计算机科学
COMPUTER SCIENCE

基于PRF-RFECV特征优选的GA-LightGBM的网络安全态势评估

任高科，莫秀良

引用本文

任高科，莫秀良. 基于PRF-RFECV特征优选的GA-LightGBM的网络安全态势评估[J]. 计算机科学, 2023, 50(6A): 220400151-6.

REN Gaoke, MO Xiuliang. Network Security Situation Assessment for GA-LightGBM Based on PRF-RFECV Feature Optimization [J]. Computer Science, 2023, 50(6A): 220400151-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220400151

基于 PRF-RFECV 特征优选的 GA-LightGBM 的网络安全态势评估

任高科[1]　莫秀良[2]
1 天津理工大学计算机科学与工程学院　天津 300384
2 天津市智能计算及软件新技术重点实验室　天津 300384
(935208706@qq.com)

摘　要　目前,在网络安全领域中,传统机器学习模型存在训练时间过长和对冗余特征高敏感性的缺点,已然处理不了日益复杂的网络空间。为针对海量、高维的网络安全要素,提高网络安全态势评估的精度和效率,提出了一种基于 PRF-RFECV 特征优选的 GA-LightGBM 的网络安全态势评估模型。首先利用并行随机森林筛选出的特征重要度,然后结合带有交叉验证的递归特征消除选出最优特征集,最后利用遗传算法的全局搜索特性选取轻度级梯度提升机模型的最优参数后进行分类。实验仿真表明,该模型在准确率和 F1 分数上均优于传统的网络安全态势评估算法,且效率更高。

关键词:网络安全态势;轻量级梯度提升机;随机森林;遗传算法
中图法分类号　TP393

Network Security Situation Assessment for GA-LightGBM Based on PRF-RFECV Feature Optimization

REN Gaoke[1] and MO Xiuliang[2]
1 School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China
2 Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology, Tianjin University of Technology, Tianjin 300384, China

Abstract　At present, in the field of cyber security, due to the shortcomings of long training time and high sensitivity to redundant features, traditional machine learning models have been unable to deal with the increasingly complex network space. To improve the accuracy and efficiency of network security situation awareness for massive and high-dimensional network security elements, a GA-LightGBM network security situation awareness model based on PRF-RFECV feature preference is proposed, which first uses parallel random forest to filter out feature importance, then combines recursive feature elimination with cross-validation to select the optimal feature set, and finally uses the global search property of genetic algorithm to select the optimal parameters of LightGBM model for classification. Experimental simulation shows that the model is more accurate and more efficient than the traditional network security situation awareness algorithm in terms of both accuracy and F1 score.

Keywords　Network security situation, Light gradient boosting machine, Random forest, Genetic algorithm

1　引言

随着以网络为中心的作战方式的作战能力不断增强,以及对云计算、大数据、物联网等技术的深入研究,当今社会对网络空间安全的定义与理解提出了更高的要求,网络安全问题也受到了更为广泛的关注。如何及时准确地掌握网络系统的整体安全信息,加固脆弱易受攻击环节是影响未来互联网发展的重要指标。

网络安全态势评估又可被称为网络安全态势理解,在网络安全态势感知的整个过程种占有重要地位[1]。其本质就是通过收集网络系统中的安全要素,在已构建的安全指标体系基础上建立合适的模型,摒弃研究单一的安全事件,而对整个网络系统已遭受和未来将遭受的安全威胁进行综合评估,动态反应网络实际的运行状态,获取宏观的网络安全态势,评估网络系统的安全等级,以协助管理者及时加固预防,达到复杂决策的目的。

近年来,关于网络安全态势评估的研究较多,Dong 等[2]针对现有网络安全态势评估方法效率低、可靠性差的问题,提出了一种基于改进 BP 神经网络的定量评估方法,该方法结合 Cuckoo 搜索算法提高了收敛速度和评估精度,但是存在评估稳定性不高的问题。为了提高网络安全态势评估性能,He[3]提出一种 KNN 和支持向量机两种算法相融合的网络安全态势评估模型,以解决支持向量机对超平面附近样本易错分的问题,减少 SVM 的误判率,提高了网络安全态势的评估

基金项目:国家基金面上-联合基金(U1536122);科技部"科技助力经济 2020"重点专项(SQ2020YFF0413781);天津市科委重大专项(15ZXDSGX00030)
This work was supported by the National Funds-Joint Fund Projects(U1536122), Key Special Project of "Science and Technology Helps Economy 2020" of the Ministry of Science and Technology(SQ2020YFF0413781) and Major Project of Tianjin Science and Technology Commission (15ZXDSGX00030).
通信作者:莫秀良(moxiuliang@163.com)

正确率，但对海量多维数据计算时间过长，牺牲了评估的实时性。Zhao 等[4]为弥补网络信息复杂的情况下网络安全态势评估准确率不高等缺陷，提出一种基于时间因子和复合 CNN 结构的网络安全态势感知模型，利用两种结构的不同优势，形成串联复合单元结构，大大减少了网络参数，缩短了网络运行时间。

综上所述，目前机器学习已经成为网络安全态势评估领域中的主流，但随着网络用户的不断增加，网络拓扑结构的日益复杂，网络安全要素呈爆炸式增长，从而导致模型训练速度和评估精度的大幅度下降，因此传统的机器学习模型已经不能满足当前网络安全所要求的高实时性和高准确性。针对以上问题，本文提出一种基于并行随机森林(Parallel Random Forest，PRF)和带有交叉验证的递归式特征消除算法(Recursive Feature Elimimation with Cross-Vaildation，RFECV)对特征进行优选、使用遗传算法(Genetic Algorithm，GA)进行参数优化过后的轻量级梯度提升机(Light Gradient Boosting Machine，LightGBM)模型进行评估的网络安全态势评估框架(PRF-RFECV-GA-LightGBM)，该框架不仅能够有效地避免过拟合，而且具有预测精度高、训练速度快以及泛化能力好的特点，可以广泛应用在网络安全态势评估问题中。

2 网络安全态势评估

2.1 数据融合

数据融合是种多层次、全方位的数据处理方法，贯穿了整个网络安全态势体系，其主要目的是将来自多种数据源的安全要素进行综合处理，使得通过数据分析所得出的结果更为精确和可靠。根据信息抽象程度可将数据融合分为三大级别，从低到高依次为数据级融合、特征级融合和决策级融合，其中决策级数据融合具有最优的实时性和容错性。

KDDCup99 数据集是从模拟的美国空军局域网上采集而来的九周网络连接数据。它共包含 41 个特征属性和 1 个标签属性，9 个特征属性为离散型，其他为连续型。标签属性包含了 1 种正常的标识类型 normal 和 22 种网络常见的攻击类型，特征属性包含离散型和连续型两种数值类型。由于连续型特征属性的各种属性的度量方法不同，为避免对评估结果产生影响，并使数据达到决策级融合，因此需要提高训练效率以增加实时性[5]。本文不但针对连续型特征属性进行标准化和归一化处理，而且将 22 种攻击类型的一级态势提取到 4 种二级态势，如表 1 所列。

表 1 决策级数据融合

Table 1 Decision-level data fusion

二级态势	一级态势
DOS	back，land，neptune，pod，smurf，teardrop
R2L	ftp_write，guess_passwd，imap，multihop，phf，spy，warezclient，warezmaster
U2R	buffer_overflow，loadmodule，perl，rootkit
Probing	ipsweep，nmap，portsweep，satan

其中，DOS 为拒绝服务攻击，R2L 表示来自远程机器的非法访问，U2R 表示普通用户对本地超级用户特权的非法访问，Probing 为端口监视和其他探测活动。

2.2 指标构建

整个网络安全态势的获取需要从海量的网络数据中将网络威胁和攻击方式按照网络安全标准进行分析，从网络流量中提取网络攻击类型。本文结合通用漏洞评分系统(Common Vulnerability Scoring System，CVSS)对 KDDCup99 数据集中攻击类型进行评分[6]，按照 Probing 寻找目标过程、R2L 从远程到本地的入侵过程、U2R 本地提权和 DOS 资源消耗四大攻击类型对整个网络系统的危害程度进行攻击影响度量化，并且按照攻击类型对网络安全的影响程度由低到高排序，如表 2 所列。

表 2 网络攻击危害分级

Table 2 Network attack hazard classification

攻击类型	攻击影响度
Probing	0.3
R2L	0.4
U2R	0.5
DOS	0.7

根据表 2 对网络系统中所有主机受到的攻击类型和攻击个数分别计算网络安全态势值。设网络中包含 N 台主机，攻击类型为 $i(1\leqslant i\leqslant 4)$种，每种攻击类型对应的影响因子为 s_i，设第 j 台主机受到的第 i 种攻击类型的数量为 m_i 个，那么第 j 台主机的态势值为：

$$M_j=\frac{1}{m_{ij}}\sum_{i=1}^{4}s_i m_{ij} \tag{1}$$

分别计算 N 台主机各自的态势值，加权求和得到整个网络系统的安全态势值为：

$$M=\sum_{j=1}^{N}q_j M_j \tag{2}$$

其中，q 代表第 j 台主机在整个网络系统中的重要程度，也就是权重。

通过以上公式计算出的整个网络系统的安全态势值可参考《国家突发公告事件总体应急预案》归一化到指定区间，以此来将网络系统态势安全划为 5 个等级[7]，具体等级划分如表 3 所列。

表 3 网络安全的态势等级分类

Table 3 Situation level classification of network security

网络安全态势等级	网络安全态势值区间
安全	[0.00，0.30]
低风险	(0.30，0.60]
中等风险	(0.60，0.90]
高风险	(0.90，1.20]
超高风险	(1.20，1.50]

2.3 模型框架

本文提出的基于 PRF-RFECV 特征优选的 GA-LightGBM 网络安全态势评估模型由态势提取模块(Situation Extraction Module，SEXM)，态势评估模块(Situation Assessment Module，SASM)，态势可视化模块(Situation Visualization Module，SVIM)三大模块构成，如图 1 所示。

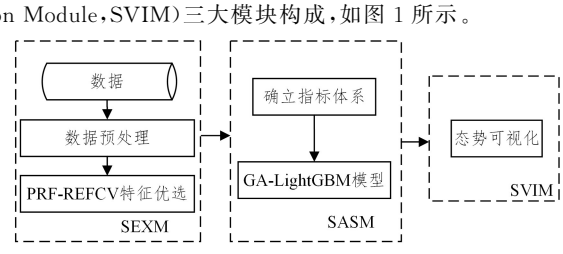


图 1 PRF-RFECV-GA-LightGBM 模型框架

Fig. 1 PRF-RFECV-GA-LightGBM model framework

数据处理模块：将数据集进行归一化、数据化等预处理

后，进行最深层次数据融合决策级融合，将一级态势构建融合为具有代表性的二级态势 DOS，R2L，U2L 和 Probing，减小了数据规模，使分类模型具备最好的实时性，然后利用并行随机森林构建多棵决策树，挖掘特征间的潜在关系，评估其重要度，最后带有交叉验证的递归式特征消除自动选出最优特征子集，排除对评估结果影响较小的特征和无关的特征，避免由于特征冗余而造成计算量爆炸和分类准确率低下等问题。

态势评估模块：选取权威性指标体系，去除对整个网络系统影响较小的因素，并将定量指标进行标准化，避免多指标综合处理的影响。由于 LightGBM 的参数种类复杂，意义多样，为避免依赖于主观经验进行调参而陷入局部最优的调参范围，利用基于进化理论和种群遗传理论的遗传算法来优化 LightGBM 的重要参数，其能够利用全局范围内的空间信息，自适应地达到逼近最优值的状态。

态势可视化模块：综合表 2 和表 3 的指标信息，可以反映宏观网络安全态势值，配合管理人员能够及时预警安全风险，减少资产损失。

3 轻量级梯度提升机

3.1 LightGBM 算法

LightGBM 算法是由 Ke 等在 2017 年基于 Histogram 算法提出的一种分布式梯度提升决策树（Gradient Boosting Decision Tree，GBDT）算法[8]。Histogram 算法的基本思想是将连续的浮点特征值离散化成 k 个整数，同时构造一个宽度为 k 的直方图，遍历时将离散化后的值作为索引在直方图中累计统计量，然后根据离散值遍历寻找最优分割点。LightGBM 算法通过使用 Histogram 算法思想，在存储数据空间、内存消耗和计算时间上都有大幅度减少。由于 GBDT 每一次迭代都需要遍历整个数据集多次，在面对海量级数据时，具有该性质的普通 GBDT 算法如 XGBoost 不能满足用户的需求。LightGBM 摒弃了大多 GBDT 算法使用的 Level-wise 的决策树生长策略，而使用带有深度限制的 Leaf-wise 算法，与 Level-wise 相比，在分裂次数相同的情况下，Leaf-wise 具有更好的精度，同时在最大深度的限制下，保证了高效率性和抗拟合性。

假设给定训练数据集 $D=\{(x_1,y_1),\cdots,(x_N,y_N)\}$，提升树模型可以表示以决策树为基本学习器的加法模型[9]。

$$F(x)=\sum_{t=1}^{T}\alpha_t f_t(x) \tag{3}$$

其中，$f_t(x)$为回归树，T 是梯度提升决策树中需要构建的树的数量，α 是对应第 t 棵树的权重。

而 LightGBM 算法的目标是极小化一个目标函数。

$$\begin{aligned}\mathcal{L}^{(t)}&=\arg\min\sum_{i=1}^{N}L(y_i,\hat{y}_i^{(t)})+\Omega(f_t)\\&=\arg\min\sum_{i=1}^{N}L(y_i,\hat{y}_i^{(t-1)}+f_t(x_i))+\Omega(f_t)\end{aligned} \tag{4}$$

其中，L 代表在建立第 t 棵树时根据真实值和预测值所得到的损失值，Ω 为其对应的正则化项。根据 LightGBM 选择如下的正则项：

$$\Omega(f_t)=\gamma^{\mathrm{T}}+\frac{1}{2}\lambda\parallel\boldsymbol{w}\parallel^2 \tag{5}$$

其中，$\boldsymbol{w}$ 是叶子节点预测值组成的向量，$\gamma,\lambda\geqslant 0$ 为超参数，用于控制比重。

在模型学习过程中使用如下的正则化策略来缓和过拟合：

$$F_t(x)=F_{t-1}(x)+\nu f_t(x) \tag{6}$$

其中，$0<\nu\leqslant 1$ 为学习率，用以更新模型，学习率越小，则需要生成更多的回归树，精度会更高，但同时也会增加训练时间，因此可以控制 γ,λ,ν 等参数来构建一个速度快且精度高的 LightGBM 模型。

为更好地解决海量数据，LightGBM 算法还采用单边梯度采样（Gradient-based One-Side Sampling，GOSS）和互斥特征捆绑（Exclusive Feature Bundling，EFB）算法，极大地减少了计算量，采用带有深度限制的按叶子生长算法，在避免一些无用的计算开销的同时还可以防止过拟合，并且通过优化后的特征并行、数据并行方法进行加速计算，提升了评估效率和精度，使该模型能够很好地适用于当前海量网络安全要素导致的评估效率低下的网络安全态势评估中。

3.2 特征重要性评估

由于当今网络流量中特征维数过多，但部分特征对网络安全态势评估的结果影响较小，甚至可能会是噪音对实验准确率造成干扰，且高维特征更加容易导致模型过拟合，大量的数据更会增加模型的复杂度，使得最终评估模型的准确率降低，所以对数据集中的原始特征进行特征优选是有必要的。随机森林算法可以根据基尼指数计算出特征的重要度，挖掘特征间的相互关系，优化下游算法性能[10]。

设样本总数 N，有 B 个特征 $X_1,X_2,X_3,\cdots,X_B$，基于基尼指数来评估特征的重要性。基尼指数的计算公式如下：

$$GI_m(p)=\sum_{k=1}^{K}p_{mk}(1-p_{mk})=1-\sum_{k=1}^{K}p_{mk}^2 \tag{7}$$

其中，k 代表共有 K 个类别，p_{mk} 代表节点 m 中类别 k 的样本权重。

特征 X_j 在节点 m 的重要性，即节点 m 分枝前后的基尼指数变化量为：

$$VIM_{jm}^{\mathrm{gini}}=GI_m-GI_l-GI_r \tag{8}$$

其中，GI_l 和 GI_r 分别代表分支后左右两个新节点的基尼指数。

如果特征 X_j 在决策树 i 中出现的节点在集合 M 中，那么特征 X_j 在第 i 棵决策树的重要性为：

$$VIM_{ij}^{\mathrm{gini}}=\sum_{m\in M}VIM_{jm}^{\mathrm{gini}} \tag{9}$$

假设随机森林一共有 n 棵决策树，那么特征 X_j 的特征重要性为：

$$VIM_j^{\mathrm{gini}}=\sum_{i=1}^{n}VIM_{ij}^{\mathrm{gini}} \tag{10}$$

最后将得到的所有重要性评分进行归一化处理。

$$VIM_j=\frac{VIM_j}{\sum_{i=1}^{B}VIM_i} \tag{11}$$

特征重要性是特征在所有决策树的重要性评分的累加和，理论上随机森林只要构造足够多的决策树就能具有较强的抗拟合能力和较高的准确度，但构建多树也会造成效率低下。并行随机森林[11]可以利用随机森林中每棵决策树单独计算等特性，构建并行随机森林，使多棵决策树并行计算，以此避免构建多棵决策树导致的训练效率低下的问题。其原理是由“管理员”（master）将完整的数据分成若干个数据块（chunk）及相应的执行程序发送给每一个进程，并行计算完成后再由 master 汇总每一份结果，并将结果整理成实际需要的相应的形式，如图 2 所示。

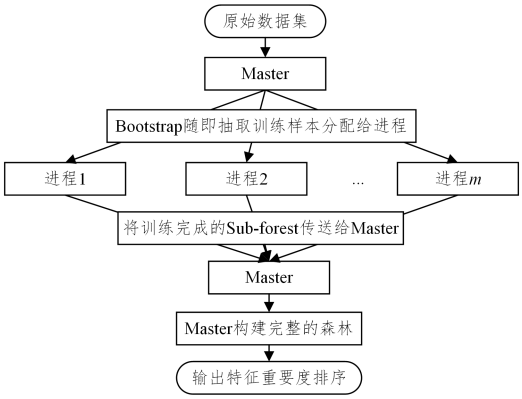


图 2 并行随机森林工作流程示例

Fig. 2 Example of parallel random forest workflow

3.3 特征子集优选

交叉验证的递归式特征消除(Recursive Feature Elimination with Cross-Vaildation,RFECV)常用于特征选择问题中。递归消除特征方法(Recursive Feature Elimination,RFE)可以描述为对基类模型进行迭代训练,每轮迭代都记录损失值,并且消除特征重要性最小的特征,再基于新的特征集进行下一轮训练,直到特征集仅为一个,最终的特征等级即为消除顺序。

RFECV 是在 RFE 的基础上对不同的特征组合进行交叉验证并行通过基类模型来计算所有子集的验证误差,选择误差率最小的子集作为最优特征子集[12]。如算法 1 所示:首先对预处理后的数据集采用并行随机森林算法构造多棵决策树,根据基尼指数得到各特征的重要性,并且从大到小进行排序;然后逐次删除一个或多个特征重要性最小的特征直到特征子集数为一个,依次记录每次迭代后的特征子集,并将其做为新特征子集;最后将所有特征子集基于基模型进行 10 折交叉验证,得到评分最高的特征子集为最优特征集。

算法 1 PRF-RFECV

```
Input:D,C,step
Output:F
1. if |C|>1 do:
2.    Feature_list=PRF(D)
3.    sub=Feature_list-Feature[-step:]
4.    C_list.appent(sub)
5. end if
6. for each sub in C_list:
7.    rfc=RandomForesClassifier()
8.    scores=cross_val_score(rfc,D,sub,CV=10)
9. end for
10. F=arg max scores
```

3.4 遗传优化

遗传算法是一种借鉴生物界自然选择和自然遗传机制的算法,具有高度并行、随机、自适应搜索等特性。在态势评估模块中,为提高模型的精准度和泛化能力,避免过拟合,利用遗传算法优化 LightGBM 的有关参数。

首先根据式(5)、式(6)对问题的潜在解进行浮点数编码,映射关系为:

$$\boldsymbol{X}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}\Rightarrow V=(v_1,v_2,\cdots,v_m)^{\mathrm{T}} \tag{12}$$

其中,$\boldsymbol{X}$ 为表现型代表决策向量,V 为基因型代表染色体,v_i 代表染色体内的一个基因,m 值等于决策变量的个数 n,设置最大进化代数,初始化第一代种群记为 $P(0)$。

选取平均绝对百分比误差(MAPE)作为适应度函数来衡量预测结果的好坏,计算公式如下:

$$M(y)=\frac{1}{n}\sum_{i=1}^{n}\frac{|y_i-\hat{y}_i|}{y_i} \tag{13}$$

其中,y_i 为真实值,$\hat{y}_i$ 为预测值,为满足非负数、连续和最大化等要求,适应度函数最终可转换为:

$$F(y)=\frac{1}{1+M(y)} \tag{14}$$

为避免陷入局部最优点,选择通用性强、效率高的锦标赛选择算法[13],根据适应度重复将最好的个体选择进入子代群体,得到的个体构成新一代种群。

选取模拟二进制交叉作为交叉算子,边界变异为变异算子。为避免概率过小的问题,导致遗传过程过早收敛,设交叉和变异概率分别为 P_c 和 P_m,限制范围为$[P_{c\min},P_{c\max}]$和$[P_{m\min},P_{m\max}]$,其中 $P_{c\min}=0$,$P_{c\max}=0.9$,$P_{m\min}=0.01$,$P_{m\max}=0.1$。设种群全部个体适应度均值为 f_{avg},交叉和变异的适应度分别为 f'和 f,得到[14]:

$$P_c=\begin{cases}P_{c\max}-\dfrac{(P_{c\max}-P_{c\min})(f_{\max}-f')}{f_{\max}-f_{\mathrm{avg}}}, & f'\geqslant f_{\mathrm{avg}}\\ P_{c\max}, & f'<f_{\mathrm{avg}}\end{cases} \tag{15}$$

$$P_m=\begin{cases}P_{m\max}-\dfrac{(P_{m\max}-P_{m\min})(f_{\max}-f')}{f_{\max}-f_{\mathrm{avg}}}, & f\geqslant f_{\mathrm{avg}}\\ P_{m\max}, & f<f_{\mathrm{avg}}\end{cases} \tag{16}$$

遗传过程不断的进化迭代,直到达到所要求的评估精度或者最大进化代数时算法停止,从而得到 LightGBM 模型最佳的学习率、最大树深和基树棵数参数。

4 实验仿真

4.1 评估指标

网络安全态势感知中的安全要素存在数据分布不均匀的问题,因此,为正确反映所构模型的真实效果,应该选取合适的评估指标。混淆矩阵常被用来评估有关网络安全模型的有效指标,如表 4 所列。

表 4 混淆矩阵

Table 4 Confusion matrix

真实类别	预测类别	
	positive	negative
positive	true positive	false negative
negative	false positive	true negative

其中,真正类(True Positive,TP)是将样本中的正类预测为正类;真负类(True Negative,TN)是将样本中的负类预测为负类;假正类(False Positive,FP)是将样本中的负类预测为正类;假负类(False Negative,FN)是将样本中的正类预测为负类。

为准确有效地评价模型在真实环境中的评估效果,引入以下指标来评估模型的有效性。精确率(Precision)表示正确检测出的负类样本与真正的负类样本的比例,计算公式为:

$$Precision=\frac{\sum_{i=1}^{N}TP_i}{\sum_{i=1}^{N}(TP_i+FP)} \tag{17}$$

召回率(Recall)表示正确检测出的负类样本占所有负类样本的比例,计算公式为:

$$Recall=\frac{\sum_{i=1}^{N}TP_i}{\sum_{i=1}^{N}(TP_i+FN_i)} \tag{18}$$

准确率(Accuracy)表示正确检测出的负类样本和正类样本占总数的比例，计算公式为：

$$Accuracy=\frac{\sum_{i=1}^{N}(TP_i+TN_i)}{\sum_{i=1}^{N}(TP_i+TN_i+FP_i+FN_i)} \tag{19}$$

F1 分数是精确率和召回率的调和平均数，常用来综合评价分类模型的精确度，计算公式为：

$$F1\text{-}score=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{20}$$

其中，N 为类别的数量。由于样本分布不均，正如真实网络中的流量数据，正常样本多，异常样本很少，为避免受到常见类别的影响，使各个类别地位相同，以上指标均采用宏平均[15]来评估模型。

4.2 实验环境和数据

为准确测试 PRF-RFECV-GA-LightGBM 算法的精度和效率，实验环境设为 Ubuntu 操作系统，python3.8 开发语言，Anaconda3 工具箱。详细硬件参数为：Intel(R) Core(TM) i7-10750H 处理器，显卡为 RTX2060，利用 CPU 核数为 4 核，64 GB 主存，16 GB 内存。

数据集 KDDCup99，是从一个模拟的美国空军局域网上采集而来的 9 个星期的网络连接数据。该数据集分为具有标识的训练数据和未加标识的测试数据，测试数据和训练数据有着不同的概率分布，测试数据集中还包含一些未在训练数据集中出现的攻击类型，更加体现了其现实性，因此被广泛用于网络威胁检测模型实验中。实验仿真选取的 KDDCup99 数据样本共有 494 010 条，其中正常流量样本数量为 97 276，攻击流量样本数量为 396 734，训练样本和测试样本数量按照 3∶1 的比例分配。详细数据分布如表 5 所列。

表 5 训练集和测试集的数据分布

Table 5 Data distribution of training set and test set

攻击类型	训练样本	测试样本	训练集占比/%	测试集占比/%
Normal	97 278	60 593	19.69	19.48
DOS	391 458	229 853	79.24	73.90
U2R	52	228	0.01	0.07
R2L	1 126	16 189	0.23	5.20
Probing	4 107	4 166	0.83	1.34

4.3 特征选择

为排除无关特征、避免过拟合和提高评估准确度，在对数据集进行归一化、数据化后使用并行随机森林和交叉验证的递归式特征消除方法来进行特征重要度评估并构造最优特征子集，使用 LightGBM 模型来验证特征优选前后评估的准确率和效率。通过 KDDCup99 训练集，基于 PRF-RFECV 特征优选的 LightGBM 模型的具体训练结果如表 6 所列。

表 6 不同的优选特征数的实验比较

Table 6 Experimental comparison of different preferred feature numbers

Number	*Accuracy*	*Precision*	*Recall*	*F1*	Time/s
27	0.99979	0.96981	0.97246	0.97113	12
35	0.98668	0.93519	0.96458	0.94966	14
41	0.99694	0.96418	0.96421	0.96419	16

由表 6 可以明显看出，只对数据进行归一化而不进行特征优选时，即总共 41 个特征时，通过 LightGBM 模型评估的准确率只有 0.99694，训练及评估时间达到 16 s。在按照特征重要度选择特征个数超过 27 个达到 35 个时，各个评估指标下跌。根据实验显示，通过 PRF-RFECV 算法进行自动特征优选后对得到的最优特征子集(27 个特征)进行训练，不但评估准确率从的0.99694提升到 0.99979，且每个评估指标均大于全特征评估，说明该算法是有效可行的，并且使用并行计算方法极大地利用了计算机已有的计算资源，充分提高了模型的训练速度。

4.4 结果对比

4.4.1 算法优化的评估性能对比

为具体分析引入 PRF-RFECV 和 GA 后对 LightGBM 模型评估性能的影响，排除偶然性影响，分别对从周一到周日的 7 组数据集训练样本进行 LightGBM 和 PRF-RFECV-GA-LightGBM 模型网络安全态势评估训练，同时根据 7 组测试数据集评估结果与实际评估结果的对比给出稳定的训练模型及其相关参数，而后分别输入 7 组测试集的准确率和 F1-score 值，并给出一周内二级态势趋势图，这有助于网络管理员进行及时加固。具体训练结果如表 7 所列。

表 7 LightGBM 和 GA-LightGBM 的仿真实验对比

Table 7 Comparison of simulation experiments between LightGBM and GA-LightGBM

日期	LightGBM		GA-LightGBM	
	准确率	*F1-score*	准确率	*F1-score*
周一	0.9965	0.7622	0.9992	0.9689
周二	0.9962	0.8063	0.9993	0.9783
周三	0.9942	0.8638	0.9993	0.9863
周四	0.9989	0.7858	0.9999	0.9821
周五	0.9976	0.7827	0.9996	0.9532
周六	0.9899	0.7931	0.9997	0.9666
周日	0.9993	0.7889	0.9996	0.9566

由表 6 可知，在以上 7 组实验数据中，相比 LightGBM 算法，GA-LightGBM 算法的网络安全态势评估准确率和 $F1$ 都更高，GA-LightGBM 对每组的评估准确率都达到了 0.999 以上，$F1$ 达到了 0.95 以上。由表 6 中的 LightGBM 的 $F1$ 值可以看出，由于测试集中异常样本极少，训练出的 LightGBM 模型对异常样本不敏感，导致每组的 $F1$ 都低于 0.9，而 GA-LightGBM 对所有数据集的 $F1$ 都大于 95，表明了 GA-LightGBM 模型的查准率和查全率更高，模型质量更好。

下面根据态势值指标构建态势趋势图，将一周内 GA-LightGBM 计算的网络安全态势值与实际态势值进行对比，可视化结果如图 3 所示。

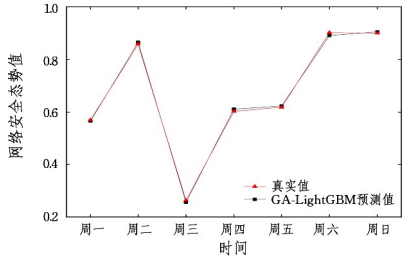


图 3 网络安全态势评估值对比

Fig. 3 Comparison of network security situation assessment values

由图 3 可看出，在一周内 GA-LightGBM 的网络安全态势评估值与真实值的拟合程度较好，只有在周六时有极小的偏差。当网络安全态势值超过 0.6 时，需要开启网络安全风险预警，超过 0.9 时就需要开启网络安全管控措施。网络安全态势值可视化有助于网络安全管理人员掌握整个网络的安全风险情况，即时发布预警管控信息和加固安全系统。

4.4.2 不同算法的评估性能对比

为进一步验证 PRF-RFECV-GA-LightGBM 模型在网络安全态势评估中的性能表现，使用 KNN[16]、SVM[17]、贝叶斯网络(BN)[18]和 BP 神经网络[19]分别在 KDDCup99 数据集中进行对比实验。由表 8 可以看出，相比其他算法，PRF-RFECV-GA-LightGBM 作为多分类模型，对每个类别上的评估准确率都有着很好的表现，即使在仅占总样本 0.07%的“U2R”上评估准确率也达到了 91.56%，远远超过 KNN，SVM 和 BP 算法。

表 8 5 种算法在每个类别的评估准确率对比

Table 8 Comparison of evaluation accuracy of five algorithms in each category

(单位：%)

类别	比例	准确率				
		KNN	SVM	BN	BP	GA-LightGBM
Normal	19.48	85.69	95.45	96.34	99.51	99.95
DOS	73.90	86.72	96.88	94.67	99.85	99.99
U2R	0.07	80.86	9.65	93.21	60.38	91.56
R2L	5.20	79.83	6.16	92.17	80.49	99.75
Probing	1.34	83.23	78.54	94.62	90.63	99.01

另外，由于 LightGBM 算法采用 Histogram，Goss 和 EFB 等优化算法，极大地减少了计算量，在已知实验环境中使用 LightGBM 对 KDDCup99 数据集中的 494010 条数据进行训练只需 16 s，使用 PRF-RFECV-GA-LightGBM 模型训练则需要长达 300 s，并且可以使用并行和分布式计算，而 SVM 和 KNN 算法在 10 万级数据量的情况下训练都已经极为困难。因此，最终实验结果表明，基于 RF-RFECV 特征优选的 GA-LightGBM 模型优于其他评估模型，更适合海量、高维数据的网络安全态势评估场景。

结束语 本文采用 PRF-RFECV 算法选出最优特征集，并使用 GA 算法优化 LightGBM 的重要参数。经过实验证明，通过 PRF-RFECV 算法对特征进行优选去除冗余特征，能够提高评估准确率和效率，并且使用 GA 算法优化 LightGBM 重要参数，提高了模型评估的准确性和稳定性。相比常见的网络态势评估算法，本文提出的 PRF-RFECV-GA-LightGBM 模型在评估准确性和评估性能上都具有很好的表现，但是由于引入 GA 算法导致模型评估性能被大幅度拉低，因此下一步将优化 GA 算法或采用性能更好的寻参算法来提升网络安全态势感知的整体评估性能，以便能够实时动态地监测网络安全的现实情况。

参 考 文 献

[1] GONG J, ZHANG X, SHU Q, et al. Survey of network security situation awareness[J]. Journal of Software, 2017, 28(4): 1010-1026.

[2] DONG G, LI W, WANG S, et al. The assessment method of network security situation based on improved BP neural network [C]// International Conference on Computer Engineering and Networks. Cham: Springer, 2018: 67-76.

[3] HE Y. Assessment model of network security situation based on K nearest neighbor and support vector machine[J]. Computer Engineering and Applications, 2013, 49(9): 81-84.

[4] ZHAO D M, SONG H Q, ZHANG H B. Network security situation based on time factor and composite CNN structure[J]. Computer Science, 2021, 48(12): 349-356.

[5] TAVALLAEE M, BAGHERI E, LU W, et al. A de-tailed analysis of the KDD CUP 99 data set[C]// 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications. IEEE, 2009: 1-6.

[6] MELL P, SCARFONE K, ROMANOSKYS. Common vulnerability scoring system[J]. IEEE Security & Privacy, 2006, 4(6): 85-89.

[7] State Council. National master plan for resp-onding to public emergencies [M]. Beijing: China Legal Publishing House, 2006: 4-6.

[8] KE G, MENG Q, FINLEY T, et al. Lightgbm: Ahighly efficient gradient boosting decision tree[J]. Advances in Neural Information Processing Systems, 2017, 30: 6-8

[9] ZHOU J Y, HE P F, QIU R F, et al. Research on intrusion detection based on random Forest and gradient[J]. Journal of Software, 2021, 32(10): 3254-3265.

[10] SCHONLAU M, ZOU R Y. The random forest algorithm for statistical learning[J]. The Stata Journal, 2020, 20(1): 3-29.

[11] NAGHIBI S A, HASHEMI H, BERNDTSSONR, et al. Application of extreme gradient boosting and parallel random forest algorithms for assessing groundwater spring potential using DEM-derived factors [J]. Journal of Hydrology, 2020, 589: 125197.

[12] SHANG Q, FENG L, GAO S. A hybrid method for traffic incident detection using random forest-recursive feature elimination and long short-term memory network with bayesian optimization algorithm[J]. IEEE Access, 2020, 9: 1219-1232.

[13] KILIÇ H, YÜZGEÇ U. Tournament selection based antlion optimization algorithm for solving quadratic assignment problem [J]. Engineering Science and Technology, an International Journal, 2019, 22(2): 673-691.

[14] WANG J H, DAN Z L, et al. NetWork securety situation assessment based on genetic optimized PNN neural netword[J]. Computer Science, 2021, 48(06): 338-342.

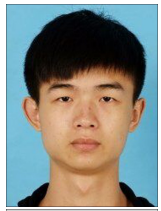
[15] OPITZ J, BURST S. Macro f1 and macro f1[J]. arXiv: 1911.03347, 2019.

[16] SHAH K, PATEL H, SANGHVI D, et al. A comparative analysis of logistic regression, random forest and KNN models for the text classification[J]. Augmented Human Research, 2020, 5(1): 1-16.

[17] TAO P, SUN Z, SUN Z. An improved intrusion detection algorithm based on GA and SVM[J]. IEEE Access, 2018, 6: 13624-13631.

[18] SHI Q, KANG J, WANG R, et al. A framework of intrusion detection system based on Bayesian network in IoT[J]. International Journal of Performability Engineering, 2018, 14(10): 2280-2293.

[19] DONG G, LI W, WANG S, et al. The assessment method of network security situation based on improved BP neural network [C]// International Conference on Computer Engineering and Networks. Cham: Springer, 2018: 67-76.



REN Gaoke, born in 1996, master. His main research interests include network security situation awareness and so on.

MO Xiuliang, born in 1969, associate professor, His main research interests include information security and artificial intelligence.