项目(2022DKC18)。

参考文献

[1] 徐辉. 基于 GA-SVM 算法的网络入侵检测研究[J]. 长春工程学院学报(自然科学版), 2021, 22(1): 101-104.

[2] 徐伟, 冷静. 基于人工蜂群算法和 XGBoost 的网络入侵检测方法研究[J]. 计算机应用与软件, 2021, 38(3): 314-318+333.

[3] 顾兆军, 李冰, 刘涛. 基于 ELM-KNN 算法的网络入侵检测模型[J]. 计算机工程与设计, 2018, 39(8): 2412-2416+2421.

[4] Jia, H., Liu, J., Zhang, M., *et al.* (2021) Network Intrusion Detection Based on IE-DBN Model. *Computer Communications*, **178**, 131-140. https://doi.org/10.1016/j.comcom.2021.07.016

[5] Ahmed, M., Seraj, R. and Islam, S.M.S. (2020) The K-Means Algorithm: A Comprehensive Survey and Performance Evaluation. *Electronics*, **9**, 1295. https://doi.org/10.3390/electronics9081295

[6] Meng, Y. and Liu, X. (2018) Quantum Inspired Evolutionary Algorithm for Community Detection in Complex Networks. *Physics Letters A*, **382**, 2305-2312. https://doi.org/10.1016/j.physleta.2018.05.044

[7] 吴迪, 郭嗣琮. 改进的 Fisher Score 特征选择方法及其应用[J]. 辽宁工程技术大学学报(自然科学版), 2019, 38(5): 472-479.

[8] KDD Cup 1999 Data. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html