基于模型堆叠的以太坊钓鱼诈骗账户识别方法

陈伟利[1], 叶明顺[1], 唐明董[1†], 郑子彬[2]

(1. 广东外语外贸大学 信息科学与技术学院, 广东 广州 510006; 2. 中山大学 软件工程学院, 广东 珠海 528478)

摘要: 近年来, 钓鱼诈骗已成为区块链平台中不可忽视的欺诈类型, 对用户金融安全构成了重大威胁. 为了解决这一问题, 本文提出了一种基于区块链交易的网络钓鱼账户检测框架, 并以以太坊为例验证了其有效性. 具体而言, 该框架通过引入数据样本过滤规则来缓解数据不均衡性以及减少计算量, 采用级联特征抽取方法以提取有效特征, 并基于模型堆叠构建集成分类算法建立模型以识别以太坊上的钓鱼诈骗账户. 实验结果表明, 该框架能够有效地识别以太坊上的钓鱼诈骗账户, 具有一定的实际应用价值.

关键词: 区块链; 以太坊; 钓鱼诈骗; 模型堆叠

引用格式: 陈伟利, 叶明顺, 唐明董, 等. 基于模型堆叠的以太坊钓鱼诈骗账户识别方法. 控制理论与应用, 2024, 41(8): 1361 – 1368

DOI: 10.7641/CTA.2023.21035

Ethereum phishing scam account identification based on model stacking

CHEN Wei-li[1], YE Ming-shun[1], TANG Ming-dong[1†], ZHENG Zi-bin[2]

(1. School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou Guangdong 510006, China; 2. School of Software Engineering, Sun Yat-sen University, Zhuhai Guangdong 528478, China)

Abstract: In recent years, phishing scams have become a type of fraud that cannot be ignored in blockchain platforms, posing a major threat to users' financial security. To solve this problem, this paper proposes a framework for phishing account detection based on blockchain transactions, and verifies its effectiveness by taking ethereum as an example. Specifically, the framework alleviates data imbalances and reduces computational effort by introducing sample filtering rules, adopts a cascading feature extraction method to extract valid features, and builds an ensemble classification algorithm based on model stacking to identify phishing accounts. The experimental results show that the framework can effectively identify phishing fraud accounts on ethereum and has certain practical application value.

Key words: blockchain; ethereum; phishing scam; model stacking

Citation: CHEN Weili, YE Mingshun, TANG Mingdong, et al. Ethereum phishing scam account identification based on model stacking. *Control Theory & Applications*, 2024, 41(8): 1361 – 1368

1 引言

比特币的出现开启了一个全新的加密货币时代. 据加密货币市场网站coinmarketcap.com提供的数据显示, 目前已经存在超过21,790种加密货币(或token), 其总市值超过8,000亿美元. 这些加密货币的核心技术是区块链技术. 一般来说, 区块链是指由点对点网络通过特殊的共识机制来维护的分布式可信数据库[1]. 区块链技术通常被用来实现加密货币或虚拟货币, 这种货币可以与其他加密货币或法定货币在交易所进行交换. 由于加密货币具有金融属性, 因此, 它们成了许多骗局的目标.

金融数据安全是区块链技术健康发展的重要基础. 在区块链生态系统中, 诈骗的扩散将阻碍用户对区块链技术的接受和使用, 从而阻碍该技术的进步. 因此, 这些诈骗的识别已成为区块链生态系统中一个紧迫而关键的问题, 并引起了研究者的极大关注[2–3]. 网络钓鱼诈骗是一种新型的网络犯罪, 随着网络业务的兴起而出现[4], 现在已经在区块链生态系统中被发现.

收稿日期: 2022−11−26; 录用日期: 2023−07−12.
†通信作者. E-mail: mdtang@126.com; Tel.: +86 13342807181.
本文责任编委: 崔巍.
国家重点研发计划项目(2020YFB1006002), 国家自然科学基金面上项目(61976061), 广东省基础与应用基础研究基金项目(2021A1515011939)资助.
Supported by the National Key Research and Development Program of China (2020YFB1006002), the National Natural Science Foundation of China (61976061) and the Basic and Applied Basic Research Foundation of Guangdong Province (2021A1515011939).

根据Chainalysis的报告, 自2017年以来, 超过50%的网络犯罪收入来自于网络钓鱼诈骗[1]. 一个众所周知的例子是蜜蜂令牌(bee token) ICO (initial coin offering)[2] 上的网络钓鱼骗局, 该网络钓鱼者最终在短短25 h内就从投资者那里骗取了约100万美元. 2022年7月, Uniswap遭遇假令牌网络钓鱼攻击, 被盗超过470万美元. 这些例子表明, 检测和预防网络钓鱼诈骗是区块链生态系统中一个紧迫的问题.

传统的网络钓鱼诈骗是指通过电子邮件、短信或社交媒体等途径, 冒充合法机构或个人的身份, 发送网站链接, 欺骗受害人提供个人敏感信息或转账汇款等行为, 以达到骗取财物或者个人信息的目的. 因此, 传统的网络钓鱼诈骗检测方法的主要目标是通过各种方式识别虚假网站, 提前警示用户在登录前识别并避免受到诈骗攻击. 然而, 网络钓鱼诈骗者在区块链中的手段和传统钓鱼诈骗有所不同. 诈骗者会创建虚假的以太坊钱包网站, 并利用虚假社交媒体帐户和电子邮件地址宣传和推广这个网站. 一旦用户相信了诈骗者的欺诈性消息, 他们会被引导到虚假的以太坊钱包网站, 并被要求创建一个新的钱包. 在这个过程中, 用户会被要求提供自己的私钥等敏感信息, 这些信息将被用来控制用户的钱包. 如果用户提供了这些信息, 诈骗者就可以控制用户的钱包, 并将其中的资金转移走. 一些诈骗者可能会创建虚假的智能合约来吸引用户参与, 但这些虚假合约可能无法执行或被用来窃取用户的资金. 诈骗者还可以通过不同的方式从以太坊网络中转移资金, 包括利用虚假合约、攻击交易所或其他以太坊应用程序. 因此, 区块链中的网络钓鱼诈骗需要采取不同的预防措施来保护用户的资产安全.

基于上述特点, 本文提出了一种新的网络钓鱼诈骗检测方法, 该方法利用区块链交易数据和机器学习算法来识别潜在的网络钓鱼诈骗. 这些检测方法可以集成到用户的加密货币钱包中, 作为区块链生态系统中管理账户和交易的工具, 为用户提供实时的风险提示和预警, 保护用户的资产安全. 然而, 解决区块链网络钓鱼诈骗的问题面临3个重要挑战: 1) 样本极度不均衡. 在区块链生态系统中, 钓鱼诈骗账户相对于正常账户来说非常少, 这导致很难建立有效的机器学习模型; 2) 账户特征缺乏. 为了构建有效的机器学习识别模型, 需要良好的特征. 然而, 在区块链上的账户由于匿名特性, 除了交易记录外, 基本没有额外的账户信息; 3) 计算量巨大. 每个账户自身大量的交易数据, 导致在进行特征提取时计算量巨大. 为了解决这些挑战, 本文提出了一个基于模型堆叠的网络钓鱼诈骗账户识别框架, 并以以太坊为例来验证方法的有效性. 图1展示了本文提出的框架图, 主要分为数据收集和处理阶段、特征工程和模型, 以及模型评价3个部分. 具体来说, 首先, 使用以太坊客户端Parity[3]下载以太坊交易信息. 然后, 通过爬取etherscan.io[4]获取所有网络钓鱼账户标签. 接着, 在数据分析的基础上, 引入了几种过滤规则来减少计算量以及缓解正负样本不平衡性. 在此基础上, 构造交易图, 并使用了一种基于交易图的级联特征提取方法. 最后, 本文提出了一个双层堆叠集成学习算法来识别以太坊上钓鱼诈骗账户. 通过与其他方法的比较, 本文评估了模型在不同数据集下的性能, 并验证了模型的有效性.

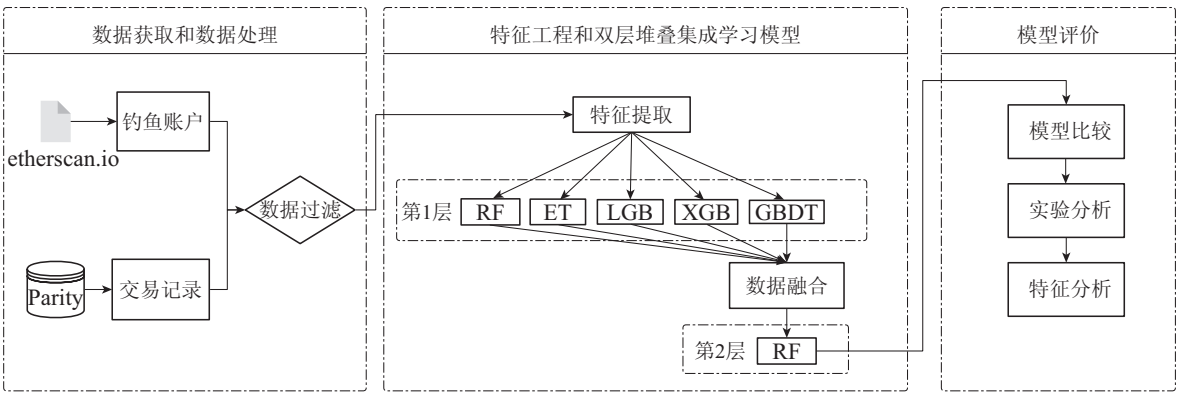

图 1 框架图

Fig. 1 The framework

本文是会议版论文[5]的扩展版. 相对于会议版本, 本文的主要差别如下: 1) 会议版本仅使用了0∼700万区块的交易记录和1,683个被标记为钓鱼诈骗的地址样本数据, 相对较少; 而本文将数据扩充到了0∼1,200万区块的交易记录, 钓鱼诈骗样本数达到了4,921个, 以便于训练出更好的模型; 2) 为了减轻扩充数据集带

[1]https://blog.chainalysis.com/the-rise-of-cybercrime-onethereum/.

[2]https://theripplecryptocurrency.com/bee-token-scam/.

[3]www.parity.io/ethereum/.

[4]https://etherscan.io/.

来巨大的运算量, 以及样本极度不平衡的问题, 本文新增了一种过滤规则. 实验结果表明, 这种方法的引入极大降低了计算量, 以及有效缓解了样本不均衡性, 同时, 保持了模型的识别效果; 3) 本文提出了一个新的模型–双层堆叠集成学习模型. 实验表明, 该模型在许多指标上取得了更好的结果.

2　背景与相关工作

以太坊是一种著名的基于区块链的平台, 它提供了一种图灵完整的语言来支持智能合约[6]. 智能合约是指在区块链上以数字形式存在的合约, 只有在满足预设条件时才能自动执行. 由于智能合约具有自动执行和不可逆性的特性, 因此可以用于实现一些经济功能. 例如, 在以太坊平台上使用智能合约创建新的数字代币[7]非常容易. 以太坊平台中的代币可以代表任何可替代的可交易商品. 为了以标准的方式实现一些基本特性, 代币与以太坊钱包和其他使用相同标准的代币兼容, 这意味着它可以方便地与以太币(以太坊的加密货币)和其他代币交换. 上述金融特性使以太坊不仅成为金融应用的平台, 而且成为网络罪犯的目标. 以太坊的第1个被广泛报道的网络犯罪事件是“DAO攻击”. “DAO(distributed autonomous organization)”是一个去中心化的自治组织, 旨在通过智能合约提供一种新的去中心化的商业模式. 为了这个组织的发展, 在成立时通过ICO筹集资金. 2016年6月, “DAO”智能合约中的一个漏洞被利用, 进一步导致了6,000万美元的损失. 这次事件最终导致了以太坊区块链的硬分叉. 以太坊的另一种严重犯罪形式是网络钓鱼诈骗, 它给普通投资者造成了很大损失. 因此, 在以太坊上开展网络钓鱼诈骗检测变得尤为重要.

近年来, 针对以太坊上的网络钓鱼诈骗, 主要出现了两类检测方法. 一类是利用传统的机器学习算法, 在会议版本[5]中重点关注统计特征(节点特征), 并通过特定的特征工程提取节点的入度、出度、最大交易值等219维的特征. 这些特征被用来识别钓鱼账户, 其中一个基于LightGBM(light gradient boosting machine)的双降采样集成机器学习算法用于分类. 另一类方法则从丰富的特征信息角度出发, 关注图嵌入(图表示学习)方法, 如DeepWalk[8], Node2Vec[9]和图卷积网络(graph convolutional networks, GCN)[10]来学习交易网络的结构信息. 其中在文献[11]中, 在基于Node2Vec的基础上提出Trans2Vec. Trans2Vec的采样过程不是随机的, 而是基于两个节点最后一次交易的偏置采样过程, 更适合于以太坊上的钓鱼检测. 在文献[12]中, 提出了一种基于图卷积网络和自动编码器的方法识别钓鱼诈骗地址, 该研究首次使用GCN学习交易网络的结构特征. 后来为了加入时间戳等信息, 在文献[13]中, 提出了一种时间交易聚合图网络方法丰富特征信息. 通过时间交易聚合图网络方法建模节点之间的历史交易记录的时间关系, 使得构建的图边带有时间戳信息. 最近, 一些基于深度学习的图神经网络方法被用来学习账户行为的深层信息, 构建更丰富的特征集, 并建立分类模型以检测钓鱼诈骗. 然而, 对于以太坊交易数据这种自然抽象的复杂图形, 标签稀缺和大量交易数据使得图神经网络(graph neural networks, GNN)方法难以应用. 为了解决这些问题, 文献[14]提出了一种自我监督的深度图学习模型(self-supervised incrEmental deep graph lEarning, SIEGE), 重点关注空间和时间两个角度, 并从大量未被标记的交易数据中学习有效的节点嵌入. 不同于这些方法, 本文的目标是针对整个区块链生态系统, 为用户提供针对网络钓鱼诈骗的早期预警.

在后续的实验部分, 本文选择了一些具有代表性的检测方法进行比较, 并进一步强调了本文提出的模型的有效性.

3　特征提取方法和模型

3.1　特征提取

由于交易记录是可用的唯一信息, 并且它们赋予了账户一个自然的图形结构, 为了提取有效的特征, 本文首先基于这些交易记录构建了一个交易图(transaction graph, TG). 具体而言, TG $= (V, E)$, 其中V是一组节点(数据集中的所有地址), 而$E = (v_i, v_j)\{v_i \in V, v_j \in V\}$是一组有序的边, 每条边表示一个地址$v_i$将一定数量的以太币交易到另一个地址$v_j$. 每条边都有两个属性: 区块编号和交易数量, 表示该边出现的时间和交易的数量. 在TG中的两个节点之间可能有多条边, 这取决于两个相关账户之间的交易数量. 此外, 账户与邻居账户(N阶邻居)的节点数据存在冗余和耦合, 本文采取特征选择等方法, 确保数据的可解释性和减少冗余信息. 基于图的特征在许多识别问题中已被证明非常有效[15–16]. 因此, 在会议版论文[5]中, 提出了一种基于交易图的级联特征提取方法, 如图2所示, 用于识别钓鱼账户. 具体方法如下: 首先将以太坊账户之间的交易视为邻居关系. 为了判断账户的类别, 不仅可以使用账户自身的信息, 还可以利用其邻居以及邻居的邻居的信息.

为了更清楚地解释, 下面给出了4个与节点相关的关键定义: 1) 节点数据: 节点数据是指该节点的交易历史记录,每个交易都包含有关交易的时间、方向和金额等信息. 交易时间以区块编号为记录, 即一个递增的整数. 每个交易有两个方向: 转出和转入, 其中一个账户的转出交易是将以太币转移到其他账户, 而一个账户的转入交易则是从其他账户接收以太币; 2) 节点特征: 节点特征是从节点数据中提取的各种信息. 本文通过各种统计方法来提取信息, 以生成节点的特

征描述; 3) N阶邻居: 一个节点的一阶邻居是直接连接到该节点的节点(两个节点之间有交易). 而一个节点的N阶邻居是指需要通过至少$N-1$个节点联结才能到达的节点; 4) N阶特征: 一个节点的0阶特征是指该节点的节点特征. 而从N阶邻居中级联提取的特征则被称为N阶特征, 用于更全面地描述节点的属性.

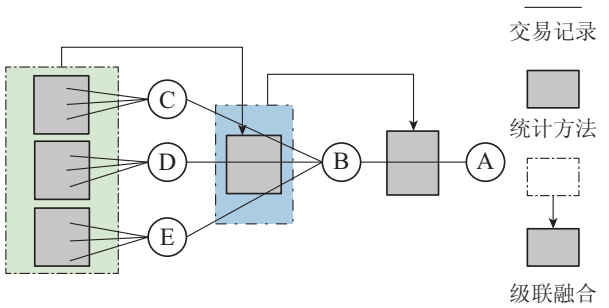

图 2 二阶特征提取过程

Fig. 2 2-order feature extraction procedure

3.1.1 节点特征

节点特征是节点数据的统计特征, 包括两种类型的数据: 交易金额和交易时间(区块编号). 根据交易性质的不同, 本文将交易分为转入和转出交易, 并使用direction_type_method来区分不同的特征. 例如, in_block_std表示所有转入交易时间的标准差. 对于交易时间, 本文只计算时间跨度(peak time peak, ptp)和标准差(standard deviation, std)两个特征; 对于交易金额, 本文计算了总和(sum)、最大值(maximum)、最小值(minimum)、平均值(mean)和标准差(std). 此外, 本文还提取了交易数量(count)、交易账户数量(unique)、两者的比率(unique_ratio)以及两者是否相等(unique_equal), 共计22个特征.

3.1.2 N阶特征

与在会议版本论文[5]中提取N阶特征的方法不同, 本文只提取节点交易金额这一类型的一阶网络特征. 如前所述, 交易的方向对于识别网络钓鱼诈骗非常重要. 因此, 本文将节点的一阶邻居分为from邻居和to邻居. 简单来说, 当交易是从节点a转出以太币到节点b时, 节点a是节点b的from邻居, 节点b是节点a的to邻居. 具体来说, 将一阶网络特征命名为neighbor_direction_statistic2_statistic1. 例如, 对于from_in_mean_max, 本文首先计算节点每个from邻居的转入交易金额的最大值. 然后, 计算节点的交易金额的平均值. 类似地, 对于to_out_std_sum, 首先, 计算节点每个to邻居的转出交易金额的总和. 然后, 计算节点交易金额的标准差. 最终, 总共得到40个特征.

3.2 基于lightGBM的双降采样集成算法

为了构建一个有效的识别模型, 在会议版论文[5]中提出了基于lightGBM的双采样集成算法(light-GBM-based dual-sampling ensemble). 如表1所示的算法提供了该算法的伪代码. 这个算法将作为本文的基线模型. 其背后的思想很简单. 类似于EasyEnsemble[17], 通过对大多数类别样本的抽样来减少类别不平衡问题. 不同之处在于, 由于该论文采用级联特征提取的方法获取大量特征, 在数据集中也采样了这些样本的特征. 这种双采样方法使得基模型具有更好的异质性, 从而保证了模型的有效性.

表 1 双采样集成算法

Table 1 Dual-sampling ensemble algorithm

输入: 少数类样本集P, 多数类样本集N, $|P| \ll |N|$, 基模型的数量k, 特征样本比例r, 特征的数量d, 选取模型最佳参数

输出: 集成的结果.

1: $i \leftarrow 0$;
2: **while** $i < k$ **do**
3: $i \leftarrow i+1$;
4: 随机抽样集合N的子集N_i, $|N_i| = \left\lfloor \frac{N}{K} \right\rfloor$;
5: 使用$P \cup N_i$训练基模型h_i, 其中特征维度为d, 特征数占比为r, 参数是最优参数;
6: **end while**
7: **return** $H(x) = \frac{1}{K}\sum_{i=1}^{T} h_i(x)$

3.3 双层堆叠集成学习算法

集成学习是一种可以用于处理数据不平衡问题的方法. 文献[18]中将模型堆叠用于网页钓鱼检测的研究. 本文利用这个思想, 提出一种双层堆叠集成学习算法, 用于以太坊上钓鱼诈骗账户的识别. 如图3所示, 堆叠集成学习模型包括两个阶段, 即训练阶段和预测阶段. 在训练阶段中, 可以包含多个层, 每个层都由一定数量的基模型组成.在堆叠集成学习中, 同一层的学习器是并行运行的, 而层与层之间的学习器是按顺序运行的, 上一层的输出作为下一层的输入. 具体来说, 在第1层有n个基学习器$e1, e2, \cdots, en$, 将数据分别输入到每一个基学习器中, 将得到的预测结果融合, 作为下一层的输入数据, 以此类推, 可以训练N层. 在预测阶段, 使用元学习层将训练阶段的最终预测结果融合, 并将其输入到元学习层的模型中进行训练, 从而得到最终的预测结果.

本文提出的双层堆叠集成学习模型(dual-layer stacked ensemble learning, DSEL) 的伪代码见表 2. 在选择第1层的基模型时需要考虑3个问题: 首先, 需要选择性能较好的模型作为基模型, 而在元学习器中可以选择一个简单的分类器, 以防止过拟合. 其次, 选取的基模型的个数不能太少, 因为第1层模型的个数等于元学习器的特征维度. 最后, 基模型一定不能选取性能表现很差的模型, 否则会影响整个模型融合的效果. 因此, 本文在第1层最终选择了随机森林(random

forest)[19]、极端随机树 (extra trees)[20]、梯度提升树 (gradient boosting decision tree)[21]、极端梯度提升(e-Xtreme gradient boosting, XGBoost)[22]和LightGBM[23]这5个模型, 并将随机森林作为元学习层的模型. 选择这些分类器的主要原因是由于在文献中存在广泛的应用, 而且这些分类器在分类数据时提供了不同的工作.

4　数据采集和处理

本章包括两个部分, 分别是数据采集和数据处理. 通过数据采集, 本文将数据集扩充到最新, 以便于训练出更好的模型. 与其他研究不同, 本文采用整个样本的数据进行实验, 而不是选取一些异常样本, 然后按比例抽取一部分正常样本作为数据集. 在数据处理部分, 为了应对数据集扩充所带来的巨大计算量和钓鱼账户与正常账户不平衡的问题, 提出了一种过滤规则. 这种方法极大地降低了计算量, 同时有效缓解了样本不均衡性. 引入这种方法有助于提高模型的性能和效率.

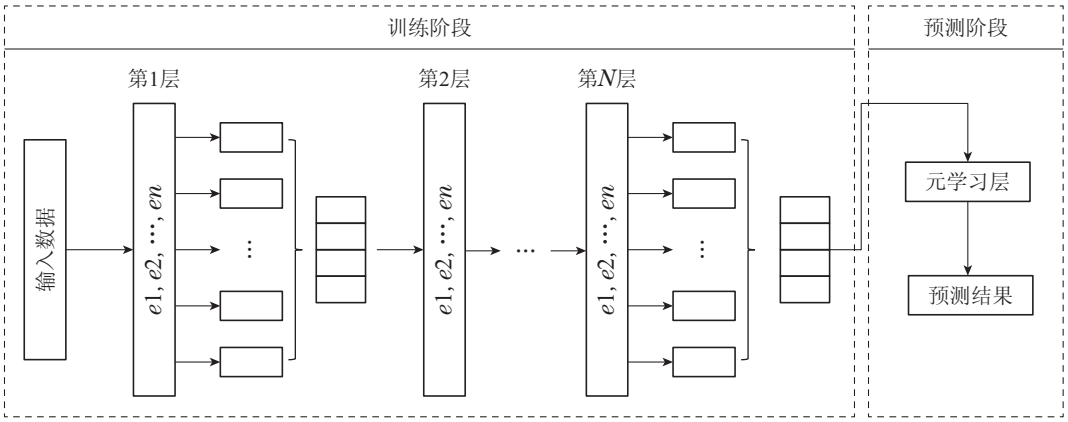

图 3　堆叠模型结构

Fig. 3　The structure of stacking model

表 2　双层堆叠集成学习算法

Table 1　Dual-layer stacked ensemble algorithm

输入: 少数类样本集P, 多数类样本集N, $|P| \ll |N|$, 第1层选取的基模型的数量n, 元学习层模型为H_l, 特征的数量d, 选取模型最佳参数

输出: 最终预测的结果

1: $i \leftarrow 0$;
2: **while** $i < k$ **do**
3: 　$i \leftarrow i + 1$;
4: 　使用$P \cup N_i$训练基模型h_i, 其中特征维度为d, 参数是最优参数;
5: **end while**
6: **return** $y = H_l(\sum_{i=1}^{T} h_i(x))$

4.1　数据集

为了检测钓鱼诈骗, 需要足够的样本来建立一个分类模型. 本文使用以太坊客户端Parity来同步以太坊上的交易信息, 并获得了2021年5月8日之前的所有以太坊块明细, 从区块高度0到12,000,000区块高度. 通过对所获得的交易情况进行分析, 得到了43,783,194个账户, 其中包括由智能合约控制的1,564,580个账户. 为了建立钓鱼诈骗识别模型, 寻找带有钓鱼诈骗标签的足够数量的地址是其中一项重要任务. 以太坊地址标签信息可在Etherscan.io网站上找到, 本文通过爬取该网站获得了所有被标记为网络钓鱼的地址. 这些地址经过验证被用于网络钓鱼诈骗. 因此, 本文获得了4,921个网络钓鱼地址, 用于训练和测试识别模型.

4.2　数据处理

在获取所有数据后, 发现样本类别非常不平衡以及大量的交易数据, 类别不平衡比率, 即负样本与正样本的规模之比, 超过26,000. 为了减少计算量以及缓解样本不平衡性, 建立有效的模型, 本文提出了4种过滤规则: 1) 过滤掉钓鱼账户的一阶和二阶交易账户及其交易记录; 2) 消除掉交易数小于10大于1,000的地址; 3) 忽略出现在块高度200万之前的所有交易; 4) 过滤涉及智能合同地址及其交易记录. 上述过滤方法基于以下原因考虑: 首先, 钓鱼诈骗账户获得非法收益后, 一个容易想到的操作是通过“洗钱”使收益合法化或套现, 这必然涉及许多用于“洗钱”的中间账户, 这部分账户本质上属于钓鱼诈骗团伙, 明显不是正常账户, 因此, 可以通过剔除这类账户以减少样本的不均衡性. 为此将过滤掉与钓鱼账户发生交易的一阶和二阶交易账户及其交易记录(过滤规则1). 其次, 通过交易记录来了解钓鱼账户的行为特征信息, 而交易记录太少不利于学习. 此外, 过多的记录表明, 该账户可能是一个钱包或其他类型的账户. 事实上, 有超过1,000条交易记录的地址中, 只有一个地址被标记为网

络钓鱼. 因此, 为了能够学习到钓鱼账户的特征, 剔除掉交易数小于10 或大于1,000的地址(过滤规则2). 接下来, 通过分析网络钓鱼地址的最开始发生的时间, 发现所有的网络钓鱼地址在2016年8月2日之后才处于活跃状态. 这可能是因为在以太坊的早期, 网络钓鱼诈骗相对较少, 被标记的数量甚至更少. 因此, 本文仅基于区块高度为200万(即2016年8月2日)后的交易记录建立模型(过滤规则3). 此外, 智能合约地址在网络钓鱼地址中的占比很小(即7%), 而且它们通常与Token 有关. 因此, 在本研究中, 忽略了智能合约地址(过滤规则4).

在本文中, 通过过滤规则得到了数据集D1, 其中包含了430,956个未标记地址和2,917个真实的网络钓鱼地址, 正负样本比约为1:150. 相比于会议版论文[5]中的数据集, 正负样本比约为1:1, 600. 本文中提出的过滤规则1能够有效减少数据集的不平衡性, 并且得到了更多的钓鱼地址样本, 因此有利于建立更鲁棒的模型. 为了证明过滤规则1对模型的影响有限以及模型的泛化能力, 将被过滤掉的账户和所有的钓鱼账户一起命名为数据集D2, 并通过过滤规则2, 3, 4后得到了529,897个未标记的地址, 正负样本比约为1:180. 值得注意的是, 本文中使用的特征工程在这两个数据集中是一样的.

5 实验结果

在本节将介绍本文所提出的模型在实验中的表现. 首先, 详细描述实验的实现方法; 然后, 通过与几种基线方法进行比较来证明本文提出的框架的有效性.

5.1 实验设置

本文的实验部分采用了数据集D1和数据集D2进行模型验证, 并将数据集随机分为80%用于模型训练和20%用于模型测试. 为了避免测试集和测试集划分造成的偶然性, 本文采用了k折交叉验证的评价方法, 其中$k = 5$. 为了更准确地反映模型的有效性, 本文采用了4个评价指标, 即Precision, Recall, F1-score和AUC(area under the curve).

5.2 实验结果分析

在本节将评估本文提出的模型是否更适合解决网络钓鱼检测问题, 并对其性能进行评估. 本节将对所提出的模型进行双层堆叠集成模型 (DSEL) 的比较, 同时, 还将与传统的机器学习算法和之前提出的DElightGBM算法进行比较. 这些实验的结果将在表3–4中展示.

表 3 在数据集D1上在DESL算法和各种算法的性能

Table 3 Performance of the DESL algorithm and various algorithms on dataset D1

Measures	ET	RF	XGB	LGB	DElightGBM	DSEL
AUC	0.6272	0.7016	0.8203	0.8095	0.7754	0.8397
Precision	0.9625	0.9721	0.9041	0.9243	0.9225	0.9532
Recall	0.2545	0.4033	0.6411	0.6193	0.5767	0.7196
F1-score	0.4026	0.5701	0.7502	0.7448	0.7071	0.7935

表 4 在数据集D2上在DESL算法和各种算法的性能

Table 4 Performance of the DESL algorithm and various algorithms on dataset D2

Measures	ET	RF	XGB	LGB	DElightGBM	DSEL
AUC	0.6595	0.6835	0.7206	0.7050	0.7588	0.7906
Precision	0.6672	0.7054	0.7232	0.7565	0.8702	0.8678
Recall	0.3249	0.3418	0.4477	0.4151	0.6524	0.6822
F1-score	0.4280	0.4870	0.5409	0.5285	0.6422	0.7577

如表3所示, 本文提出的双层堆叠集成模型 (DSEL)与传统的机器学习算法以及之前提出的DElightGBM算法在D1数据集上的性能. 从表1显示的结果可以看出, 在这些单模型中, 机器学习算法极端随机树和随机森林除了Precision这个指标以外, 其他3个指标表现不佳. 相反, 在采用集成学习 LightGBM, XGBoost和DElightGBM算法之后, 各个指标都取得了较好的表现. 由此可见, 对于不平衡的数据集, 利用集成学习会有更好的效果. 因为集成学习主要用bagging的方法, 将多个分类器融合在一起, 然后用投票的方法得到最后的结果, 此举能够提高模型的泛化性能, 对不平衡问题有一定的帮助. 而本文提出的DSEL模型融合了各种分类器, 最终使得所有指标均有了显著的提高. 其中Precision达到95.32%, AUC达到了83.97%等, 从各个数据来看, 本文所提出的模型具有更好的效果.

如表4所示, 本文提出的DSEL算法与传统机器学习算法以及之前提出的DElightGBM算法在数据集D2

上的性能. 从表4中的结果可以看出, 所提出的DSEL模型表现出了很好的性能, 除了Precision略低于DElightGBM之外, 其他指标均优于其他模型. 例如, AUC达到了79.06%, F1-score达到了75.77%. 这表明本文所提出的模型对于不平衡的数据集仍然具有很好的效果. 这些结果表明了一个值得注意的现象, 本文所提出的模型在不同的数据集上都表现出很好的性能. 其中, DSEL模型在D2数据集上的表现仅比在D1数据集上的指标低了约3%, 这不仅证明了本文提出的新的过滤规则的合理性, 还证明了本文提出的模型具有良好的迁移性. 这意味着本文所提出的模型可以部署在实际的钱包中, 用于实时预警.

5.3　特征分析

本文采用了级联特征提取的方法, 得到了大量的特征. 在DSEL模型中, 使用特征重要性分析方法, 得到了前15个重要特性, 如图4所示. 本文选取排名前两名的重要特征进行分析, 证明了本文提出特征工程的有效性, 充分挖掘了节点本身和不同邻居节点的特征, 并进一步提高了模型的性能.

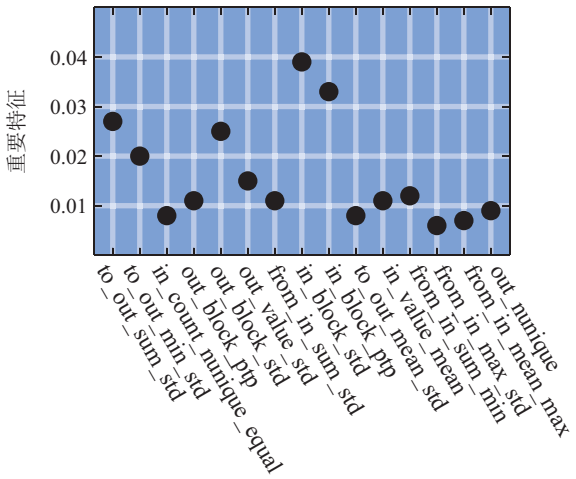

图 4　排名前15的重要特征

Fig. 4　The top 15 important features

in_block_std是指节点转入交易中交易时间(区块编号)的标准差. 该特征反映了地址的转入交易强度. 如果在短时间内有大量的转入交易, 这些交易的区块编号将非常接近. 这个特征的重要性很容易理解. 对于钓鱼地址, 在开始钓鱼后的一段时间内, 交易数量突然增加. 然而, 在被发现之后, 基本就没有转入交易了. 因此, 可以发现钓鱼地址的交易基本上只存在于一个小的时间周期内.

in_block_ptp是指节点转入交易中交易时间(区块编号)的时间跨度. 如前所述, 钓鱼地址的交易基本上只存在于一个小的时间周期内, 因此, 这个特征可以很好地表示这个属性.

6　结论与未来工作

在区块链生态系统中, 各种骗局非常猖獗, 严重威胁到相关用户的资产安全. 为了解决这个问题, 本研究提出了一种基于模型堆叠的框架, 并以以太坊系统中的钓鱼诈骗账户识别为例验证了框架的有效性. 具体而言, 首先, 使用Parity客户端同步了所有以太坊交易记录, 并通过爬取etherscan.io收集了标记为钓鱼欺诈的所有地址, 构建了数据集. 其次, 为缓解了数据极度不平衡性, 本文提出了一种新的过滤规则, 并对其进行了验证, 在此基础上, 基于账户间的交易记录, 构造了交易图, 并通过基于图的级联特征提取方法, 提取了许多有用的特征. 接下来, 本文提出了一个双层堆叠集成学习模型来检测钓鱼账户. 最后, 本文从多个角度对该模型进行了评估, 结果表明了该模型的有效性. 该方法在实际应用中具有指导意义, 并可有效提高交易平台的安全性和用户体验.

在未来的工作中, 一方面, 需要进一步优化模型, 改进模型存在的误判问题. 未来将结合图分类、多特征学习等方法进一步优化模型, 降低模型的误判率和漏判率等指标; 另一方面, 将进一步扩充数据集, 验证模型的泛化能力, 改进模型可能存在的过拟合与欠拟合问题. 此外, 在实际应用中, 可结合更多的方法以增强实际应用价值. 一个典型的方式是建立黑白名单规则, 结合过滤规则, 这将进一步缓解样本不均衡问题.

参考文献:

[1] ZHENG Z, XIE S, DAI H, et al. Blockchain challenges and opportunities: A survey. *International Journal of Web and Grid Services*, 2018, 14(4): 352 – 375.

[2] BARTOLETTI M, CARTA S, CIMOLI T, et al. Dissecting ponzi schemes on ethereum: Identification, analysis, and impact. *Future Generation Computer Systems*, 2020, 102(1): 259 – 277.

[3] CHEN W, ZHENG Z, CUI J, et al. Detecting ponzi schemes on ethereum: Towards healthier blockchain technology. *In Proceedings of the 2018 World Wide Web Conference*. Lyon, France: International World Wide Web Conferences Steering CommitteeRepublic and Canton of Geneva, 2018: 1409 – 1418.

[4] LIU J, YE Y. Introduction to e-commerce agents: Marketplace solutions, security issues, and supply and demand. *In E-Commerce Agents*. Berlin Heidelberg, Germany: Springer, 2001: 1 – 6.

[5] CHEN W, GUO X, CHEN Z, et al. Phishing scam detection on ethereum: Towards financial security for blockchain ecosystem. *In Proceedings of the 29th International Joint Conference on Artificial Intelligence*. Yokohama Yokohama, Japan: International Joint Conferences on Artificial Intelligence, 2021: 4506 – 4512.

[6] SZABO N. Smart contracts: Building blocks for digital markets. *EXTROPY: The Journal of Transhumanist Thought*, 1996, 18(2): 28.

[7] GARG R. Ethereum based smart contracts for trade and finance. *International Journal of Economics and Management Engineering*, 2022, 16(11): 619 – 629.

[8] PEROZZI B, AL-RFOU R, SKIENA S. DeepWalk: online learning of social representations. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. New York, USA: Association for Computing Machinery, 2014: 701 – 710.

[9] GROVER A, LESKOVEC J. Node2vec: Scalable feature learning for networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*. New York, USA: Association for Computing Machinery, 2016: 855 – 864.

[10] KIPF T, WELLING M. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*. Toulon, France: OpenReview.net, 2017.

[11] WU J, YUAN Q, LIN D, et al. Who are the phishers? Phishing scam detection on ethereum via network embedding. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2022, 52(2): 1156 – 1166.

[12] CHEN L, PENG J, LIU Y, et al. Phishing scams detection in ethereum transaction network. *ACM Transactions on Internet Technology (TOIT)*, 2020, 21(1): 1 – 16.

[13] LI S, GOU G, LIU C, et al. TTAGN: Temporal transaction aggregation graph network for ethereum phishing scams detection. *Proceedings of the ACM Web Conference 2022 (WWW '22)*. New York, USA: Association for Computing Machinery, 2022: 661 – 669.

[14] LI S, XU F, WANG R, et al. Self-supervised incremental deep graph learning for ethereum phishing scam detection. *ArXiv Preprint*, 2021, arXiv: 2106.10176.

[15] CHATZAKOU D, KOURTELLIS N, BLACKBURN J, st al. Mean birds: Detecting aggression and bullying on twitter. *In Proceedings of the ACM on web science conference*. New York, USA: Association for Computing Machinery, 2017: 13 – 22.

[16] RAMALINGAM D, CHINNAIAH V. Fake profile detection techniques in large-scale online social networks: A comprehensive review. *Computers & Electrical Engineering*, 2018, 65(1): 165 – 177.

[17] LIU X, WU J, ZHOU Z. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2008, 39(2): 539 – 550.

[18] KALABARIGEL R, RAO R, ABRAHAM A, et al. Multilayer stacked ensemble learning model to detect phishing websites. *IEEE Access*, 2022, 10: 79543 – 79552.

[19] BREIMAN L. Random forests. *Machine Learning*, 2001, 45(1): 5 – 32.

[20] GEURTS P, ERNST D, WEHENKEL L. Extremely randomized trees. *Machine Learning*, 2006, 63(1): 3 – 42.

[21] FRIEDMAN J, HASTIE R, TIBSHIRABI R, et al. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 2000, 28(2): 337 – 407.

[22] CHEN T, GUESTRIN C. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. New York, USA: Association for Computing Machinery, 2016: 785 – 794.

[23] KE G, MENG Q, FINLEY T, et al. Lightgbm: A highly efficient gradient boosting decision tree. *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc, 2017: 3146 – 3154.

作者简介:

陈伟利 博士, 研究生导师, 副研究员, 中国计算机学会(CCF)会员, 主要研究领域为区块链、异常检测、机器学习等, E-mail: mathutopia@163.com;

叶明顺 硕士, 主要研究领域为区块链、数据挖掘等, E-mail: mingshunye@126.com;

唐明董 博士, 研究生导师, 教授, 中国计算机学会(CCF)会员, 主要研究领域为区块链、软件服务、数据挖掘等, E-mail: mdtang@126.com;

郑子彬 博士, 研究生导师, 教授, 中国计算机学会(CCF)会员, 主要研究领域为区块链、智能合约、数据挖掘、软件可靠性等, Email: zhzibin@mail.sysu.edu.cn.