基于多分类器集成的区块链网络层异常流量检测方法

戴千一[1,2]，张斌[1,2]，郭松[1]，徐开勇[1]
（1. 信息工程大学密码工程学院，河南 郑州 450001；2. 河南省信息安全重点实验室，河南 郑州 450001）

摘 要：为提升对区块链网络层混合型攻击流量的综合泛化特征感知能力，增强异常流量检测性能，提出一种具有支持异常数据综合判决机制和强泛化能力的基于多分类器集成的区块链网络层异常流量检测方法。首先，为扩大所用基分类器的输入特征子集差异度，提出基于区分度和冗余信息量特征子集选择算法，特征筛选过程中激励高区分度子集项输出，同时抑制冗余信息生成。其次，在 Bagging 集成算法中引入随机方差缩减梯度算法动态调整各基模型投票权重，提升对混合型攻击流量的检测泛化能力。最后，为了将集成算法输出的低维数值向量向高维空间映射，提出基于数据场概念的局部离群因子算法，并基于数据点间势差放大各样本数据点空间密度分布差异性，提升异常数据点检测召回率。实验结果表明，相较于单一分类检测器集成方法，所提方法的异常检测准确率、召回率分别平均提升 1.57%、2.71%。

关键词：区块链网络层；集成学习；机器学习；异常流量检测
中图分类号：TN393
文献标志码：A
DOI: 10.11959/j.issn.1000–436x. 2023066

Blockchain network layer anomaly traffic detection method based on multiple classifier integration

DAI Qianyi[1,2], ZHANG Bin[1,2], GUO Song[1], XU Kaiyong[1]
1. Department of Cryptogram Engineering, Information Engineering University, Zhengzhou 450001, China
2. Henan Province Key Laboratory of Information Security, Zhengzhou 450001, China

Abstract: To improve the comprehensive generalized feature perception ability of mixed attack traffic on the blockchain network layer, and enhance the performance of abnormal traffic detection, a blockchain layer traffic anomaly detection method was proposed that supported the comprehensive judgement of data anomaly with a strong generalisation capability. Firstly, to expand the difference of the input feature subset of the base classifier used, a feature subset selection algorithm based on discrimination degree and redundant information was proposed, and the output of high sensitivity subset terms was stimulated during the feature screening process, while the generation of redundant information was suppressed. Then, the stochastic variance reduction gradient algorithm was introduced into the bagging integration algorithm to realize the dynamic adjustment of the voting weights of each base modeland improve thecapability in detecting the generalised hybrid abnormal attack traffic. Finally, LBoF algorithm was proposed to map the low-dimensional numerical vector output by the integrated algorithm to a high-dimensional space. The discrepancy of data point spatial density distribution of various samples were amplified based on the potential difference between data points to increase the recall rate of anomalous data point detection. The experimental results show that in detecting multiple hybrid attack traffic on blockchain layers, the proposed method presents an increase in the anomaly detection accuracy and recall rate, which is 1.57% and 2.71%, respectively, compared with methods based on a single classifier integration.

Keywords: blockchain network layer, ensemble learning, machine learning, anomaly traffic detection

收稿日期：2022–10–12；**修回日期：**2023–03–02
基金项目：信息保障技术重点实验室开放基金资助项目（No.KJ-15-109）；信息工程大学新兴科研方向培育基金资助项目（No.2016604703）；信息工程大学科研基金资助项目（No.2019f3303）
Foundation Items: The Open Fund Project of Information Assurance Technology Key Laboratory (No.KJ-15-109), The New Research Direction Cultivation Fund of Information Engineering University (No.2016604703), The Research Project of Information Engineering University (No.2019f3303)

0　引言

区块链作为一种去信任的新型分布式计算范式为用户提供可信服务，其网络层面临的安全问题愈发严峻[1-2]。区块链技术存在网络体系非结构化、网络流量动态变化性强、管控难度大和网络协议标准不统一等缺点，导致以 DDoS 攻击、Eclipse 攻击和 Erebus 攻击为代表的混合型攻击流量已成为区块链网络层的重要威胁[3-5]。为避免攻击者破坏区块链系统算力平衡性和稳定性，有必要研究区块链网络层异常流量检测方法[6-7]。

区块链网络层异常流量检测技术基于数据挖掘和机器学习等方法挖掘样本的潜在攻击迹象以判断区块链网络层是否存在异常运行或攻击行为，从而降低恶意攻击流量对区块链系统造成的影响[8-9]。基于集成学习的异常流量检测方法通过整合各基分类器的异质输出结果，以构建具有强泛化能力的异常检测模型，并可发挥各基分类器对特定类型攻击流量的针对性检测优势，提升集成模型输出特征的稳健性和稳定性[10]。

根据基模型集成模式的不同，传统的异常流量检测采用的集成学习算法分为序贯方式集成学习算法（如 Boosting 算法）和并行化集成学习算法（如 Bagging 算法）两类[11-13]。Boosting 算法通过对样本重采样为各基分类器生成相应特征子集，使用弱分类检测器对各特征子集进行训练，并使用串行策略进行层次式基分类器集成[14]，每个基分类器基于前序学习器输出结果进行模型参数调整[15]。Boosting 算法对不稳定型弱分类器有较强的集成效果，且集成泛化误差较低[16]。但 Boosting 算法中后序模型参数依赖前序模型，训练过程中噪声数据易造成因前序基模型过拟合而导致 Boosting 算法整体收敛效果降低[17]。Bagging 算法使用有放回式策略随机选取基模型的特征子集，采取独立并行方式训练各基分类器，并对输出分类决策结果进行投票表决[18]。Bagging 算法对易受样本扰动的强差异基学习器具有较可靠的集成效果，可降低样本数据波动性导致的集成算法泛化误差，并可减少因样本数据随机偏移而导致集成异常检测结果出现波动性[19]。但经典 Bagging 算法采用等概率抽样和等权重投票模式，无法侧重训练特征集中的特定实例，导致强作用特征子项和重点基分类器权重欠缺激励针对性[20-21]。

攻击者会利用区块链网络层固有缺陷动态选择流量攻击的方法和策略，该特点会扩大区块链网络层中各类型攻击流量分类检测模型之间的差异程度，而 Bagging 算法对差异度高、异构性强的基分类器具有较好的集成效果。除此之外， Bagging 算法可有效应对因少数类流量样本变化而导致集成算法突变的问题，并可提升对混合型攻击流量样本核心特征的提取能力和检测泛化能力。特别地，基于多分类器集成学习的异常检测模型相比于单个分类检测器具有更好的泛化能力和检测性能。因此，本文选择并改进 Bagging 算法，对区块链网络层中混合型攻击流量进行集成学习，针对特征选择和模型集成 2 个过程分别改进特征选择算法和集成学习算法，并形成可综合感知区块链网络层混合型攻击流量的异常检测模型，以提升区块链网络层应对混合型攻击流量的主动防御能力。

本文主要贡献如下。

1) 为解决经典 Bagging 算法因采用随机采样策略导致生成弱差异度化、高冗余度特征子集的问题，提出区分度和冗余信息量特征子集（D&RFS, discernibility and redundancy of feature subset）选择算法。该算法结合 Bootstrap 算法进行样本采样，并对各基模型的输入样本特征实现有监督输入特征扰动，激励各基模型提升输入特征差异性的同时放大各基分类模型输出结果差异度；在 D&RFS 特征筛选过程中基于冗余信息惩罚策略降低输出特征子集复杂度，为各基分类器生成强稳健性特征子集，进而提升集成算法泛化性能。

2) 为解决经典 Bagging 算法因等权重投票表决方式而造成集成模型异常检测泛化能力下降的问题，提出改进 Bagging 算法。该算法在集成过程中引入动态自适应权重投票集成策略，增加集成算法对混合型攻击流量核心特征的综合感知能力，并基于自学习式随机方差缩减梯度（SVRG , stochastic variance reduction gradient）算法调整各基分类器投票权重，避免集成算法陷入局部最优困境，以获得集成算法最优稳定输出模型。理论分析表明，所提集成算法泛化性能优于单一分类检测器。

3) 为提升改进 Bagging 算法输出的低维概率数值向量的异常数据点检测能力，提出基于数据场概念的局部离群因子（LBoDF, LOF based on data field）算法。该算法通过 Gaussian 核方法将低维数值向量映射为高维数值空间以优化低维线性不可

分的样本空间分布，基于势概念模型重定义空间距离以扩大稠密向量集中正常数据点和异常数据点的空间分布差异性，有效放大异常数据点的重定义局部离群值，提升对密度分布不均匀型异常样本的检测准确率。

1 基于改进 Bagging 算法的多分类器集成异常检测方法

首先，采用 D&RFS 特征选择算法提取低冗余高区分度的特征子集，扩大各基模型输入特征的差异性；然后，通过改进 Bagging 算法动态调整各分类器投票权重，归一化输出检测结果；最后，使用 LBoDF 算法对集成结果输出值进行离群点检测，挖掘输入样本的异常数据点。根据提出的异常检测模型并结合异常流量检测和机器学习的关键步骤，设计基于改进 Bagging 算法的多分类器集成异常检测方法，其总体结构如图 1 所示。

1.1 D&RFS 算法的特征子集构建方法

经典 Bagging 算法由基模型和集成算法 2 个部分构成。假设全体基模型所生成特征合集为 $D=\{(x_i,y_i)\mid x_i\in\mathbb{R}^m,y_i\in\{-1,1\}\}$，$i\in\{1,2,\cdots,n\}$，$y\in\mathbb{R}$ 为输出变量，x_i 为 m 维特征向量，y_i 为标签项，–1 和 1 分别为异常样本和正常样本。经典 Bagging 算法通过多轮 Bootstrap 算法进行样本重采样以生成训练子集$\{D_1,D_2,\cdots,D_N\}$，并根据相应子集生成具有差异性输出的基模型$\{\ell_1,\ell_2,\cdots,\ell_N\}$，分类标签集合记为 $Y=\{-1, 1\}$，但该算法使生成的各输入训练集与原始数据集的特征重合度较高。而弱差异化特征子集不利于扩大各基模型输出差异性和集成算法的泛化性能，所以在特征子集生成阶段引入 D&RFS 算法。该算法结合 Bootstrap 算法进行样本重采样，可实现基模型的有监督输入特征扰动，扩大各基模型输入特征子集的差异性，可提升特征子集特征丰富性并平衡各基分类器的特征差异性。特别地，在特征子

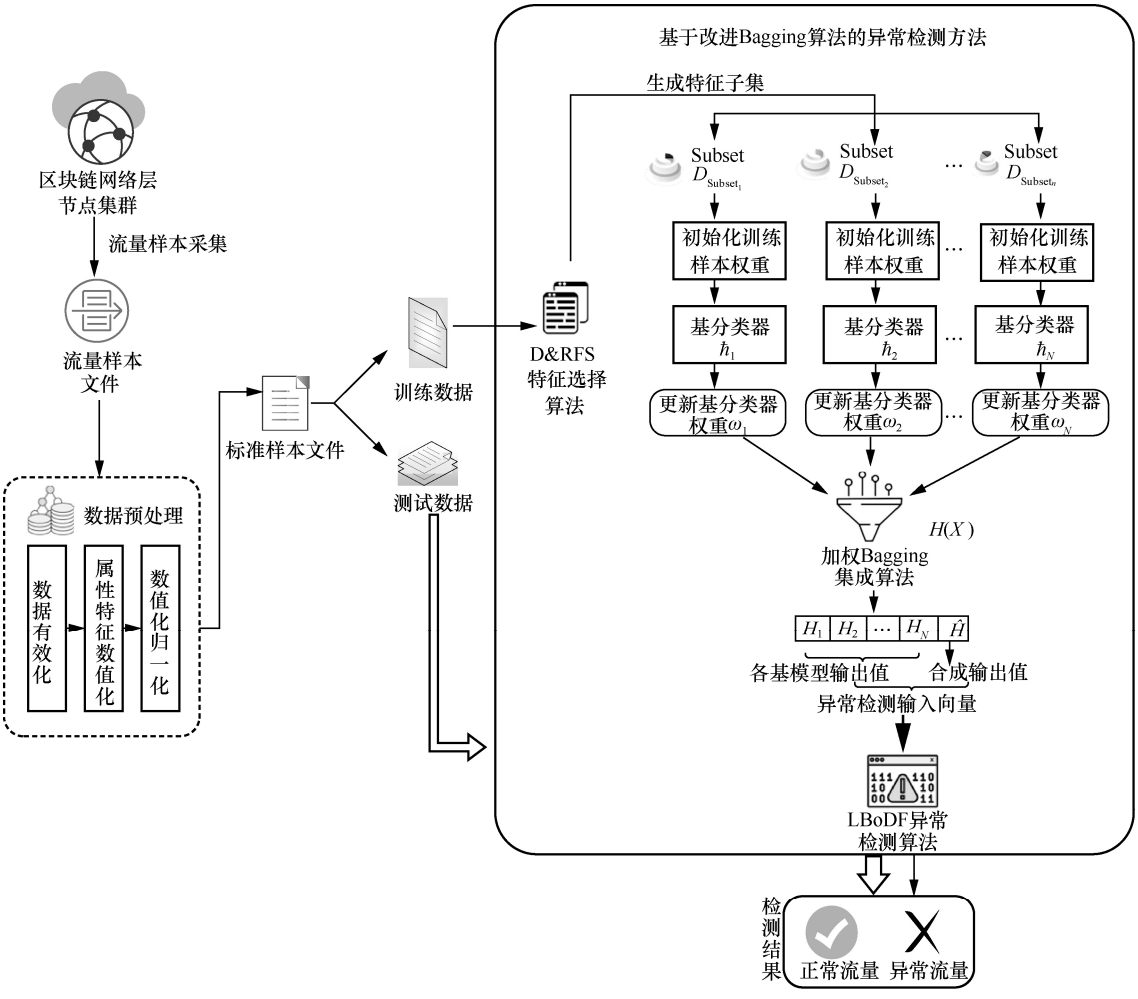


图 1 改进 Bagging 算法的多分类器集成异常检测方法总体结构

集筛选过程中特征间冗余信息和噪声会降低特征子集稳健性，所以必须兼顾处理各子集内非必要冗余信息。

D&RFS 算法采用最大特征独立性和最小特征间冗余信息量的惩罚策略提取高区分度特征子集。首先，D&RFS 算法在经典 DFS（discernibility of feature subset）框架基础上生成特征子集，从输入样本特征中筛选子集间强区分性特征项，从而支持改进 Bagging 算法提升泛化性能；然后，计算类内特征冗余信息量并在特征选择框架中进行抑制和惩罚，以降低子集内特征冗余度，进而降低集成算法计算开销和提升输出特征子集稳健性。

对于包含 κ 类输出标签的特征子集 S，在 DFS 特征选择框架中特征子集区分度计算式为

$$\mathrm{DFS}(S)=\frac{\sum_{i=1}^{\kappa}(\bar{\boldsymbol{x}}^{(i)}-\bar{\boldsymbol{x}})^2}{\sum_{i=1}^{\kappa}\frac{1}{n_i-1}\sum_{j=1}^{n_i}(\boldsymbol{x}_j^{(i)}-\bar{\boldsymbol{x}}^{(i)})^2} \tag{1}$$

其中，$\bar{\boldsymbol{x}}$ 为样本特征的均值向量，$\bar{\boldsymbol{x}}^{(i)}$为第 i 类样本特征的均值向量，$\boldsymbol{x}_j^{(i)}$ 为第 j 类样本的第 i 类特征向量，DFS(S)的输出值范围为[0,1]。式(1)的分子项描述子集内特征稀疏度，分母项描述子集间特征聚集度，故 DFS(S)值越大说明特征子集的类间区分度和差异性越大，利于各基模型学习差异化输入特征向量。为描述子集内特征间冗余信息量，使用互信息量来描述特征冗余度，其输出值范围为[0,1]，记为

$$\mathrm{Red}'(S)=\frac{1}{|S|^2\,k}\sum_{i=1}^{k}\sum_{\alpha_1,\alpha_2\in S}\sum_{j=1}^{n_i}p([x_j^{(i)}]_{\alpha_1},[x_j^{(i)}]_{\alpha_2})\cdot \log\frac{p([x_j^{(i)}]_{\alpha_1},[x_j^{(i)}]_{\alpha_2})}{p([x_j^{(i)}]_{\alpha_1})p([x_j^{(i)}]_{\alpha_2})} \tag{2}$$

其中，$\|\cdot\|$表示集合中元素数量，$0\leqslant\|S\|\leqslant m$；$[x_j^{(i)}]_{\alpha_1}$ 和 $[x_j^{(i)}]_{\alpha_2}$ 分别表示第 i 个特征向量组中特征项 α_1 和 α_2 在第 j 个样本中的特征值，Red′(S)表示 S 中特征间冗余信息量。区分度和冗余度是 2 个互斥的特征相关性评价标准。为改进 DFS 特征选择框架过程中仅考虑类间相关性而忽略类内特征冗余的情况，定义最大化类间特征独立度和最小化冗余信息量特征选择框架以筛选最优特征子集。D&RFS 特征选择框架定义如式(3)所示，其输出值范围为[0,1]。

$$\mathrm{D\&RFS}(S)=\mathrm{DFS}(S)-\mathrm{Red}'(S) \tag{3}$$

D&RFS 特征选择算法如算法 1 所示。首先，算法从空集开始进行搜索，将特征合集中具有最强类间区分能力特征 feature 加入特征子集 D_{Subset} 中；其次，将待选特征项与已筛选的强类间区分能力特征子集 D 中各特征项进行组合形成更新特征子集；最后，将更新特征子集 D'输入生成集成机器模型 f 中进行检测，根据分类准确率判定是否将待选特征项保留。通过对每个特征项组合执行迭代，直至所有特征项组合完成验证测试。

算法 1　D&RFS 特征选择算法

输入　$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$

输出　D_{Subset}

1) $D_{\mathrm{Subset}}\leftarrow\Phi$
2) $\mathrm{ACC}_{\max}=0$
3) 提取 D 中所有特征项{feature$_1$, feature$_2$,$\cdots$, feature$_m$}
4) for i=1 to m:
5)　if　DFS(feature$_i$)为最大值项
6)　　$D_{\mathrm{Subset}}\leftarrow$feature$_i$ //选择 DFS 值最大项，加入 D_{Subset} 空集中
7) 使用基分类器进行测试，计算 D_{Subset} 的分类准确率 $\mathrm{ACC}_{\max}$
8) for i =1 to m
9)　使用 Bootstrap 算法选取特征项 $x^{(i)}$
10)　$D_{\mathrm{Subset}}=D_{\mathrm{Subset}}\cup\{x^{(i)}\}$
11)　使用基分类器进行测试，计算 D_{Subset} 的分类准确率 $\mathrm{ACC}_{\mathrm{temp}}$
12)　if ($\mathrm{ACC}_{\max}<\mathrm{ACC}_{\mathrm{temp}}$):
13)　　$\mathrm{ACC}_{\max}=\mathrm{ACC}_{\mathrm{temp}}$
14)　　else
15)　$D_{\mathrm{Subset}}=D_{\mathrm{Subset}}\backslash\{x^{(i)}\}$

由于各基分类器对不同类型攻击流量进行针对性检测，各基分类器所定义的原始攻击流量特征集并不一致。特征合集中不仅包含部分重复定义或近似特征项，也包含各基分类器针对不同类型攻击流量的特定特征项。DFS 特征选择算法对特征合集进行采样和扰动，为各基分类器所生成的特征子集具有一定的随机性。D&RFS 特征选择算法充分保留了 DFS 算法为各基分类器所生成特征子集的高

随机性特点，并可发挥激励高区分度子集项输出能力，降低各基分类器所获得的特征子集冗余度，使各基分类器可充分感知不同特征项对集成模型输出结果带来的影响，并可提升改进 Bagging 集成算法对混合型特征集的检测泛化能力。D&RFS 特征选择算法输出结果示例如图 2 所示。该图示意了特征选择算法过程中随机选取特征项及输出特征集扰动的过程。

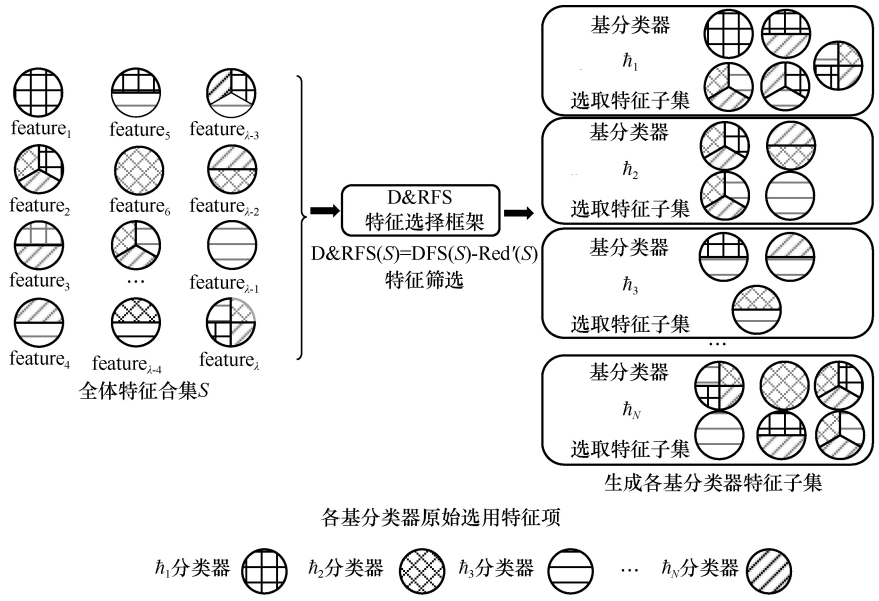


图 2 D&RFS 特征选择算法输出结果示例

1.2 基于动态自适应加权的 Bagging 集成算法

为整合各基分类器的差异性输出结果，增强核心学习器在改进 Bagging 集成算法中的输出效果，本文在集成过程中引入动态自适应加权集成方法，动态调整每个基分类器投票权重，以提升基于改进 Bagging 算法的多分类器集成异常检测模型泛化能力。基于改进 Bagging 算法的多分类器集成过程如图 3 所示。

对于 D&RFS 算法生成的特征子集，$D_{\text{Subset}}=\{(D_1,y_1),(D_2,y_2),\cdots,(D_n,y_n)\}, y^{(i)}\in\{1,2,\cdots,\kappa\}$。其中，$\kappa$ 为标签索引值，$D_i\in\mathbb{R}^{m'}$，每个特征子集维度为 m'。默认每个基分类器均使用 Softmax 作为回归函数，相应假设函数为

$$\ell(D)=[p(y^{(1)}\mid D),p(y^{(2)}\mid D),\cdots,p(y^{(\kappa)}\mid D)]^{\mathrm{T}}=\frac{1}{\sum_{j=1}^{\kappa}\mathrm{e}^{\boldsymbol{\theta}_j^{\mathrm{T}}D}}[\mathrm{e}^{\boldsymbol{\theta}_1^{\mathrm{T}}D},\mathrm{e}^{\boldsymbol{\theta}_2^{\mathrm{T}}D},\cdots,\mathrm{e}^{\boldsymbol{\theta}_\kappa^{\mathrm{T}}D}]^{\mathrm{T}} \tag{4}$$

其中，$\boldsymbol{\theta}_i^{\mathrm{T}}$为参数向量，$\boldsymbol{\theta}_j^{\mathrm{T}}D$代表各参数项与数据集 D 的向量积。$\ell(D)$的输出由类标签 $y^{(i)}$的后验概率构成，假设第 i 个基模型的输出为 $\hbar_i(D)$，根据文献[11-12]中训练基于 Bagging 算法的异常检测模型经验，将 Softmax 输出的区分值定为 0.5。

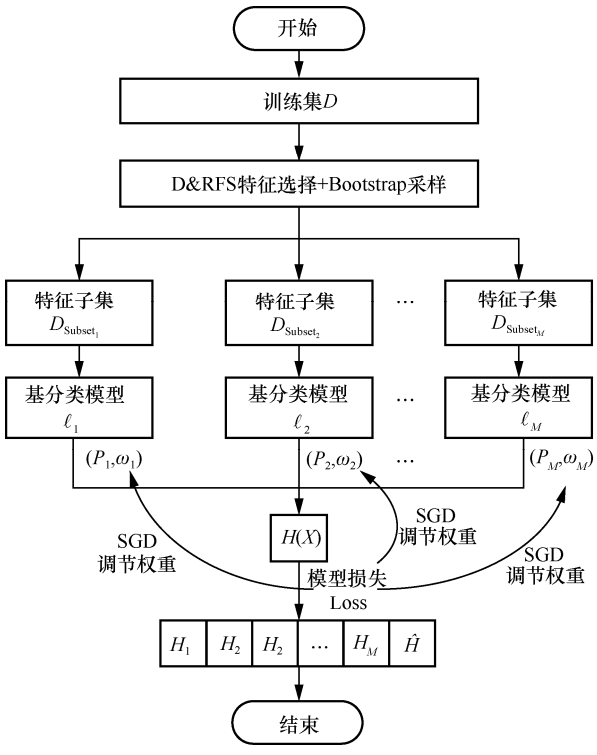


图 3 基于改进 Bagging 算法的多分类器集成过程

$$\hbar_i(D)=\begin{cases}1 & ,\ell(D)\geqslant 0.5\\ -1 & ,\ell(D)<0.5\end{cases} \tag{5}$$

考虑到输出向量加权集成后可能存在不满足概率分布的情况，故使用 Softmax 函数对 $\hbar(D)$再次进行回归输出，记为

$$\boldsymbol{P}=\mathrm{Softmax}\left(\sum_{i=1}^{M}\omega_i\hbar_i(D_i)\right) \tag{6}$$

其中，M 为基模型个数，$\boldsymbol{P}=[p_1,\ p_2,\cdots,p_\kappa]^{\mathrm{T}}$，$p_i\in[0,1]$，$\sum_{i}^{\kappa}p_i=1$。重新赋权重后的假设函数输出值转换为

$$H(x)=\arg\max_i(P_i) \tag{7}$$

将各基分类器输出结果记为向量 $\boldsymbol{H}=[H_1, H_2,\cdots, H_M]^{\mathrm{T}}$。在设计模型损失函数时，使用交叉熵函数作为集成算法损失函数，记为

$$\mathrm{Loss}=-\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k}\mathrm{sgn}(H(x^{(i)})=y_j)\ln p_j \tag{8}$$

其中，Loss 是间接关于投票权重$(\omega_1, \omega_2,\cdots,\omega_M)$的函数，将权重组合记为向量 $\boldsymbol{\omega}=(\omega_1, \omega_2,\cdots,\omega_M)$。为避免梯度下降训练过程中梯度训练模型陷入局部最优困境，采用迭代型 SVRG 方法动态调整投票权重向量[22]，表达式为

$$\tilde{g}_i^{(t)}\leftarrow\frac{\partial\mathrm{Loss}}{\partial\omega_i^{(t)}}-\left(\frac{\partial\mathrm{Loss}}{\partial\omega_i}-\frac{1}{M}\sum_{i=1}^{M}\frac{\partial\mathrm{Loss}}{\partial\omega_i}\right) \tag{9}$$

$$\omega_i^{(t+1)}\leftarrow\omega_i^{(t)}-\alpha\tilde{g}_i^{(t)} \tag{10}$$

$$\frac{\partial\mathrm{Loss}}{\partial\omega_l}=-\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k}\mathrm{sgn}(H(x^{(i)})=y_j)[h_l(x)^{(i)}]_j\ln(1-p_j) \tag{11}$$

$$\widehat{H}=\sum_{i=1}^{M}\omega_iH_i \tag{12}$$

其中，$[\cdot]_j$ 为权重向量第 j 个分量。通过迭代训练获得最优权重 $\boldsymbol{\omega}^*=(\omega_1^*, \omega_2^*,\cdots, \omega_M^*)$。将各基模型输出结果与 H 组合形成改进 Bagging 算法的输出向量 $\boldsymbol{\theta}=(H_1, H_2,\cdots,H_N, \hat{H})$，$H_i\in[0,1]$，$\hat{H}\in[0,1]$。将历次输出结果记为 $\boldsymbol{\Theta}=(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2,\cdots,\boldsymbol{\theta}_\tau)$，$\tau$ 为模型训练次数。

1.3　基于数据场概念的改进 LOF 算法

经典的 LOF 算法是基于 Euclid 距离定义下的非监督型异常检测算法，其通过对比每个检测数据和其邻域的空间密度分布情况来计算数据点所在区域空间密集程度，进而判断输入数据是否为离群型异常数据点[23]。LOF 算法检测准确率在一定条件下依赖输出向量的维度，维度较高有利于 LOF 算法挖掘样本数据的空间分布规律。

区块链网络层攻击流量会采用流量伪装和低速渗透等攻击手段，导致异常攻击流量与正常流量存在低维特征相似现象，使改进 Bagging 算法输出的组合向量存在低维空间数值分布空间密度分布不均匀现象。默认空间中等权重距离分布的经典 LOF 算法在低维稠密空间对正常数据点与异常数据点空间分布区分度的描述能力存在欠缺，导致异常检测准确率较低。为扩大数据点间空间分布差异程度，提出基于数据场概念的 LOF 算法。基于数据场概念的样本数据点分布示意如图 4 所示。首先，通过 Gauss 核方法将低维输入向量向高维空间映射；然后，通过数据场中的势概念来量化数据点分布的稀疏程度，势差越小则数据点密集程度越高，势差越大则数据点离群程度越高；最后，通过计算各数据点在场空间的平均势差来确定各数据点 Euclid 距离权重，自适应地放大各数据点空间距离以放大正常与异常数据点空间分布的区分度和离群点检测的精确度，缓解因不合理的距离定义所导致的内部空间密度稠密化现象及检测误报率和漏报率升高的情况。

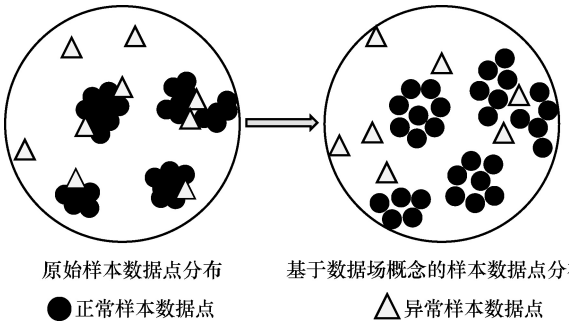


图 4　基于数据场概念的样本数据点分布示意

对原始流量进行多次采样，分别输入改进 Bagging 集成算法进行迭代训练。将算法历次输出的向量结果 $\boldsymbol{\theta}$ 进行组合，形成输出向量集合 $\boldsymbol{\Theta}=(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2,\cdots, \boldsymbol{\theta}_\tau)$。通过对原始流量样本进行大量采集和多次实验，以提升集合 $\boldsymbol{\Theta}$ 的丰富性。基于 LBoDF 算法学习改进 Bagging 算法输出向量的统计分布规律以构建相应数据点空间分布模型，通过对比各数据点离群值来判定输入样本是否为异常数据点。LBoDF 算法包含以下定义。

定义 1　数据场空间中各数据点势能函数为

$$\mathrm{PE}(\boldsymbol{\theta})=\sum_{i=1}^{\mu}\mathrm{e}^{-\frac{(\boldsymbol{\theta}-\boldsymbol{\theta}_i)^2}{2\delta^2}} \tag{13}$$

定义 1 结合了径向基函数（RBF, radial basis function）中高斯核函数方法，可以将低维数据映射到高维空间，以放大数据点空间分布差异度。数据场空间中数据点间的辐射能为 $\mathrm{e}^{-\frac{(\boldsymbol{\theta}_i-\boldsymbol{\theta})^2}{2\delta^2}}$，$i\in[1,2,\cdots,\mu]$，$\delta$ 为势能函数辐射因子，用于在不同数据集中调整输出值。数据点 $\boldsymbol{\theta}$ 的势能函数为 $\boldsymbol{\Theta}$ 中所有数据点至该点辐射能的叠加。

定义 2 $\boldsymbol{\Theta}$ 的数据场空间平均势差为

$$\overline{\nabla}\mathrm{PE}_{\boldsymbol{\Theta}}=\frac{1}{\tau^2}\sum_{\boldsymbol{\theta}_p,\boldsymbol{\theta}_q\in\boldsymbol{\Theta}}|\mathrm{PE}(\boldsymbol{\theta}_p)-\mathrm{PE}(\boldsymbol{\theta}_q)| \tag{14}$$

对于 $\boldsymbol{\Theta}$ 中任意两数据点 $\boldsymbol{\theta}_p$ 和 $\boldsymbol{\theta}_q$，通过定义 1 计算两数据点势差 $\mathrm{PE}(\boldsymbol{\theta}_p)$–$\mathrm{PE}(\boldsymbol{\theta}_q)$。$\boldsymbol{\Theta}$ 平均势差为 $\boldsymbol{\Theta}$ 中任意两数据点间势差总和的加权平均值。

定义 3 加权空间距离为

$$\mathrm{weight\_dist}(\boldsymbol{\theta}_p,\boldsymbol{\theta}_q)=\chi\sqrt{(\boldsymbol{\theta}_p-\boldsymbol{\theta}_q)^2} \tag{15}$$

$$\chi=\begin{cases}\eta, & |\mathrm{PE}(\boldsymbol{\theta}_p)-\mathrm{PE}(\boldsymbol{\theta}_q)|\geqslant\overline{\nabla}\mathrm{PE}_{\boldsymbol{\Theta}}\\ 1, & |\mathrm{PE}(\boldsymbol{\theta}_p)-\mathrm{PE}(\boldsymbol{\theta}_q)|<\overline{\nabla}\mathrm{PE}_{\boldsymbol{\Theta}}\end{cases} \tag{16}$$

式(15)在 Euclid 距离的基础上自适应放大 $\boldsymbol{\theta}_p$ 和 $\boldsymbol{\theta}_q$ 之间的距离，放大权重 χ 由式(16)定义。式(16)中通过判断两数据点势差大于 $\boldsymbol{\Theta}$ 的平均势差，针对性放大稠密样本集群中数据点空间距离，进而通过加权空间距离实现放大正常数据点与异常数据点之间的空间分布。

定义 4 $\boldsymbol{\theta}_p$ 的改进 k-距离。对于任意 $k\in R^+$，$\boldsymbol{\theta}_p$ 的改进 k-距离记为 $\mathrm{dist}'_k(\boldsymbol{\theta}_p)$，$\boldsymbol{\theta}_p$ 与某数据点 $\boldsymbol{\theta}_o$ 之间的加权空间距离记为 $\mathrm{dist}'_k(\boldsymbol{\theta}_p,\boldsymbol{\theta}_o)$。当 $\boldsymbol{\theta}_o$ 满足以下条件时，$\mathrm{dist}'_k(\boldsymbol{\theta}_p)=\mathrm{dist}'_k(\boldsymbol{\theta}_p,\boldsymbol{\theta}_o)$。

1) 至少存在 k 个数据点 $\boldsymbol{\theta}'_o\in\boldsymbol{\Theta}$，满足

$$\mathrm{dist}'_k(\boldsymbol{\theta}_p,\boldsymbol{\theta}'_o)\leqslant\mathrm{weighted\_dist}(\boldsymbol{\theta}_p,\boldsymbol{\theta}_o) \tag{17}$$

2) 至多存在 k 个数据点 $\theta'_o\in\boldsymbol{\Theta}$，满足

$$\mathrm{dist}'_k(\boldsymbol{\theta}_p,\boldsymbol{\theta}'_o)>\mathrm{weighted\_dist}(\boldsymbol{\theta}_p,\boldsymbol{\theta}_o) \tag{18}$$

定义 5 $\boldsymbol{\theta}_p$ 的改进 k-距离邻域。以点 $\boldsymbol{\theta}_p$ 为圆心，并以 $\boldsymbol{\theta}_p$ 的改进 k-距离为半径，圆内所包含的数据点为 $\boldsymbol{\theta}_p$ 的改进 k-距离邻域记为 $N'_k(\boldsymbol{\theta}_p)$，定义为

$$N'_k(\boldsymbol{\theta}_p)=\{\boldsymbol{\theta}_q\in\boldsymbol{\Theta}\,|\,\mathrm{weighted\_dist}(\boldsymbol{\theta}_p,\boldsymbol{\theta}_q)\leqslant\mathrm{dist}'_k(\boldsymbol{\theta}_p)\} \tag{19}$$

定义 6 $\boldsymbol{\theta}_p$ 相对 $\boldsymbol{\theta}_o$ 的可达邻域。对于参数 k，数据点 $\boldsymbol{\theta}_p$ 相对于 $\boldsymbol{\theta}_o$ 的可达距离记为 $\mathrm{Reach\_Dist}'_k(\boldsymbol{\theta}_p,\boldsymbol{\theta}_o)$，定义为

$$\mathrm{Reach\_Dist}'_k(\boldsymbol{\theta}_p,\boldsymbol{\theta}_0)=\max\{\mathrm{dist}'_k(\boldsymbol{\theta}_p,\boldsymbol{\theta}_o),\mathrm{weighted\_dist}(\boldsymbol{\theta}_p,\boldsymbol{\theta}_o)\} \tag{20}$$

定义 7 $\boldsymbol{\theta}_p$ 的局部可达密度（LRD, local reachability density）。$\boldsymbol{\theta}_p$ 的局部可达密度为 $\boldsymbol{\theta}_p$ 相对于其改进 k-距离邻域的平均可达距离倒数，记为

$$\mathrm{LRD}(\boldsymbol{\theta}_p)=\frac{\left\|N'_k(\boldsymbol{\theta}_p)\right\|}{\sum_{\boldsymbol{\theta}_o\in N'_k(\boldsymbol{\theta}_p)}\mathrm{Reach\_Dist}'_k(\boldsymbol{\theta}_p,\boldsymbol{\theta}_o)} \tag{21}$$

定义 8 $\boldsymbol{\theta}_p$ 的局部异常因子为

$$\mathrm{LBoDF}_k(\boldsymbol{\theta}_p)=\frac{\sum_{\boldsymbol{\theta}_o\in N'_k(\boldsymbol{\theta}_p)}\frac{\mathrm{LRD}(\boldsymbol{\theta}_o)}{\mathrm{LRD}(\boldsymbol{\theta}_p)}}{\left\|N'_k(\boldsymbol{\theta}_p)\right\|} \tag{22}$$

$\mathrm{LBoDF}_k(\boldsymbol{\theta}_p)$表示在参数 k 条件下 $\boldsymbol{\theta}_p$ 的离群分布异常程度，其值越大说明该样本点离群程度越高，其样本是异常数据的可能性越大。

根据定义 1~定义 8，给出 LBoDF 算法过程。

Step1 计算各输入数据点势能函数 $\mathrm{PE}(\boldsymbol{\theta}_i)$。

Step2 计算 $\boldsymbol{\Theta}$ 的数据场空间平均势差 $\overline{\nabla}\mathrm{PE}_{\boldsymbol{\Theta}}$。

Step3 确定 $\boldsymbol{\Theta}$ 的离群因子阈值 ζ。

Step4 计算并存储各数据点 $\boldsymbol{\theta}_p$ 的改进 k-距离邻域。

Step5 计算每个数据点的局部可达密度，即 $\mathrm{LRD}(\boldsymbol{\theta}_p)$和 $\mathrm{LBoDF}(\boldsymbol{\theta}_p)$值。

Step6 对各数据点的 $\mathrm{LBoDF}(\boldsymbol{\theta}_p)$值进行排序，大于离群因子阈值 ζ 的数据点标记输出为异常。

1.4 改进 Bagging 算法错误率界分析

错误率可反映机器模型输出值与实际值间的误差，即模型的精准度和拟合能力。由于所提改进 Bagging 算法采用加权投票策略进行集成，故通过分析集成方法的错误率界和期望错误率界来分析模型偏移变化情况，进而证明所提改进 Bagging 算法相较于单一分类检测器针对混合型流量样本可降低错误率界。

引理 1 加权投票型 Bagging 集成算法 f 对 D 的训练错误率界应满足以下严格界。

$$\frac{1}{N}|\mathrm{sgn}_i(H(x^{(i)})\neq y_i)|\leqslant\frac{1}{N}\prod_{j=1}^{M}Z_j \tag{23}$$

其中，有

$$Z_j=\sum_{i=1}^{N}\exp(-\omega_j y_i \hbar_j(x^{(i)})) \tag{24}$$

证明　如果 $H(x^{(i)})\neq y_i$，则 $f(x^{(i)})y_i\leqslant 0$，即 $\exp(-f(x^{(i)})y_i)\geqslant 0$，可得

$$\begin{aligned}|\mathrm{sgn}_i(H(x^{(i)})\neq y_i)|&\leqslant\sum_{i=1}^{N}\exp(-y_i f(x^{(i)}))=\\&\sum_{i=1}^{N}\exp(-y_i\sum_{j=1}^{M}\omega_j\hbar_j(x^{(i)}))=\\&\sum_{i=1}^{N}\prod_{j=1}^{M}\exp(-\omega_j y_i\hbar_j(x^{(i)}))\leqslant\\&\prod_{j=1}^{M}\sum_{i=1}^{N}\exp(-\omega_j y_i\hbar_j(x^{(i)}))=\prod_{j=1}^{M}Z_j\end{aligned} \tag{25}$$

进而可得

$$\frac{1}{N}|\mathrm{sgn}_i(H(x^{(i)})\neq y_i)|\leqslant\frac{1}{N}\prod_{j=1}^{M}Z_j \tag{26}$$

式(26)左边为集成算法的训练错误率，右边为基分类器 $\hbar_j$ 基于权重 ω_j 所生成错误率。证毕。

引理 2　加权投票型 Bagging 集成算法 f 对 D 的训练错误率界应满足以下宽松界。

$$\frac{1}{N}|\mathrm{sgn}_i(H(x^{(i)})\neq y_i)|\leqslant\frac{1}{N}\prod_{j=1}^{M}\Gamma_j \tag{27}$$

其中，有

$$\Gamma_j=\frac{1}{N}\sum_{i=1}^{N}\exp\left(-y_i\hbar_j(x^{(i)})\right)^{\omega_j} \tag{28}$$

证明　由引理 1 可知，有

$$\begin{aligned}\frac{1}{N}|\mathrm{sgn}_i(H(x^{(i)})\neq y_i)|&\leqslant\frac{1}{N}\sum_{i=1}^{N}\prod_{j=1}^{M}\exp(-\omega_j y_i\hbar_j(x^{(i)}))=\\&\frac{1}{N}\sum_{i=1}^{N}\prod_{j=1}^{M}\left(\exp(-y_i\hbar_j(x^{(i)}))\right)^{\omega_j}\end{aligned} \tag{29}$$

由于 $\sum_{i=1}^{M}\omega_j=1$，根据 Hölder 不等式可得

$$\begin{aligned}&\sum_{i=1}^{N}\prod_{j=1}^{M}\left(\exp(-y_i\hbar_j(x^{(i)}))\right)^{\omega_j}\leqslant\\&\prod_{j=1}^{M}\left(\left(\sum_{i=1}^{N}\exp(-y_i\hbar_j(x^{(i)})\right)\right)^{\omega_j}\end{aligned} \tag{30}$$

将式(25)代入式(30)，则有

$$\frac{1}{N}|\mathrm{sgn}_i(H(x^{(i)})\neq y_i)|\leqslant\frac{1}{N}\prod_{j=1}^{M}\Gamma_j \tag{31}$$

证毕。

通过引理 2 中的训练错误率宽松界可用于估计加权投票型 Bagging 集成算法期望错误率界。由于 $\hbar_i(x_i)\in\{-1,1\}$，因此可将 Γ_j 中 $\sum_{i=1}^{N}\exp(-y_i\hbar_j(x^{(i)}))$ 项分解。

$$\begin{aligned}\sum_{i=1}^{N}\exp(-y_i\hbar_j(x^{(i)}))&=\sum_{i:y_i f(x^{(i)})=1}\exp(-y_i\hbar_j(x^{(i)}))+\\&\sum_{i:y_i f(x^{(i)})=-1}\exp(-y_i\hbar_j(x^{(i)}))=\sum_{i:y_i f(x^{(i)})=1}\mathrm{e}^{-1}+\sum_{i:y_i f(x^{(i)})=-1}\mathrm{e}\end{aligned} \tag{32}$$

假设每个基分类器模型的泛化误差为 P_ε，即 $P(\hbar_i(x)\neq f(x))=P_\varepsilon$。当样本数量 N 足够充分时，则有

$$\begin{cases}\left\|\{\mathrm{sgn}(y_i f(x^{(i)}))=1\}\right\|=(1-p_\varepsilon)N\\\left\|\{\mathrm{sgn}(y_i f(x^{(i)}))=-1\}\right\|=p_\varepsilon N\end{cases} \tag{33}$$

结合式(33)和 Γ_j 的定义，可得

$$\begin{aligned}\prod_{j=1}^{M}\Gamma_j&=\prod_{j=1}^{M}\left(\frac{N(1-p_\varepsilon)\mathrm{e}^{-1}+Np_\varepsilon\mathrm{e}}{N}\right)^{\omega_j}=\\&(1-p_\varepsilon)\mathrm{e}^{-1}+p_\varepsilon\mathrm{e}\end{aligned} \tag{34}$$

结合引理 2 可得

$$\frac{1}{N}|\mathrm{sgn}_i(H(x^{(i)})\neq y_i)|\leqslant(1-p_\varepsilon)\mathrm{e}^{-1}+p_\varepsilon\mathrm{e} \tag{35}$$

式(35)给出了加权投票集成策略对于数据集 S 的粗略上界，式(35)说明如果各基分类器降低错误率，则改进 Bagging 算法的总体错误率会相应下降。记 $e_s=\frac{1}{N}|\mathrm{sgn}_i(H(x^{(i)})\neq y_i)|$，对式(35)两边求期望可得改进 Bagging 算法的期望界。

$$P_{j\sim D}\left[H(x)\neq y\right]=\mathrm{E}_{j\sim D}(e_s)\leqslant\mathrm{E}\left(\prod_{j=1}^{M}\Gamma_j\right) \tag{36}$$

式(36)说明，如果各 $\hbar_i(x)$ 的错误率相互独立，则 Γ_j 互相独立，所提 Bagging 集成算法的错误率期望界取得最小值。由此可延伸出以下结论：1) 为获得更好的 Bagging 集成算法输出效果，应扩大各基分类器互相独立性，以提升降低集成模型错误率的界；2) 所提 Bagging 集成算法错误率的期望界小于单一独立分类器模型错误率的期望界，即代表改进 Bagging 集成算法的拟合能力强于单一分类检测器。

由于 Γ_j 中反映基分类器错误率的 $-y_i\hbar_i(x^{(i)})$ 出现在式(28)指数项，因此 Bagging 集成算法的期望

错误率界随各基分类器错误率的下降以指数级形式下降，并且集成模型越丰富该界越低。这说明扩大基分类器数量和扩大输出特征子集差异度可降低改进 Bagging 算法的模型误差。

2 实验与分析

搭建区块链网络实验环境，模拟真实区块链交易行为流量和区块链网络层 DDoS 攻击、Eclipse 攻击、Erebus 攻击等多类型攻击行为流量。首先，对所提基于异常检测方法进行训练，验证本文方法的可行性；其次，分析所提出的 D&RFS 算法和 LBoDF 算法有效性，确定相关算法最佳参数并验证参数设置合理性；最后，为检验本文方法的性能优势，与经典机器学习的异常检测方法、单一分类检测器和现有区块链网络层异常流量检测方法在相同条件下进行实验性能对比。

特别地，集成算法的基分类器模型（ℓ_1～ℓ_3）由课题组前期所提出的 DDoS 攻击[24]、Eclipse 攻击[25]和 Erebus 攻击 3 种检测方法组成；同时，为扩大各基模型输出差异度，引入 SVM、C4.5 和 Naive Bayes 这 3 种经典机器学习分类模型分别作为基分类器模型 ℓ_4、ℓ_5 和 ℓ_6。

2.1 实验评估指标

模型采用准确率（ACC, accuracy）、召回率（recall）、精准率（precision）、误判率（FAR, false alarm rate）、F_1-score、AUC（area under curve）和平均绝对误差（MAE, mean absolute error）等指标评价检测性能。

$$\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{37}$$

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{38}$$

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{39}$$

$$\mathrm{FAR}=\frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}} \tag{40}$$

$$F_1\text{-score}=2\times\frac{\mathrm{recall}\times\mathrm{precision}}{\mathrm{recall}+\mathrm{precision}} \tag{41}$$

$$\mathrm{AUC}=1-\frac{1}{m^+m^-}\cdot\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(F((f(x^+)<f(x^-))+\frac{1}{2}F((f(x^+)=f(x^-))\right) \tag{42}$$

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}\left|y_i-\widehat{y_i}\right| \tag{43}$$

其中，TP（true positive）表示被正确归类的攻击样本，FP（false positive）表示被错误归类的攻击样本，TN（true negative）表示被正确归类的正常样本，FN（false positive）表示被错误归类的正常样本；$\widehat{y_i}$ 为模型预测值，y_i 为实际值；AUC 为受试者工作特征曲线（ROC, receiver operating characteristic curve）下包面积，其纵坐标为真阳性率（TPR, true positive rate），横坐标为假阳性率（FPR, false positive rate）；m^+、m^- 为相应正例、反例统计数量；D^+、D^- 为相应正例集、反例集；$f(x^+)$ 为检测为正类样本概率，$F(x)$ 为指标函数，x 为真时 $F(x)$ 取值为 1。AUC 值越大说明模型检测性能越好。

$$\mathrm{FPR}=\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \tag{44}$$

$$\mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{45}$$

2.2 实验环境与样本数据

以 Xu 等[26]所提实现环境为基础，模拟真实场景下区块链网络层流量交互行为，其中网络流量环境包括正常区块链网络层流量环境和攻击流量环境。实验数据源自真实区块链系统运行过程中采集的流量，样本特征以文献[14]选择所设定的 28 维常规流量特征为基础，并通过 Wireshark 捕获真实 UDP 和 TCP 数据包集以形成.pcap 文件。在 Wireshark 中添加 Ethereum Devp2p 协议 Dissector 插件，用于解析以太坊数据包负载信息。采集的正常流量约 15.4 万条，攻击流量约 2.8 万条。

本文分别采用 Ethereum 2.0 交易流量和 Hyperledger Fabric1.4 运行流量来模拟区块链网络层正常交易行为流量。其中，Ethereum 2.0 环境由 5 台虚拟主机构成，在 Ubuntu 18.04 上运行 Geth 核心程序来进行区块链矿机间交易过程，其 IP 网段为 192.168.108.0/24；Hyperledger Fabric1.4 环境由 8 台虚拟主机构成，在 Ubuntu18.04 的 Docker 容器中运行区块链程序，IP 网段为 192.168.106.0/24。

为模拟真实条件下的攻击场景，使用 2 种方式生成攻击流量。1) 基于 Yersinia 攻击工具挖掘根据网络协议产生的漏洞，通过伪造特定的协议负载信息和数据包来实现欺骗网络邻居节点，并生成 Eclipse 和 Erebus 路由覆盖攻击流量以破坏网络路由拓扑结构。2) 基于 Python 3.9 编写基于 Scapy 库

的攻击脚本，其中包括周期性生成 ping、pong、findnode、neighbors 和 ADDR 构建等 Eclipse 攻击和 DDoS 攻击命令，攻击周期为 1 s。攻击脚本通过周期性持续执行上述命令来攻击受害子网节点和路由器以强制构建错误路由，屏蔽受害子网节点路由。攻击环境主机 IP 网段设置为 192.168.13.0/24。实验模拟的拓扑环境如图 5 所示。

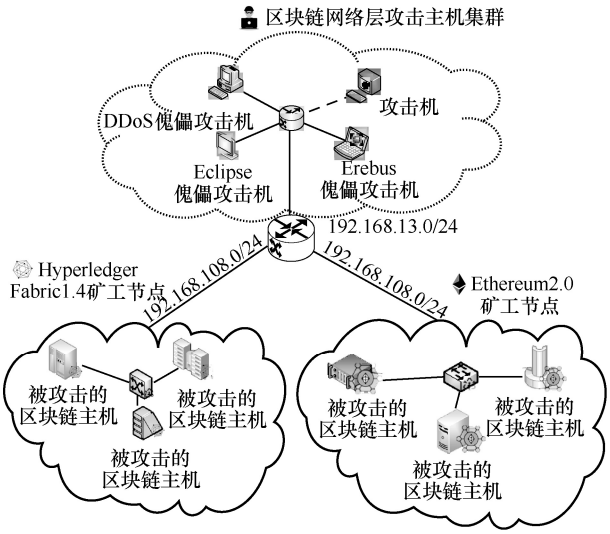


图 5　实验模拟的拓扑环境

2.3　实验参数选取和结果分析

模型训练过程基于 mini-batch 模式进行训练，对区块链网络层环境中采样流量进行预处理和标记后进行随机独立抽样形成多个新数据集，并分割为训练集和测试集。按模型结构对训练集进行模型训练，通过反向训练参数调优以获取最佳参数集合；并使用测试集验证模型性能。在批训练过程中，模型每训练 50 次进行一次样本迭代，每次迭代抽取 5 000 条流量样本量，其中正常流量和攻击流量分别占 80%和 20%。实验过程中对数据集进行多次独立重复实验，在测试集之间进行交叉验证，并取每个数据集的检测结果平均值，以提升实验结果的无偏见性。实验参数值选取如表 1 所示。

图 6 给出了在表 1 设定参数条件下不同训练次数下模型平均绝对误差。从图 6 可知，训练集和测试集在模型训练初期的 MAE 较高，随着训练次数的增加，MAE 快速下降，最终模型收敛并趋于稳定，说明模型可以在给定的数据集中完成模型收敛并形成较好的稳定输出模型。为验证训练样本中基于改进 Bagging 集成算法的有效性，图 7 给出了从原始数据集中随机抽取的 300 组样本使用所提检测方法的 LBoDF 输出值，并采用 *t*-SNE[27]方法将数据点分布降至 2 维空间进行可视化分析；图 8 给出了随机抽样样本的改进 *k*-距离邻域，以反映采样样本的空间密度分布。异常检测输出结果中正常流量样本多数呈聚类状，数据空间分布较均匀，相应离群因子值较小；异常流量样本点在采样集外部呈零星分布，与距离采样中心点的空间距离较大，导致其离群因子值较高。正常样本数据点的改进 *k*-距离邻域和空间密度较高，异常样本反之。这说明所提检测方法可在数据场定义下有效描述正常与异常流量样本间的空间分布差异，并可反映样本数据空间聚类分布特点。

表 1　　实验参数值选取

参数	参数说明	取值
τ	改进 Bagging 算法训练模型次数	1 815
η	空间距离加权倍数	8
ζ	离群因子阈值	1.17
N	输入样本集数量	5 000
M	基模型数量	6
δ	场模型中势能函数辐射因子	1.0
μ	组合向量维度	6
κ	模型输出样本类别	2
μ	组合向量维度	6
m	原始特征合集特征项数	126
m'	特征筛选后特征项数	65

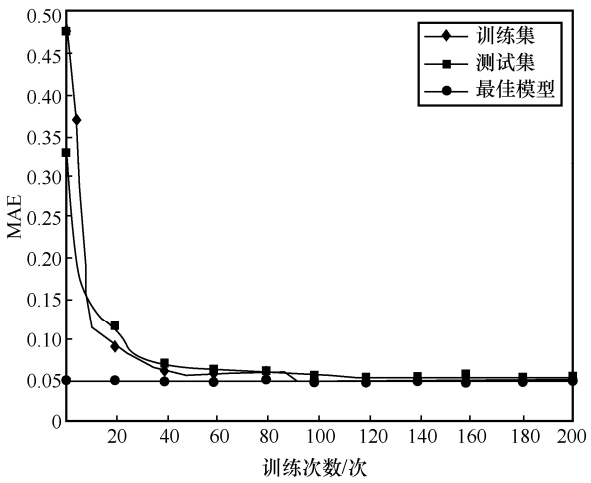


图 6　不同训练次数下模型平均绝对误差

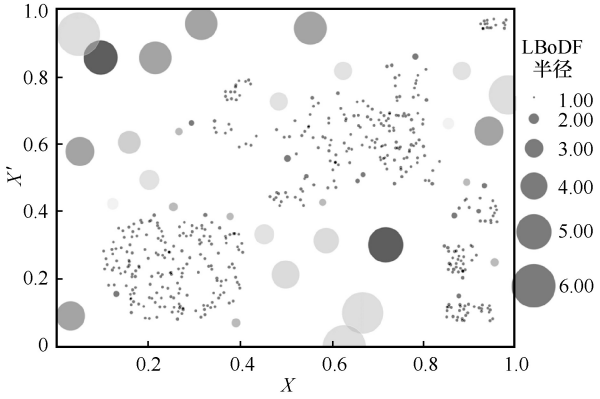


图 7　对随机抽样样本使用所提异常检测方法的 LBoDF 输出值

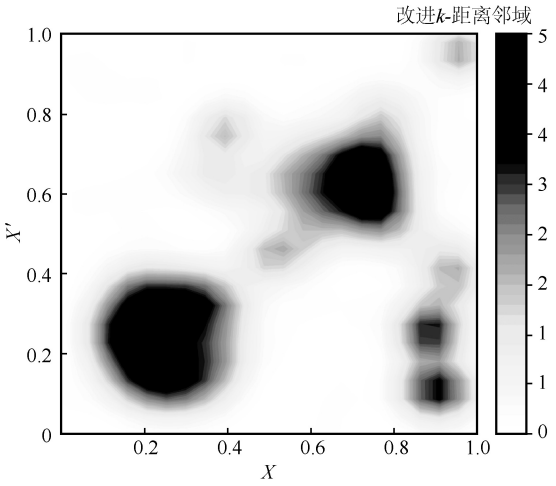


图 8　随机抽样样本的改进 k-距离邻域

图 9 给出了采样样本中不同改进离群因子阈值下的改进 k-邻域距离样本点分布。其中，正常数据点和异常数据点的 LBoDF 值均呈正态分布，流量样本的空间分布合理，当 LBoDF 值为 1.17 左右时，邻域值最大，说明样本数据在该参数设定下对正常和异常数据点区分度较好且置信度较高，因此将所提 LBoDF 算法的阈值 ζ 设置为 1.17。

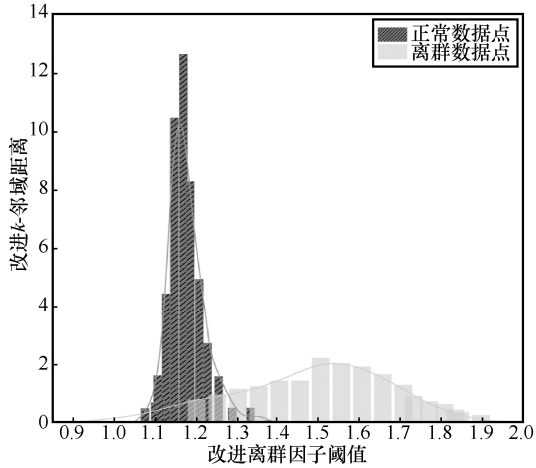
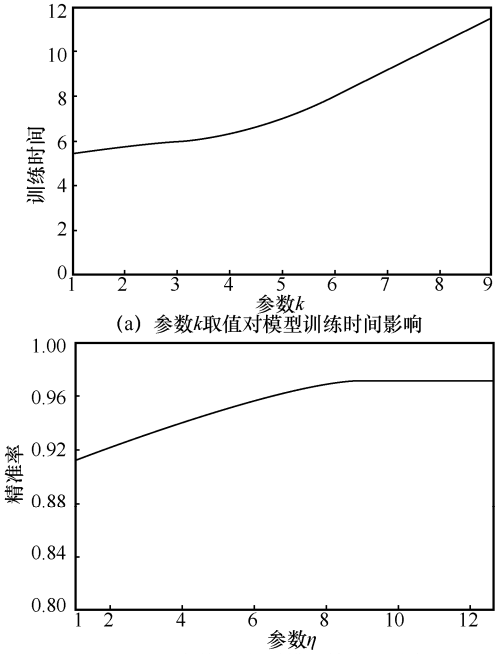


图9　采样样本中不同改进离群因子阈值下的改进 k-邻域距离样本点分布

图 10 给出了 k 和 η 取值对模型训练结果的影响。对于参数 k，当 $1<k<5$ 时，模型运行时间在 6 s 左右；当 $k>5$ 时，训练时间呈指数增长。对于参数 η，所提模型的精准率随着 η 增加相应增加；当 $\eta>9$ 时，模型精准率趋于稳定，说明 η 对模型的影响有极限，并且 $\eta=9$ 时精准率取最优值 0.968。



(a) 参数k取值对模型训练时间影响

(b) 参数η取值对模型精准率影响

图 10　k 和 η 取值对模型训练结果的影响

图 11 给出了不同特征选择算法对区块链网络层样本进行特征选择后，输出特征子集维度对检测模型精准率的影响。从图 11 中可知，当输出特征子集维度约为 10 维时，4 种算法精准度均超过 0.95，随着输出特征子集维度的增加，模型精准率会逐渐下降。因此，实验将 D&RFS 算法输出特征维度定为 10 维，此时精准率达到最大值 0.982。

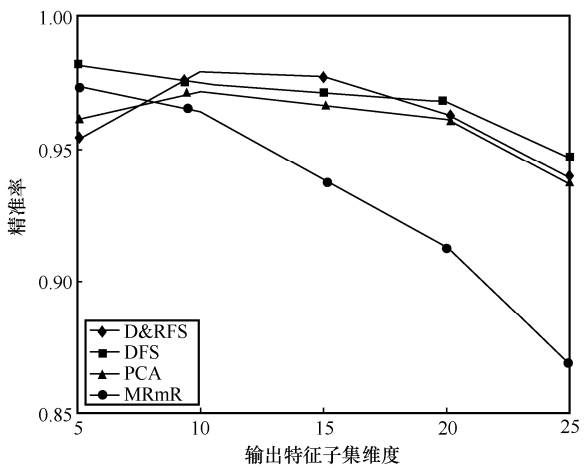
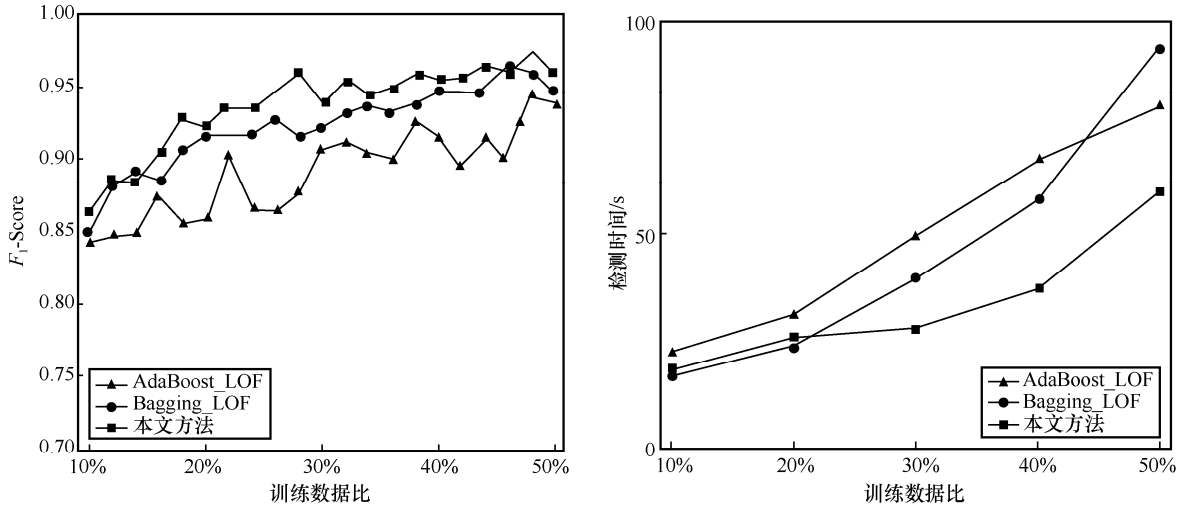


图 11　输出特征维度对模型检测精准率的影响

2.4　检测方法性能对比

2.4.1　与经典集成学习异常检测方法性能对比

为验证本文方法相较于现有经典集成学习异常检测方法的检测能力的优势，实验使用相同基模型对相同采样数据集进行异常检测。不失一般性地，选取基于 AdaBooost_LOF 和基于 Bagging_LOF 两类经典集成学习异常检测方法，检验不同检测方法在不同训练数据比情况下的 F_1-Score 性能和训练时间。如图 12(a)所示，3 种检测方法随着训练数据比增加而缓慢增加，当训练数据比达到 40%后，3 种方法的 F_1-Score 性能趋于稳定，且本文方法的 F_1-Score 优于其他 2 种方法（平均上升 3.01%），说明本文方法相较于经典机器学习算法有较高的综合性能优势。产生该优势的原因如下。本文方法基于不同基分类器对混合型攻击流量采用多角度的集成学习，对于混合型攻击流量的核心特征感知能力更强，因而生成的集成模型对混合型攻击流量样本的检测泛化能力提升；并且所提 LBoDF 相较于 LOF 算法对稠密向量有较强的异常检测区分能力，可有效感知正常与异常样本的核心区别。如图 12(b)所示，3 种检测方法的检测时间随训练数据比增加也逐渐增加，当训练数据比相对较少时，本文方法的检测时间相对较多；但随训练数据比上升，本文方法的检测时间相较于另 2 种经典机器学习方法较少。



(a) 3种异常模型F_1-Score性能对比　　(b) 3种异常模型所需检测时间

图 12　3 种检测方法在不同训练数据比的下 F_1-Score 性能和检测时间对比

2.4.2　与现有区块链网络层异常检测方法性能对比

1) 与单一分类检测器性能对比

为验证本文方法相较于单一类型区块链网络层攻击流量分类检测的性能提升情况，本节对前期所提 DDoS[24]、Eclipse[25]和 Erebus 这 3 种稳定输出的攻击检测模型进行横向性能对比。实验从区块链网络层流量数据集中抽取 2 000 组进行随机抽样，包含 50%正常流量、20%DDoS 攻击流量、10%Eclipse 攻击流量和 10%Erebus 攻击流量。图 13 绘制了 4 种检测方法的准确率、精准率、召回率和 F_1-Score 检测指标平均性能。

从图 13 中可知，本文方法的 F_1-Score 和召回率均高于其他 3 种检测方法，平均增加 1.44%和 2.11%，说明本文方法有较好的综合检测性能和较低的漏报率，并且集成检测方法相较于单一分类检测器对于多类型混合攻击样本有较高的异常检测敏感度，可有效检测攻击流量的存在。准确率方面，本文方法的性能最好，但优势稍弱，平均增加 1.57%；精准率方面，本文方法稍逊于 DDoS 和 Eclipse 方法，但仍超过 95.26%。本文方法由于强调针对混合攻击流量的异常检测泛化性，对特定种类攻击流量检测的特征感知能力稍有欠缺；但其在真实区块链网络层环境中仍具备较好的异常检测适应性，可有效从大规模流量样本中检测异常流量行为。综合图 13 结果可说明，本文方法相较于单一分类检测器已经有相当的性能上升，在保留各单一分类检测针对性的同时可提升集成模型的检测泛化能力，特别是在检测召回率和 F_1-Score 这 2 个指标方面反映了本文方法对异常流量样本的检测漏报情况较少，综合检测性能得到提升。

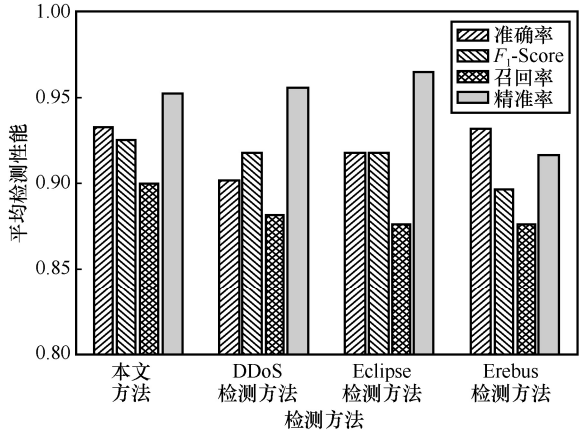


图 13　4 种检测方法性能对比

2) 与集成模型异常检测方法性能对比

为了检验与现有针对区块链网络层异常流量检测方法在不同标记样本比例下的性能对比结果，设置对照实验与文献[9]、文献[20]和文献[21]等所提区块链网络层综合型异常检测方法进行对比；同时与基于多 SVM_AdaBooost 和基于多 SVM_Bagging 的异常检测方法进行横向综合性能对比，检验准确率、精准率和召回率等指标性能。从数据集中随机选取部分样本数据，设定异常流量比例为 20%，并构建相应样本进行模型集成学习和训练，对比结果如图 14 所示。从图 14 中可知，本文方法在输入特征子集前进行了针对性采样，并把动态权重设置方法引入成员分类器集成环节中，从而实现有效区分正常流量样本和异常流量样本的同时提高检测性能。因此检测精准率、准确率和召回率在不同标记样本比例上都比对照方法要好（准确率平均提升 11.7%，精准率平均提升 10.9%，召回率平均提升 6.7%）。特别是当标记样本比例较低时（如 5%），本文方法准确率仍能超过 80%。

为验证本文方法与横向对照方法在不同时间开销下 F_1-Score 性能的变化情况，进行横向对比实验，实验结果如图 15 所示。从图 15 中可知，随着训练时间增加，6 种检测方法的 F_1-Score 性能逐渐增加，在时间开销达到 600 s 后均趋于稳定；并且本文方法使用较少时间即可获得较高 F_1-Score（结果超 90%），在同样时间开销下 F_1-Score 性能均优于对照方法，说明本文方法可以快速准确检测出区块链网络层异常流量。

为进一步验证本文方法相较于文献[9]和文献[21]方法的检测性能优势，通过计算 AUC 值来进行实验对比。如图 16 所示，本文方法的 AUC 值为 0.976，接近于 1，高于另外 2 种方法的 AUC 值（0.854 和 0.862），说明本文方法相比现有检测方法更具有综合性优势。

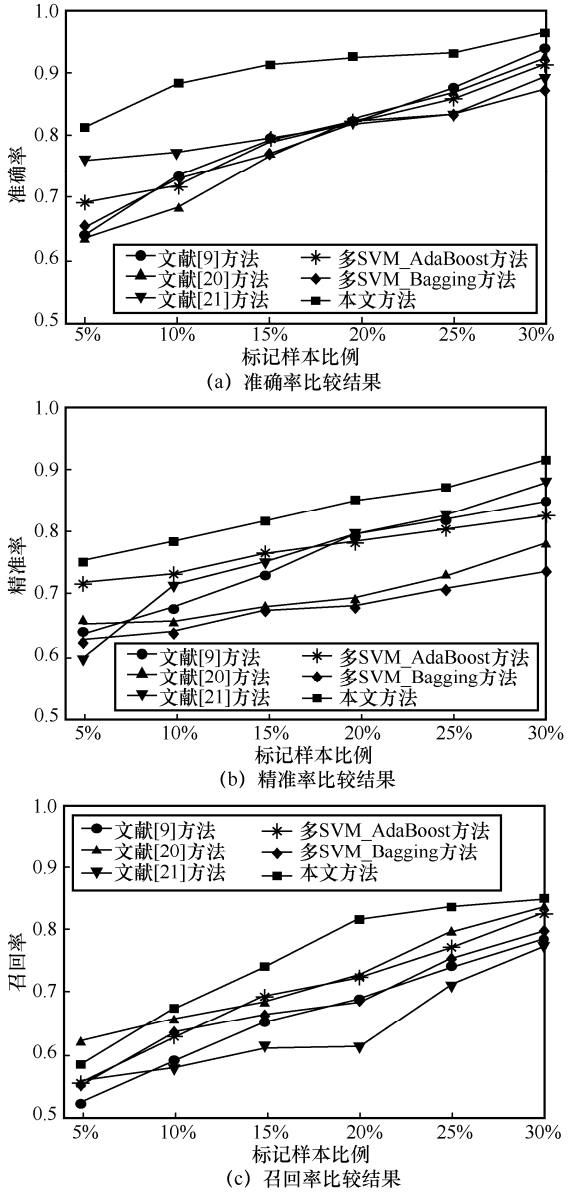


图 14　6 种检测方法在不同标记样本比例下的准确率、精准率和召回率检测性能对比

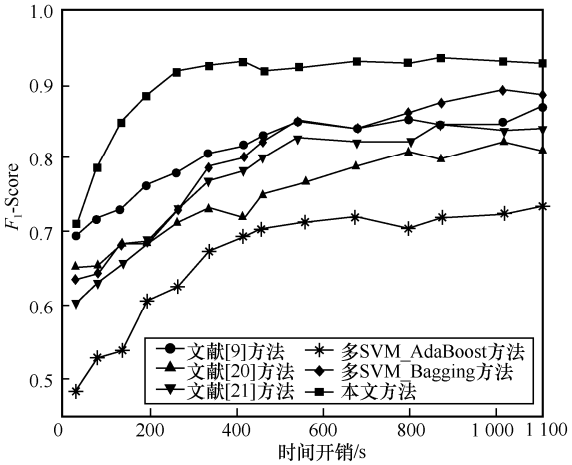


图 15　不同时间开销下的 6 种检测方法的 F_1-Score

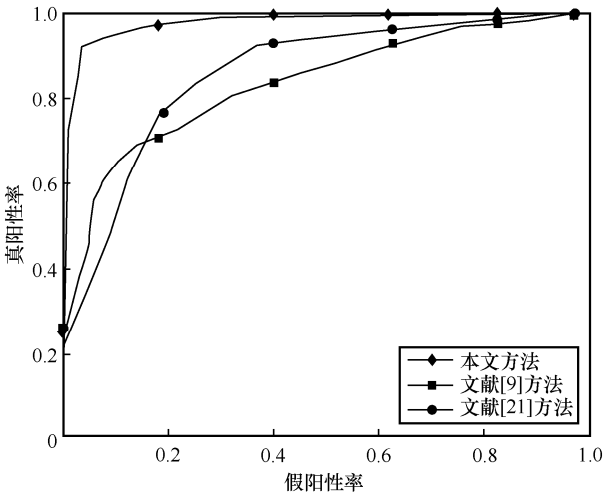


图 16　3 种检测方法的 ROC 曲线对比

综上，与现有区块链网络层异常检测方法和基于集成算法的 LOF 检测方法相比，本文方法能在包含攻击流量的区块链网络层环境中有效分离出正常和异常流量，且具有检测性能优势。

3　结束语

本文在深入研究区块链网络层异常流量检测相关方法的基础上，针对现有检测方法存在检测性能受限的问题，提出一种基于多分类器集成的区块链网络层异常流量检测方法。首先，基于 D&RFS 算法从区块链网络层流量中提取高区分度、低冗余的特征子集，提升各特征子集区分度和差异性，并减少计算复杂度和特征冗余度；其次，基于动态自适应加权的 Bagging 集成算法对多类型区块链网络层异常流量分类器模型进行强泛化集成，提升模型对混合型攻击流量的综合感知能力；最后，提出 LBoDF 异常检测方法，将集成算法输出低维向量进行高维空间映射，扩大正常数据点与异常数据点的空间分布以提升异常检测准确率。实验结果表明，本文方法对区块链网络层中的混合攻击流量进行异常流量检测，具有较好的检测性能。下一步，笔者将考虑基于动态自学习型非监督深度学习模型对区块链网络层异常流量进行检测，提高未知区块链网络层异常攻击流量的检测性能。

参考文献：

[1] 沈鑫, 裴庆祺, 刘雪峰. 区块链技术综述[J]. 网络与信息安全学报, 2016, 2(11): 11-20.
SHEN X, PEI Q Q, LIU X F. Survey of block chain[J]. Chinese Journal of Network and Information Security, 2016, 2(11): 11-20.

[2] STEPHEN R, ALEX A. A review on blockchain security[J]. IOP Conference Series: Materials Science and Engineering, 2018: doi.org/10.1088/1757-899x/396/1/012030.

[3] 韩璇, 袁勇, 王飞跃. 区块链安全问题: 研究现状与展望[J]. 自动化学报, 2019, 45(1): 206-225.
HAN X, YUAN Y, WANG F Y. Security problems on blockchain: the state of the art and future trends[J]. Acta Automatica Sinica, 2019, 45(1): 206-225.

[4] 江沛佩, 王骞, 陈艳姣, 等. 区块链网络安全保障：攻击与防御[J]. 通信学报, 2021, 42(1): 151-162.
JIANG P P, WANG Q, CHEN Y J, et al. Securing guarantee of the blockchain network: attacks and countermeasures[J]. Journal on Communications, 2021, 42(1): 151-162.

[5] TRAMÈR F, BONEH D, PATERSON K G. Remote side-channel attacks on anonymous transactions[C]//Proceedings of the 29th USENIX Conference on Security Symposium. New York: ACM Press, 2020: 2739-2756.

[6] 叶聪聪, 李国强, 蔡鸿明, 等. 区块链的安全检测模型[J]. 软件学报, 2018, 29(5): 1348-1359.
YE C C, LI G Q, CAI H M, et al. Security detection model of blockchain[J]. Journal of Software, 2018, 29(5): 1348-1359.

[7] 曾诗钦, 霍如, 黄韬, 等. 区块链技术研究综述：原理、进展与应用[J]. 通信学报, 2020, 41(1): 134-151.
ZENG S Q, HUO R, HUANG T, et al. Survey of blockchain: principle, progress and application[J]. Journal on Communications, 2020, 41(1): 134-151.

[8] HASSAN M U, REHMANI M H, CHEN J. Anomaly detection in blockchain networks: a comprehensive survey[J]. arXiv Preprint, arXiv: 2112.06089, 2021.

[9] KIM J, NAKASHIMA M, FAN W J, et al. Anomaly detection based on traffic monitoring for secure blockchain networking[C]// Proceedings of 2021 IEEE International Conference on Blockchain and Cryptocurrency (ICBC). Piscataway: IEEE Press, 2021: 1-9.

[10] SYARIF I, ZALUSKA E, PRUGEL-BENNETT A, et al. Application of bagging, boosting and stacking to intrusion detection[C]// Proceedings of International Workshop on Machine Learning and Data Mining in Pattern Recognition. Berlin: Springer, 2012: 593-602.

[11] 凌玥, 刘玉岭, 姜波, 等. 基于双层异质集成学习器的入侵检测方法[J]. 信息安全学报, 2021, 6(3): 16-28.
LING Y, LIU Y L, JIANG B, et al. Intrusion detection method based on double-layer heterogeneous ensemble learner[J]. Journal of Cyber Security, 2021, 6(3): 16-28.

[12] GAIKWAD D P, THOOL R C. Intrusion detection system using bagging ensemble method of machine learning[C]//Proceedings of 2015 International Conference on Computing Communication Control and Automation. Piscataway: IEEE Press, 2015: 291-295.

[13] DONG X B, YU Z W, CAO W M, et al. A survey on ensemble learning[J]. Frontiers of Computer Science, 2020, 14(2): 241-258.

[14] 杨晓晖, 张圣昌. 基于多粒度级联孤立森林算法的异常检测模型[J]. 通信学报, 2019, 40(8): 133-142.
YANG X H, ZHANG S C. Anomaly detection model based on multi-grained cascade isolation forest algorithm[J]. Journal on Communications, 2019, 40(8): 133-142.

[15] 黄金超, 马颖华, 齐开悦, 等. 一种基于集成学习的入侵检测算法[J]. 上海交通大学学报, 2018, 52(10): 1382-1387.
HUANG J C, MA Y H, QI K Y, et al. An ensemble-based intrusion

detection algorithm[J]. Journal of Shanghai Jiao Tong University, 2018, 52(10): 1382-1387.

[16] 李小剑, 谢晓尧, 徐洋. 网络流量异常检测方法: SSAE-IWELM-AdaBoost[J]. 武汉大学学报(理学版), 2020, 66(2): 126-134.
LI X J, XIE X Y, XU Y. Network traffic anomaly detection method: SSAE-IWELM-AdaBoost[J]. Journal of Wuhan University (Natural Science Edition), 2020, 66(2): 126-134.

[17] ZHANG J Q, WANG Z Z, MENG J J, et al. Boosting positive and unlabeled learning for anomaly detection with multi-features[J]. IEEE Transactions on Multimedia, 2019, 21(5): 1332-1344.

[18] 刘金平, 何捷舟, 马天雨, 等. 基于 KELM 选择性集成的复杂网络环境入侵检测[J]. 电子学报, 2019, 47(5): 1070-1078.
LIU J P, HE J Z, MA T Y, et al. Selective ensemble of KELM-based complex network intrusion detection[J]. Acta Electronica Sinica, 2019, 47(5): 1070-1078.

[19] REDDY D K K, BEHERA H S, PRATYUSHA G M S, et al. Ensemble bagging approach for IoT sensor based anomaly detection[C]// Proceedings of the Intelligent Computing in Control and Communication. Berlin: Springer, 2021: 647-665.

[20] OFORI-BOATENG D, DOMINGUEZ I S, AKCORA C, et al. Topological anomaly detection in dynamic multilayer blockchain networks[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin: Springer, 2021: 788-804.

[21] ZHANG R, ZHANG G F, LIU L, et al. Anomaly detection in bitcoin information networks with multi-constrained meta path[J]. Journal of Systems Architecture, 2020: doi.org/10.1016/j.sysarc.2020.101829.

[22] JOHNSON R, ZHANG T. Accelerating stochastic gradient descent using predictive variance reduction[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. New York: ACM Press, 2013: 315-323.

[23] 钱景辉, 梁栋. 一种基于多标记的局部离群点检测算法[J]. 微电子学与计算机, 2017, 34(10): 110-114.
QIAN J H, LIANG D. Local outlier detection algorithm based on multi-label learning[J]. Microelectronics & Computer, 2017, 34(10): 110-114.

[24] DAI Q Y, ZHANG B, DONG S Q. A DDoS-attack detection method oriented to the blockchain network layer[J]. Security and Communication Networks, 2022, 2022: 1-18.

[25] DAI Q Y, ZHANG B, DONG S Q. Eclipse attack detection for blockchain network layer based on deep feature extraction[J]. Wireless Communications and Mobile Computing, 2022, 2022: 1-19.

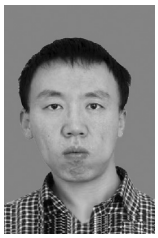
[26] XU G, GUO B, SU C, et al. Am I eclipsed? a smart detector of eclipse attacks for Ethereum[J]. Computers & Security, 2020, 88: 101604.

[27] LAURENS V D M, HINTON G. Viualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(2605): 2579-2605.

[作者简介]

戴千一（1994− ），男，陕西西安人，信息工程大学博士生，主要研究方向为区块链安全、区块链网络层流量检测、区块链系统应用、机器学习。

张斌（1969− ），男，河南南阳人，博士，信息工程大学教授、博士生导师，主要研究方向为信息系统安全。



郭松（1985− ），男，河北保定人，博士，信息工程大学讲师，主要研究方向为信息系统安全。

徐开勇（1963− ），男，河南信阳人，信息工程大学研究员，主要研究方向为信息安全与可信计算。