文章编号：2096-1472(2019)-06-33-05
DOI:10.19644/j.cnki.issn2096-1472.2019.06.009

基于量子人工鱼群和模糊核聚类算法的网络入侵检测模型研究

李 根
(广东工商职业学院计算机应用技术系，广东 肇庆 526020)

摘 要：针对基于传统模糊C均值聚类的网络入侵检测模型存在分类效果不佳，且容易出现局部极值的问题，提出了一种基于量子人工鱼群的半监督模糊核聚类算法。该算法使用少量的标记数据和大量未知标记数据生成网络入侵检的分类，并通过核距离的方式构建了模糊C均值聚类算法的新目标函数，此外，结合了量子人工鱼群算法来解决模糊核聚类算法的全局最优解问题，适用于并行执行架构。在KDD Cup 99网络入侵检测数据上的仿真实验结果表明，相比于基于FCM和PSO−FCM的入侵检测模型，以及基于此提出的算法入侵检测模型具有更好的检测率。

关键词：网络安全；入侵检测；量子人工鱼群；半监督学习；C均值聚类
中图分类号：TP393.8 **文献标识码**：A

Research on Network Intrusion Detection Model Based on Quantum Artificial Fish School and Fuzzy Kernel Clustering Algorithm

LI Gen
(*Department of Computer Application Technology*,*Guangdong College of Business and Technology*,*Zhaoqing* 526020,*China*)

Abstract: Aiming at the problem that the network intrusion detection model based on traditional fuzzy C-means clustering has poor classification effect and the local extremum is easy to occur,the paper proposes a semi-supervised fuzzy kernel clustering algorithm based on quantum artificial fish school optimization.The algorithm uses a small amount of tag data and a large amount of unknown tag data to generate the classification of network intrusion detection,and constructs a new objective function of fuzzy C-means clustering algorithm by means of kernel distance.In addition,it is combined with quantum artificial fish school optimization algorithm to solve the global optimal problem of the fuzzy kernel clustering algorithm,which is applicable to the parallel execution architecture.The simulation results on the KDD Cup 99 network intrusion detection data show that the intrusion detection model based on the proposed algorithm has better detection rate than that based on FCM and PSO-FCM.

Keywords:network security;intrusion detection;quantum artificial fish school;semi-supervised learning;C-means clustering

1 引言(Introduction)

伴随着互联网技术的不断普及，网络支付和网购等的得带了飞速的发展，个人、企业和政府部门对网络的依赖性已经越来越高，网络安全的重要性问题也是日益突出。但是，网络的安全问题其实涉及很多方面，且由于先天性发展问题，安全漏洞在网络协议中无处不在，导致黑客入侵事件和计算机病毒的泛滥[1−3]。如果不能有效遏制这种现象，将会给整个国家乃至整个社会带来巨大的灾难。因此，网络安全问题已成为世界各界十分最关注的问题之一。

现阶段的网络安全防护技术大致可以分为两种[4,5]：被

动式的和主动式的。入侵检测作为一种积极主动的安全防护技术，是传统防火墙的有力补充。早期的入侵检测系统存在的虚警率和漏报率偏高的问题。随着人工智能技术的进步，研究人员提出将机器学习技术应用到入侵检测之中，主要的方式为归纳、分类和数据聚类。例如，朱琨等人[6]对各个机器学习算法在入侵检测应用中的优势和缺点进行了较为详尽的分析比较。目前机器学习主流分为监督学习、无监督学习、半监督和强化学习[7]。计算机网络安全系统中基于无监督学习或者监督学习的入侵检测模型均存在较大的局限性，这是因为基于无监督学习的入侵检测模型通常检测率较差，而基于监督学习的入侵检测模型需要大量的标记数据[8]。因此，广泛采用的是半监督式模型。Haweliya J[9]等人提出一种基于半监督支持向量机的网络入侵检测方法。Vahid S等人[10]提出一种基于K均值聚类和多分类器的网络入侵检测混合学习方法。YY Wee等人[11]采用半监督模糊C-均值聚类来实现网络入侵检测。群智能优化算法和半监督聚类的有效结合，可以提高入侵检测的精确度，是一个新的研究方向。王雪松等人[12]采用改进蚁群算法来优化支持向量机从而提高网络入侵检测的性能。Sreejini K S等人[13]提出了粒子群算法(Particle Swarm Optimization,PSO)和模糊C均值(Fuzzy C-Means，FCM)聚类的PSO-FCM结合方法。

因此，按照群智能优化算法和半监督聚类相结合的思路，本文提出了一种基于量子人工鱼群的半监督模糊核聚类算法。传统的FCM算法采用欧式距离来构建目标函数[14]，导致其分类能力不强。模糊核聚类算法通过选择高斯核函数构造新的特征向量将输入模式空间映射到高维特征空间，提高非线性处理能力。但是，模糊核聚类算法也同样容易陷入局部最优的问题，所以采用量子人工鱼群算法来实现模糊核聚类算法的最优求解，这是因为量子人工鱼群算法比PSO算法具有更强的并行性和分布式优势，且解决了全局优化和收敛速度快。使用少量的标记数据和大量未知标记数据生成网络入侵检的分类。在KDD Cup 99网络入侵检测数据上的仿真实验结果验证提出的算法有效性和可行性。

2 模糊核函数聚类算法(Fuzzy kernel clustering algorithm)

传统的模糊C均值聚类算法的目标函数采用欧式距离作为测度，导致其分类能力不强，无法有效应对非线性样本情况。因此，Wu Yi-quand等人[15]提出把满足Mercer条件的函数核函数引入到聚类分析中，通过非线性映射实现更准确的聚类效果。其原理是利用非线性变换使样本集映射到高维空间F内，则模糊核函数聚类的目标函数J_m^{Φ}为：

$$J_m^{\Phi}=\sum_{i=1}^{c}\sum_{k=1}^{N}u_{ik}^{m}\left\|\Phi(x_k)-\Phi(v_i)\right\|^2 \tag{1}$$

其中，整数c（$2\le c\le n$）为待分类数，集合$X=\{x_1,x_2,...,x_n\}$为样本的集合，n为聚类样本的个数，$V=\{v_1,v_2,...,v_c\}$为样本集合的聚类中心，$\Phi(\cdot)$为非线性变换函数，$\Phi(x_k)$为样本在特征空间F中的映射，$\Phi(v_i)$为聚类中心在特征空间F中的映射，$U=\left[u_{ik}^{m}\right]$为样本集$x_k$的隶属度矩阵，$m\in[1,\infty)$为平滑因子。

对比传统的模糊C均值聚类算法可以看出，模糊核聚类算法使用$\|\Phi(x_k)-\Phi(v_i)\|$替换了原有的欧式距离$\|x_k-v_i\|$。

$$\begin{aligned}&\left\|\Phi(x_k)-\Phi(v_i)\right\|^2\\&=\left(\Phi(x_k)-\Phi(v_i)\right)^T\left(\Phi(x_k)-\Phi(v_i)\right)\\&=\Phi(x_k)^T\Phi(x_k)-\Phi(v_i)^T\Phi(v_i)\\&\quad-\Phi(x_k)^T\Phi(v_i)+\Phi(x_k)^T\Phi(v_i)\\&=K(x_k,x_k)K(v_i,v_i)-2K(x_k,v_i)\end{aligned} \tag{2}$$

以目标函数J_m^{Φ}最小值为聚类准则，令其对u_{ik}和$\Phi(v_i)$的偏导数为零，如下：

$$\frac{\partial J_m^{\Phi}}{\partial u_{ik}}=m\left(u_{ik}^{-m}\left\|\Phi(x_k)-\Phi(v_i)\right\|^2\right)=0 \tag{3}$$

$$\frac{\partial J_m^{\Phi}}{\partial \Phi(v_i)}=-2\sum_{k=1}^{N}u_{ik}^{m}\left(\Phi(x_k)-\Phi(v_i)\right)=0 \tag{4}$$

其中，核函数选用了满足Mercer条件的高斯核函数，即：

$$K(x_i,x_j)=\exp\left(-\frac{\left\|x_i-x_j\right\|^2}{2\sigma^2}\right) \tag{5}$$

其中，σ为尺度参数。因此，公式(2)可以写成：

$$\left\|\Phi(x_k)-\Phi(v_i)\right\|^2=2-2K(x_k,v_i) \tag{6}$$

将公式(6)代入公式(3)和公式(4)中，可以得到：

$$u_{ik}=\frac{(1-K(x_k,v_i))^{-\frac{1}{m-1}}}{\sum_{j=1}^{c}(1-K(x_k,v_i))^{-\frac{1}{m-1}}} \tag{7}$$

$$\Phi(v_i)=\frac{\sum_{k=1}^{N}u_{ik}^{m}\Phi(x_k)}{\sum_{k=1}^{N}u_{ik}^{m}} \tag{8}$$

公式(7)即为隶属度矩阵的表达式，根据核函数$K(x_i,x_j)=\langle\Phi(x_i),\Phi(x_j)\rangle$，结合公式(8)可以推导出：

$$K(x_i,v_i)=\frac{\sum_{k=1}^{N}u_{ik}^{m}K(x_k,x_j)}{\sum_{k=1}^{N}u_{ik}^{m}} \tag{9}$$

其中，$K(x_k,x_j)$可以通过公式(5)计算得到。

如果$\|U_{new}-U_{old}\|<\varepsilon$，则终止迭代，得到要求的聚类中心$v_i$和隶属度矩阵$u_{ik}$，否则转到公式(7)继续更新隶属度矩阵$u_{ik}$。$\varepsilon>0$为迭代截止误差值。

3　量子人工鱼群算法(Quantum artificial fish school algorithm)

人工鱼群算法是一种基于模拟鱼群行为的群体智能算法[16]，而量子人工鱼群算法采用量子比特[17]与人工鱼群相结合的方式对传统人工鱼群算法进行改进，设种群大小为n，其信息素用量子位$P=(p_1,p_2,...,p_n)$表示。基于人工鱼群三大基本行为的算法流程如下：

(1)觅食行为：设该人工鱼此刻的状态为P_i，在其感知范围内随机挑选一个状态P_j。判断是否移动条件的数学表达式为：

$$P_i'=\begin{cases}P_i+Rand\cdot step\cdot\dfrac{P_j-P_i}{d_{ij}} & Y_i<Y_j\\ P_i+Rand\cdot step & Y_i\geq Y_j\end{cases} \tag{10}$$

其中，$Rand$表示一个取值范围为(0,1)的随机数，$step$表示步长，d_{ij}表示人工鱼的当前邻域，Y_i和Y_j分别为两个不同状态下的适应值。

(2)聚群行为：设人工鱼群中人工鱼的d_{ij}内感知到的伙伴数目为n_f且中心位置状态为P_c，数学表达式为：

$$P_i'=\begin{cases}P_i+Rand\cdot step\cdot\dfrac{P_c-P_i}{d_{ic}} & \dfrac{Y_c}{n_f}<\delta Y_j\\ 觅食行为 & \dfrac{Y_c}{n_f}\geq\delta Y_i\end{cases} \tag{11}$$

其中，δ表示拥挤度因子。

(3)追尾行为；设人工鱼当前的d_{ij}内感知到食物浓度最大状态是$P_{\max}$，其数学表达式为：

$$P_i'=\begin{cases}P_i+Rand\cdot step\cdot\dfrac{P_{\max}-P_i}{d_{i\max}} & \dfrac{Y_{\max}}{n_f}<\delta Y_j\\ 觅食行为 & \dfrac{Y_{\max}}{n_f}\geq\delta Y_i\end{cases} \tag{12}$$

(4)随机行为：设人工鱼此刻的状态是P_i，在感知范围$Visual$内随机挑选另外一个状态P_j并朝着此方向前进一步。人工鱼群中个体的量子行为更新方式为[14]。

$$P_j^t=\left[\begin{array}{c|c|c|c}\alpha_1^t & \alpha_2^t & ... & \alpha_m^t\\ \beta_1^t & \beta_2^t & ... & \beta_m^t\end{array}\right] \tag{13}$$

其中，P_j^t代表鱼群中第t次迭代的第j个个体，α_i和β_i是一对复数，表示量子态的概率幅，且满足$|\alpha_i|^2+|\beta_i|^2=1,i=1,2,...,m$。

4　基于量子人工鱼群的半监督模糊核函数聚类算法（Semi-supervised fuzzy kernel clustering algorithm based on quantum artificial fish school）

通过上述模糊核函数聚类算法原理的分析，能够看出如果聚类中心的初始化选择随机分布，容易导致算法陷入局部最优陷阱，而难以获得较好的全局收敛性能。因此本文采用量子人工鱼群来代替来指导其初始聚类中心的选取，从提高分类检测的精确度。

4.1　半监督式模糊聚类

设标记数据和未知标记数据可以表示为矩阵的形式：

$$\boldsymbol{X}=\{\underbrace{x_1^l,x_2^l,...,x_l^l}_{label}\mid\underbrace{x_1^u,x_2^u,...,x_u^u}_{unlabel}\}=\boldsymbol{X}^l\cup\boldsymbol{X}^u \tag{14}$$

模糊聚类差异定义为：

$$J=\sum_{j=1}^{n_u}\sum_{i=1}^{c}\left(u_{ij}^u\right)^m\left\|x_j^u-v_i\right\|_c^2 \tag{15}$$

其中，n_u为未知标记数据样本的数量。

设置量子人工鱼群的适应度函数为：

$$f=(U^u,v)=J+\alpha\cdot J_m^{\Phi} \tag{16}$$

其中，α是调整标记数据和未知标记数据之间平衡的因子。可以看出，f为模糊核函数聚类算法的目标函数和模糊聚类差异的加权求和。

4.2　算法步骤

基于量子人工鱼群的模糊核函数聚类算法的迭代过程为：

(1)设定聚类类别数c、平滑因子m和尺度参数σ，初始化隶属度矩阵u_{ik}，并设定一个任意足够小的迭代截止误差值ε。

(2)量子人工鱼群算法，利用公式(16)计算每个人工鱼的适应度。初始化P_i和P_j，然后利用公式(10)—公式(13)模拟执行人工鱼的群体行为，并更新人工鱼的状态。

(3)P_j'作为初始聚类中心并根据公式(7)更新模糊核函数聚

类的隶属度矩阵u_{ik}，根据公式(8)得到新的聚类中心P_s'。

(4)如果$f(P_s')>f(P_j')$，则用P_s'替换P_j'，否则不做替换。

(5)如果$\|U_{new}-U_{old}\|<\varepsilon$或者达到迭代次数，则终止迭代，根据每个样本得到的隶属度矩阵进行分类，否则转到步骤(3)。

5 实验及结果分析(Experiment and result analysis)

5.1 实验环境

本文实验使用的网络入侵检测数据集为KDDCup99。该数据集是从一个模拟的美国空军局域网上采集来的9个星期的网络连接数据，分成具有标识的训练数据和未加标识的测试数据，其标识类型如表1所示。

表1 KDDCup99入侵检测实验数据的标识类型

Tab.1 Identification types of KDDCup99 intrusion detection experiment data

标识类型	含义	具体分类标识
Normal	正常记录	Normal
DOS	拒绝服务攻击	back、land、neptune、pod、smurf、teardrop
Probing	监视和其他探测活动	ipsweep、nmap、portsweep、satan
R2L	来自远程机器的非法访问	ftp_write、guess_passwd、imap、multihop、phf、spy、warezclient、warezmaster
U2R	普通用户对本地超级用户特权的非法访问	buffer_overflow、loadmodule、perl、rootkit

实验硬件环境为：Windows 7操作系统，Intel(R) Core(TM) i5 CPU、4GB RAM、500G硬盘。实验软件环境为：Eclipse8.5，Cloud Sim3.0.3。基于量子人工鱼群的半监督模糊核函数聚类算法的参数设置如表2所示。

表2 实验参数设值

Tab.2 Experimental parameter setting

参数	数值
Visual	6
步长	0.4
δ	0.8
c	2
m	2
ε	0.001

5.2 数据预处理

考虑到不通过类型入侵数据分别具有不同数量的属性和度量单位，为了不影响聚类分析的结果，所以在数据预处理阶段必须对数据进行归一化处理，具体方式为：

$$new\_s[j]=\frac{\log\left(\frac{ide}{\min\_col+dis}\right)+\log(col+dis)}{\log\left(\frac{ide}{\min\_col+dis}\right)+\log(\max\_col+dis)} \tag{17}$$

其中，*ide*表示数值为2的参数，*dis*表示数值为1.5的参数，min_*col*和max_*col*分别表示[*j*]列属性的最小值和最大值。

5.3 检测结果分析

每次实验重复5次并去检测率的平均值结果，当量子人工鱼群的迭代次数不断增加时，平均检测率的结果如表3所示。可以看出，基于量子人工鱼群的半监督模糊核聚类网络入侵检测模型的平均检测率，会随着迭代次数的增加而缓慢增长，当达到180时趋于稳定，达到84.57%。

表3 检测率与迭代次数的关系表

Tab.3 Relationship between detection rate and number of iterations

项目	数值				
迭代次数	20	60	100	140	180
平均检测率	80.07%	81.12%	82.73%	83.42%	84.57%

将FCM算法[14]、PSO-FCM算法[13]和本文提出算法在KDDCup99数据集中的入侵检测性能进行比较，实验结果对比如表4所示。可以看出，相比于其他两种算法，本文提出的算法在已知攻击和未知攻击方面均具有更高的平均检测率，分别达到了92.47%和83.82%，即表现出更佳的聚类效果。

表4 三种算法的检测率结果比较

Tab.4 Comparison of detection rates among the three algorithms

检测算法	已知攻击	未知攻击
FCM	80.26%	51.07%
PSO-FCM	86.71%	67.25%
本文	92.47%	83.82%

6　结论(Conclusion)

本文提出了一种基于量子人工鱼群的半监督模糊核聚类算法，将群智能优化算法和半监督聚类有效结合到一起。通过选择高斯核函数构造新的目标函数，解决了传统的FCM算法分类能力不强和非线性处理的问题。此外，采用量子人工鱼群算法来实现模糊核聚类算法的最优求解，解决了容易陷入局部最优的问题。网络入侵实验使用少量的标记数据和大量未知标记数据生成半监督分类检测，克服了单一基于监督学习或无监督学习的入侵检测算法的局限。但是对于某些固定的标识类型，如R2L，本文算法仍表现不够理想，存在一定的误检情况；此外量子人工鱼群其他参数对检测性能的影响，后续将针对这两方面进行进一步研究。

参考文献(References)

[1] Chen G,Gong Y,Xiao P,et al.Physical Layer Network Security in the Full-Duplex Relay System[J].IEEE Transactions on Information Forensics & Security,2015,10(3):574-583.

[2] Shin S,Wang H,Gu G.A First Step Toward Network Security Virtualization:From Concept To Prototype[J].IEEE Transactions on Information Forensics & Security,2015,10(10):2236-2249.

[3] Zhou Q,Luo J.The Study on Evaluation Method of Urban Network Security in the Big Data Era[J].Intelligent Automation & Soft Computing,2017(5):1-6.

[4] Liyanage M,Abro A B,Ylianttila M,et al.Opportunities and Challenges of Software-Defined Mobile Networks in Network Security[J].IEEE Security & Privacy,2016,14(4):34-44.

[5] Xiang W,Shi W,Yang X,et al.Efficient Network Security Policy Enforcement With Policy Space Analysis[J].IEEE/ACM Transactions on Networking,2016,24(5):2926-2938.

[6] 朱琨,张琪.机器学习在网络入侵检测中的应用[J].数据采集与处理,2017,32(3):479-488.

[7] Giusti A,Guzzi J,Dan C C,et al.A Machine Learning Approach to Visual Perception of Forest Trails for Mobile Robots[J].IEEE Robotics & Automation Letters,2017,1(2):661-667.

[8] Buczak A L,Guven E.A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection[J].IEEE Communications Surveys & Tutorials,2017,18(2):1153-1176.

[9] Haweliya J,Nigam B.Network Intrusion Detection using Semi Supervised Support Vector Machine[J].International Journal of Computer Applications,2014,85(9):27-31.

[10] Vahid S,Ahmadzadeh M.KCMC:A Hybrid Learning Approach for Network Intrusion Detection using K-means Clustering and Multiple Classifiers[J].International Journal of Computer Applications,2015,124(9):18-23.

[11] Wee Y Y,Cheah W P,Tan S C,et al.A method for root cause analysis with a Bayesian belief network and fuzzy cognitive map[J].Expert Systems with Applications,2015,42(1):468-487.

[12] 王雪松,梁昔明.改进蚁群算法优化支持向量机的网络入侵检测[J].计算技术与自动化,2015(2):95-99.

[13] Sreejini K S,K.Govindan V.Severity Grading of DME from Retina Images:A Combination of PSO and FCM with Bayes Classifier[J].International Journal of Computer Applications,2014,81(16):11-17.

[14] Zhou K,Yang S.Exploring the uniform effect of FCM clustering:A data distribution perspective[J].Knowledge-Based Systems,2016,96(C):76-83.

[15] WU Yi-quan,CAO Peng-xiang,WANG Kai,et al.Meat Image Segmentation Using Fuzzy Local Information C-Means Clustering for Generalized or Mixed Kernel Function[J].Modern Food Science & Technology,2015,31(7):130-136.

[16] Duan Q,Mao M,Pan D,et al.An improved artificial fish swarm algorithm optimized by particle swarm optimization algorithm with extended memory[J].Kybernetes,2016,45(2):210-222.

[17] Fei T,Zhang L,Sun Y,et al.The Location Selection of Distribution Centre Based on DNA Artificial Fish Swarm Algorithm[J].Journal of Computational & Theoretical Nanoscience,2016,13(2):1406-1414.

作者简介：

李　根(1981-)，男，硕士，讲师.研究领域：大数据分析，网络安全.