面向区块链平台的庞氏骗局模式检测方法

毛典辉,梁秀霞,赵　爽,郝治昊
(北京工商大学 计算机学院,北京 100048)

摘　要:区块链技术的出现给各行各业带来了新的变革,同时也给诈骗提供了新的平台。作为金融诈骗的代表形式——庞氏骗局借助智能合约在二代区块链上给人们制造了巨大的损失,这不仅影响区块链技术的发展,同时也在一定程度上扰乱了正常的社会经济秩序,因此,对借助区块链技术实施庞氏骗局的相关平台进行监管势在必行。该文选取区块链平台以太坊作为研究对象,设计了一种基于智能合约混合特征的庞氏骗局检测算法。首先根据交易主体间的关联特征判断其是否符合庞氏骗局中回报不公平的金字塔交易形式,提取智能合约交易特征;其次根据智能合约的操作代码在庞氏骗局合约和其他合约出现频率设计了一种新的 ITF 算法,提取区分庞氏骗局智能合约的操作码特征;最后采用 Catboost 集成学习算法来训练庞氏骗局检测模型,算法强调多个特征之间的联系,并解决检测算法训练过程中梯度偏差以及预测偏移问题。与其他算法相比,该算法在庞氏骗局检测上具有较高的准确率(精确率=0.89、召回率=0.78、F1 值=0.82)。

关键词:庞氏骗局;区块链;以太坊;智能合约;TF-IDF;集成学习;Catboost
中图分类号:TP301　　**文献标识码**:A　　**文章编号**:1673-629X(2022)05-0153-07
doi:10.3969/j.issn.1673-629X.2022.05.026

Ponzi Scheme Pattern Detection Method for Blockchain Platform

MAO Dian-hui,LIANG Xiu-xia,ZHAO Shuang,HAO Zhi-hao
(School of Computer,Beijing Technology and Business University,Beijing 100048,China)

Abstract:Blockchain technology has brought new changes to all walks of life,but also provides a new platform for fraud. As a representative form of financial fraud,Ponzi scheme has created huge losses to people on the second-generation blockchain with the help of smart contracts,which not only affects the development of blockchain technology,but also disrupts the normal social and economic order to a certain extent. Therefore,it is imperative to regulate the relevant platform in a Ponzi scheme which use blockchain technology. Ethereum,a blockchain platform,is selected as the research object to design a Ponzi scheme detection algorithm based on the mixed features of smart contracts. Firstly,according to the correlation characteristics of the transaction subjects,it is judged whether it conforms to the pyramid trading form of Ponzi scheme with unfair returns,and the transaction characteristics of smart contract are extracted. Secondly,a new ITF algorithm is designed according to the occurrence frequency of the operation codes of smart contracts in Ponzi scheme contracts and other contracts to extract the characteristics of the operation codes that distinguish the Ponzi scheme smart contracts. Finally,Catboost ensemble learning algorithm is used to train Ponzi scheme detection model. The algorithm emphasizes the relationship between multiple features,and solves the gradient deviation and prediction offset problems in the training process of detection algorithm. Compared with other algorithms,the proposed algorithm has higher accuracy in Ponzi scheme detection(precision rate=0.89,recall rate=0.78,F1 value=0.82).

Key words:Ponzi scheme;blockchain;Ethereum;smart contract;TF-IDF;ensemble learning;Catboost

1　概　述

2008 年,中本聪发表了《Bitcoin: A Peer-to-Peer Electronic Cash System》[1]一文,讨论了一个电子现金系统,它是以区块链[2-3]为底层架构的虚拟货币平台,由此奠定了区块链技术发展的基础。由于区块链具有不依赖于第三方管理机构,可通过分布式进行数据的核算和存储,具有去中心化、不可篡改、可追溯和匿名性等特点,为区块链奠定了坚实的“信任”基础。正是由于去中心化特性,区块链技术游走于法律灰色地带,缺乏有关部门的监管;匿名性隐藏了用户的真实身份,更是增加了区块链监管难度。因此,各种诈骗组织借助区块链技术[4]大肆进行金融犯罪,如洗钱[5]、钓鱼

收稿日期:2021-04-15　　**修回日期**:2021-08-17
基金项目:国家社会科学基金(18BGL202)
作者简介:毛典辉(1979-),男,博士,教授,从事区块链与 AI 融合研究;通讯作者:梁秀霞(1995-),女,硕士,CCF 会员(B4644G),从事区块链与诈骗检测研究。

网址[6-7]和庞氏骗局[8]等。根据加密分析专家 Chainalysis 的最新报告[9],以区块链作为底层技术的以太坊平台(ETH)是深受欢迎的二代区块链,已经成为区块链诈骗的首选加密平台。因此,加强对以太坊为代表的区块链平台的金融诈骗监管迫在眉睫。

基于智能合约[10-11]的以太坊诈骗又称为智能庞氏骗局。在庞氏骗局持续过程中,随着投资者的加入,合约将获得的投资优先返还最先加入的人,后续投资者无法获得回报而失去他们的投资。前者和后者使用相同的投资金额却没有相同金额的回报,这是极其不公平的。显然,这种合约给大多数后来者造成了损失,骗取了大量的财富。

对庞氏骗局智能合约进行监测时面临了以下挑战:(1)在特征提取方面,要求提取贴近诈骗本质的特征,使庞氏骗局能够具有更好的区分度;(2)在检测分类方面,要求能够提出减小目标泄漏导致的预测偏移问题,提高分类准确率的算法。

针对上述挑战,该文设计了一种基于智能合约混合特征的庞氏骗局检测算法,提取交易主体间符合庞氏骗局中回报不公平的金字塔形式交易特征和能够区分庞氏骗局的智能合约操作码特征,并采用 Catboost 集成学习[12]的分类算法解决算法训练过程中的梯度偏差以及预测偏移问题。

2 技术方案

首先从 Etherscan. io 爬取智能合约相关交易及合约代码,从交易记录中可获得合约账户得到投资的具体金额和时间戳,以及在回报分配上的比例,统计相关信息得到交易特征。由于智能合约以字节码的形式保存于以太坊平台,该文通过反汇编将字节码转换为操作码,利用改进的 TF-IDF[13]获取合约的代码特征。将交易特征和代码特征结合成混合特征向量,进入由若干个弱分类器组成的 Catboost[14]集成学习,得到庞氏骗局分类训练模型,最终实现对新交易合约的预测。因此,设计的庞氏骗局智能合约检测算法技术路线如图 1 所示。

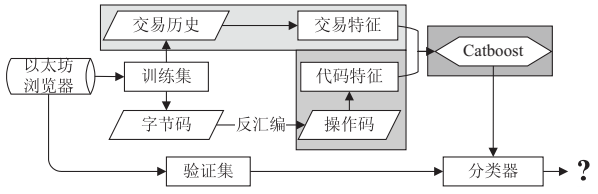


图 1 整体技术流程

2.1 交易特征

为吸引更多投资者的加入,前期的庞氏骗局账户频繁与投资者互动,互动形式表现为给前期投资者返还回报利息。因此,庞氏骗局的账户余额通常保持在较低水平,而展现在投资者面前的则是一个守信的合约账户,它能够时常给投资者返还回报利息,并且保证所有的投资者都能够获得应得的回报。因此,以下交易特征对分类效果有很好的作用:

活跃度:记录智能合约账户与外部账户(即投资者账户)频繁交易的特点。

账户余额:智能合约账户余额。

投资交易数:对智能合约进行投资交易的次数。

返利交易数:从智能合约获取返利交易的次数。

返利者比率:返利者占投资者的百分比。

最大返利次数:参与者获得返利的最大次数。

通过观察交易记录,发现在庞氏骗局中,大部分的投资进入合约创建者的口袋,而使得回报总返利金额远低于总投资金额。完整诈骗过程中,在所有投资者获得回报之前庞氏骗局便宣告破产,使得多数迟到的投资者得不到回报。此外,获得回报的投资者,其回报利息的高低也受投资时间影响,使得回报差异呈现出一个高水平。基于此,提取了可结合投资与回报相关特点的特征,如下:

总投资额:智能合约账户获得的投资总额。

总返利额:智能合约账户指出的返利总额(诈骗者需要从合约中获取利益。大部分庞氏骗局不会将全部投资作为返利返回给参与者,而是将一部分投资作为手续费返利给合约创建者,因此可以看到庞氏骗局中回报交易的金额高于投资金额)。

总投资人:进行投资的地址账户数。

总返利人:获得回报的地址账户数(并不是所以投资者都可以获得相应的回报。所以,获得返利的账户地址会少于投资账户地址)。

返利标准差:智能合约账户返利的标准差(投资者回报金额的标准差反映了回报的不平衡,并由此判断处庞氏骗局的回报金额呈现金字塔结构)。

2.2 代码特征

以太坊上庞氏骗局以字节码的形式存在,字节码可反汇编为操作码。由于庞氏骗局的行为与普通智能合约的经济行为有一定的区别,在操作码中表现为对庞氏骗局的重要性不同。获取操作码重要指标 ITF 的准备阶段需将从以太坊区块链浏览器上爬取的字节码进行反汇编以得到操作码集合,之后通过 ITF 算法得到操作码的重要性。详细过程如下:

(1)使用 pyevmasm 工具将 EVM 字节码反汇编成操作码,通过预处理操作,最终得到可作为研究使用的数据集合。

(2)统计在特定合约内特定操作码出现的次数以及合约操作码数,通过计算 TF 实现对于操作码数量的归一化处理。

(3)通过统计特定操作码在合约数据集合中出现的频率和合约数据集内操作码的数量来度量该词语对于合约代码特征的重要程度。

(4)计算特定操作码的重要指标 ITF 值,即特定操作码对特定合同的重要性。

在特定的智能合约操作码片段中,操作码出现的频率呈现一定的概率,高频率的特定操作码是庞氏骗局的代码特征。该文首先对操作码数量进行归一化处理,以得到庞氏骗局中各操作码的频率。以操作码 m 为例,对 m 进行归一化处理,如公式(1):

$$\mathrm{TF_m} = \frac{n_\mathrm{m}}{N} \tag{1}$$

其中, $\mathrm{TF_m}$ 为归一化处理后的频率; n_m 为特定操作码在某一个庞氏骗局中出现的次数; N 为特定操作码在所有智能合约中出现的次数。

之后,公式(2)通过统计“m”在合约数据集合中出现的频率来度量该词语对于合约代码特征的重要程度,即对该词语的 IDF 值进行计算:

$$\mathrm{IF_m} = \log\left(\frac{S}{D_\mathrm{m}+1}\right) \tag{2}$$

其中, S 为智能合约数据集合中所有操作码总数; D_m 表示智能合约数据集合中操作码“m”的数量。

最后,在上述基础上,通过公式(3)即可计算得出词语“m”的 ITF 值,若该词语在庞氏骗局中出现频率较高,而在智能合约数据集合(即其他合约)中出现频率较低,则认为该词语对庞氏骗局代码具有良好的表征能力,有利于与其他智能合约进行区分并实现分类。

$$\mathrm{ITF_m} = \mathrm{TF_m} * \mathrm{IF_m} \tag{3}$$

其中,$\mathrm{TF_m}$和 $\mathrm{IF_m}$分别来自公式(1)、公式(2)。ITF 与传统的 TF-IDF 算法并不十分相同,ITF 中采用的是结合操作码“m”分别在庞氏骗局的词频与在所有数据集中的词频表示其对于庞氏骗局的重要程度,这种算法更适合于提取操作码的特征。

3　分类模型

由于交易是一种复杂的经济行为,庞氏骗局与其他经济行为相似度高,易混淆。普通的机器学习在庞氏骗局检测上得不到优秀的准确率,一般的集成学习也容易在一定程度上产生预测偏离的问题,使得最终的准确率并不让人信服其分类效果。

该文采用了 Catboost 集成学习算法来训练庞氏骗局检测模型。与其他的集成学习方法相比,Catboost 为训练集生成一个随机序列处理特征类别,同时使用完全对称的二叉决策树[15]作为基础预测器解决预测偏移的问题,提高了分类效果的准确性。图 2 是分类器的技术框架。

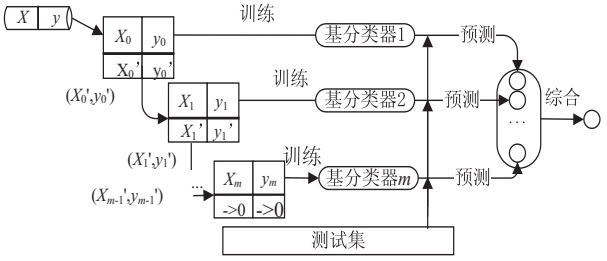


图 2　分类模型技术框架

把爬取到的数据视为一个数据集 $D=\{(x_k,y_k)\mid k=1,2,\cdots,n\}$,其中 $x_k=(x_k^1,x_k^2,\cdots,x_k^m)$是 m 个特征的随机向量, y_k 是数据 x_k 的标签,若 $y_k=1$,表示 x_k 是数据集里的一个庞氏骗局账户,反之则表示 x_k 是其他正规的账户。每个(x_k , y_k)都是独立的,它根据某个未知的分布 P(.,.)恒等分布。设置 $\sigma=(\sigma_1,\sigma_2,\cdots,\sigma_n)$ 对数据集进行随机排列。根据式(4)得到数值型特征:

$$\frac{\sum_{j=1}^{p-1}[x_{\sigma_j,k}=x_{\sigma_p,k}]Y_{\sigma_j}+ap}{\sum_{j=1}^{p-1}[x_{\sigma_j,k}=x_{\sigma_p,k}]+a} \tag{4}$$

其中,$[x_{\sigma_j,k}=x_{\sigma_p,k}]$是一个艾弗森括号,当$x_{\sigma_j,k}=x_{\sigma_p,k}$时可得结果 $k=1$,否则等于 0。这里表示随机排列的数据集与原来的数据集进行匹配, p 是一个先验值,其权值为 a ,其中 $a>0$。

需要训练近似函数 $F:R_m->R$ 来尽可能降低在特征转换过程的损失:

$$L(F):=\mathrm{EL}(y,F(x)) \tag{5}$$

其中, $L(.,.)$是平滑损失函数,(x,y)是独立于训练集 D 的测试集 P 中的样例。

根据近似函数 $F^t=F^{t-1}+\partial h^t$,可以得到一系列迭代 $F^t:R^m->R$,其中 α 是步长,函数 h^t 是一个基础预测器,可以将预期的损失降到最低。

$$h^t=\underset{h\in H}{\operatorname{argmin}}L(F^{t-1}+\partial h^t)=\underset{h\in H}{\operatorname{argmin}}\mathrm{EL}(y,F^{t-1}(x)+h(x))$$

使用最小二乘近似,负梯度步近似最小化问题:

$$h^t=\underset{h\in H}{\operatorname{argmin}}(-g^t(x,y)-h(x))^2 \tag{6}$$

在 Catboost 的基本预测器中,完全对称二叉决策树递归地将特征空间 R^m 划分为若干个独立区域 R_j (树节点), b_j 为叶子节点,每个叶子节点被分配一个值,该值是被预测的类别。根据 $x^k>t$ 判断最终所属类别。决策树 h 可以写成:

$$h(x)=\sum_{j=1}^{J}b_jI\{x\in R_j\} \tag{7}$$

4　实验结果

4.1　实验环境及参数

提取特征及分类模型训练实验都是在 anaconda

内使用 Python 3.6 环境下来完成的。在实验过程中，采用 Catboost 深度学习框架构建庞氏骗局检测模型并进行模型训练。使用了 anaconda 自带的一些工具包数据的处理分析，如 pandas、numpy 等。

4.2 数据集

获取的代码数据有重复或破损，手动检查并删除了无效地址，最终实际得到 1 393 个非庞氏骗局的地址和 123 个庞氏骗局的地址。在数据集中交易数据与合约数据分别保存在不同的文件夹。其中，每个合约地址都有对应的若干个历史交易。每个交易都含有时间戳、交易金额和手续费等信息。合约操作码以文本文档文件保存，每一个智能合约都是一个字典格式。

4.3 评判标准

需要通过优秀的性能指标来判断分类模型的性能。使用精确率(Precision)、召回率(Recall)和 F1 分数(F1-score)来评判分类模型的性能。各项指标计算公式如下：

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

$$\text{F1 - score} = \frac{2 * \text{Preision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

4.4 实验结果与分析

4.4.1 特征提取结果

该文利用交易数据提取了 11 个相关特征。提取的特征中不仅有与合约相关的余额、活跃度等特征，同时包含从参与者角度的投资和返利相关数据。最后以庞氏骗局和非庞氏骗局为集合分别将提取的特征值计算得到均值和标准差，结果如表 1 所示。

表 1 交易特征

特征		Ponzi		Nonponzi	
		均值	标准差	均值	标准差
活跃度		0.51	0.24	0.18	0.34
余额		3.35	16.34	360.65	1 945.66
返利比		0.59	0.21	0.21	0.36
最大返利次数		72.11	135.66	158.48	922.97
交易数	投资	46.77	107.95	619.14	2 115.80
	返利	144.52	412.67	276.90	1 395.65
总额	投资	266.00	966.00	296.00	4 690.00
	返利	485.00	2 690.00	387.00	5 550.00
人数	投资	13.60	31.78	123.38	585.63
	返利	7.22	12.81	26.19	317.21
返利标准差		511.00	367.00	1 960.00	43 200.00

表 1 中清楚地展示了各项交易特征的均值和标准值。发现庞氏骗局各项标准差都小于非庞氏骗局，表示庞氏骗局各项交易特征的值比较接近，更有可能具有相似的行为。例如：庞氏骗局的活跃度标准差小于非庞氏骗局的活跃度标准差，而其均值却相对大得多。活跃度越高表示庞氏骗局希望通过与投资者互动返利以吸引越多的投资者。

该文参考了 TF-IDF 的思想，利用改进的新算法 ITF 提取合约的操作码特征。操作码特征分别在庞氏骗局和非庞氏骗局中的均值如表 2 所示，其中，Ponzi 为 1 的列数据表示庞氏骗局的数据。

表 2 智能合约部分 ITF 取值

操作码	1	0
GASLIMIT	0.014 7	0.000 3
EXP	0.134 3	0.069 6
CALLDATALOAD	0.041 8	0.034 8
SLOAD	0.174 0	0.076 6
CALLER	0.053 5	0.023 7
LT	0.076 1	0.020 0
GAS	0.018 1	0.009 4
MOD	0.022 5	0.001 1
MSTORE	0.147 9	0.110 0

在表 2 中最后一项，当 Ponzi = 1 时表示为庞氏骗局，Ponzi = 0 表示为非庞氏骗局。表中所有的数据表示为代码特征的均值，可以看到此九个操作码在庞氏骗局的 ITF 值较非庞氏骗局的高，将其作为特征进行分类有极大可能增加分类的准确率。

4.4.2 分类结果

使用的 Catboost 集成学习分别与 Knn、SVM[16]、DT、XGBoost[17]和 RF[18]做了对比实验，实验结果如表 4 所示，其中 Tra 表示分类过程仅使用交易特征进行实验，Code 表示仅使用代码特征进行实验，而 Com 表示用混合特征进行分类实验。

从表 3 中可以看到，提出的混合特征和基于 Catboost 的庞氏骗局检测方法性能更好，原因在于 Catboost 对新增的特征有很好的适应性，其完全对称树可在最大程度上避免预测便宜问题。提出的 Catboost 模型虽然在仅使用交易特征的实验中结果并不好，但结合代码特征却能够得出最好的结果。在使用混合特征的 Catboost 实验各项评判指标均达到最优，因此可以认为 Catboost 能够很好地利用交易特征与代码特征，使得结果最优化。若是仅使用代码特征进行实验，XGBoost 与 RF 的表现似乎都不错，但是不要忽视了在 Recall 指标上 XGBoost 比 RF 高了 3 个百分点，其表明 XGBoost 更容易得到庞氏骗局的正确分类。如果仅使用代码特征结果都不算好。分析各种模型后，发现 DT 在仅使用交易特征的实验中可以得出

相对较好的结果,尤其在 Recall 和 F1-score 指标上明显强于其他分类模型。在缺少智能合约的庞氏骗局上可能能够得到最优的结果,因此 DT 可能更加适应于 Bitcoin 或其他第一代区块链的诈骗检测。

表 3　分类结果对比实验

Algorithm	Tra			Code			Com		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
KNN	0.51	0.20	0.26	0.74	0.80	0.73	0.58	0.33	0.38
SVM	0.48	0.11	0.17	0.77	0.63	0.69	0.73	0.36	0.47
DT	0.61	0.62	0.63	0.62	0.71	0.67	0.35	0.27	0.29
XGBoost	0.60	0.35	0.43	0.86	0.76	0.79	0.85	0.73	0.77
RF	0.64	0.23	0.31	0.91	0.73	0.80	0.88	0.72	0.79
Catboost	0.68	0.30	0.40	0.86	0.77	0.80	0.89	0.78	0.82

采用 Catboost 在特征提取前后的实验效果对比如图 3 所示。其中,其他交易特征指的是仅与合约相关的余额、活跃度等特征,文中交易特征在其他交易特征的基础上考虑了提取投资与回报相关特征;其他代码特征值操作码词频,文中代码特征是提取操作码的 ITF 值;相应的混合特征则是将交易特征与代码特征混合形成特征向量。

图 3(a)、(b)、(c)表示仅使用交易特征、仅使用代码特征和使用了混合特征。新提取的特征能够在各项指标上胜于先前提取的特征,这表明了提取的交易特征与代码特征都对庞氏骗局有更好的表征能力。且从图 3(d)中可以看到,代码特征可以辅助代码特征,使得结合二者的混合特征可以在代码特征的实验上略胜一筹。鉴于此,在所有的诈骗中,即使代码特征可以很好地作用检测的特点,也不能忽视交易数据的帮助。

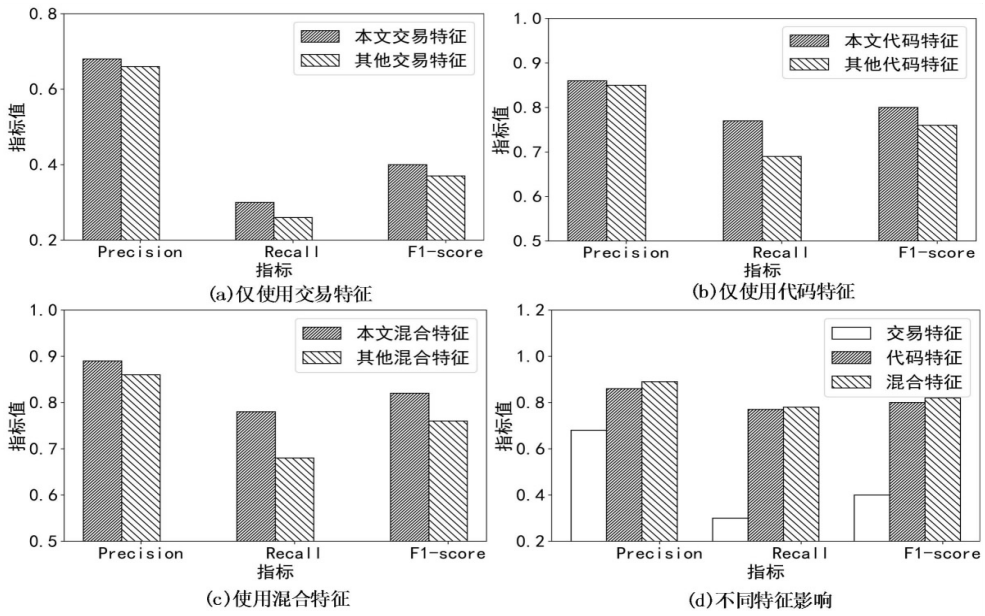


图 3　特征提取前后实验结果

4.4.3　模型参数变化影响分析

在使用 Catboost 模型对庞氏骗局进行检测分类时,参数会影响分类模型的性能,需要通过实验选取最好的参数值。该文选取对模型性能影响较大的参数进行实验,包括交叉验证的 K 值、损失函数、树的深度 depth 以及学习率 learning rate。各参数对实验结果的影响如图 4 所示。

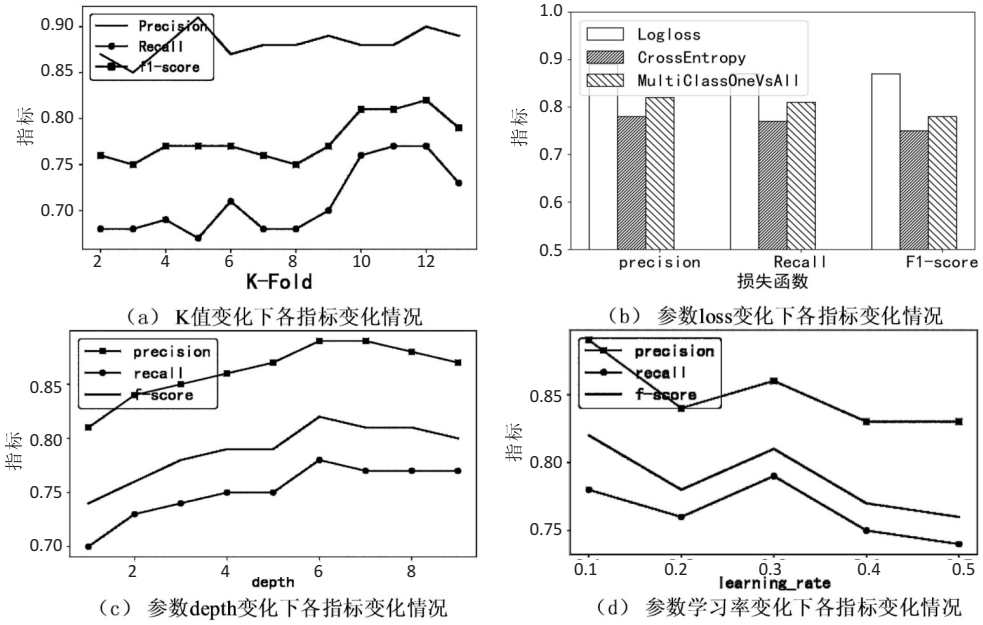
(1)K 折交叉验证(K-fold cross validation)指的是把训练数据 D 分为 K 份,用其中的 $K-1$ 份训练模型,剩余的 1 份数据用于评估模型的质量。将这个过程在 K 份数据上依次循环,并对得到的 K 个评估结果进行合并,如回归问题求平均或分类问题投票。由图 4 中可以看到,混合特征各项性能指标随交叉验证 K 值变化而变化。

图 4(a)中,在 $K=5$ 时,Precision 达到最高,然而此时 Recall 与 F1-score 指标过低,庞氏骗局大量被误判为非庞氏骗局,分类模型最好的情况是三项指标均接近 1;在 $K=12$ 时,各指标分别为:Precision = 0.90,Recall = 0.77,F1-score = 0.82,几乎都已达到最优。若 K 值进一步增加,三项指标迅速下滑,几乎可以认为,

在 $K=12$ 时提出的分类模型达到最优的值。

(2)在 loss 参数的选择上,Catboost 适用的 loss 函数包括 Logloss、CrossEntropy 和 MultiClassOeaVsAll,Logloss 损失函数能非常好地表征概率分布,对于分类问题可以很好地判断分类结果属于每个类别的置信度,非常适合于该文的分类问题。由图 4(b)可以看出,Logloss 函数对分类确实得到了良好的效果,因此,该文选用 Logloss 作为分类模型的损失函数。

(3)在基学习器的分类中,二叉树的深度对分类效果有一定的影响。过深的 depth 不仅不会提高分类的效果,同时极有可能造成分类结果的过拟合。从图 4(c)可以看出,在 depth=9 之前,随着 depth 深度的提高,分类效果呈现先增后减的趋势,同时在 depth=6 时达到最优的性能,因此,选取 depth=6 作为实验过程中的深度参数。



(a) K值变化下各指标变化情况　(b) 参数loss变化下各指标变化情况

(c) 参数depth变化下各指标变化情况　(d) 参数学习率变化下各指标变化情况

图 4　参数对实验结果的影响

(4)学习率作为监督学习中重要的超参,其决定着目标函数能否收敛到局部最小值以及何时收敛到最小值。合适的学习率能够使目标函数在合适的时间内收敛到局部最小值。从图 4(d)可以看出,随着学习率的递增,分类效果的三项指标均大致呈现一个递减的趋势,虽然在 0.3 时有一定的提升,但是提升效果没有超过 0.1 时的,因此,断定在 learning_rate=0.1 分类效果达到最优。

在实验的最后发现,部分庞氏骗局所有投资人都获得了回报,从交易特征观察似乎并不符合庞氏骗局。这样的庞氏骗局占比达到 1∶5。查看这些账户的源代码,发现其回报分配依旧呈现金字塔形态。另外,这些账户的投资者数量都太小。因此可以判断,这部分庞氏骗局是被正确分类的,但是由于是创建初期,其仍属于吸引投资时期,需要靠给投资者分配回报增加它的影响力。

5　结束语

在对以太坊平台的庞氏骗局进行检测中,提出了一种新的特征提取方法。在交易特征上还原庞氏骗局的金字塔回报率形式。在代码特征提取方面,借助 TF-IDF 的思想,提出了一种结合频率的操作码重要指数的特征提取方式,提取的代码特征从数据集出发,表示其对庞氏骗局的重要程度。针对训练过程梯度偏差以及预测偏移问题,提出了新的集成学习方法 Catboost,使用完全二叉树作为基础分类器在最大程度上解决计算梯度估计时存在目标泄漏导致的预测偏移问题。

该方法在一定程度上仍存在少许不足,如召回率只达到了 82%,意味 18% 的庞氏骗局被遗漏。在未来工作中,需对分类模型进行改进,以期达到更高的准确率。此外,类别不平衡对实验结果仍存在一定的影响,后续工作中不可将类别不平衡问题的处理全部寄托于模型中,可考虑扩大庞氏骗局类别数量,一定最大限度类别不平衡带来的影响。

参考文献:

[1] NAKAMOTO S. Btcoin:a peer-to-peer electronic cash system[EB/OL]. 2019. https://bitcoin. org/bitcoin. pdf Manubot.

[2] 李贵洪. 基于区块链的云存储数字取证[J]. 网络安全技术与应用,2021(4):155-156.

[3] 郭上铜,王瑞锦,张凤荔. 区块链技术原理与应用综述[J]. 计算机科学,2021,48(2):271-281.

[4] 陈伟利,郑子彬. 区块链数据分析:现状,趋势与挑战[J]. 计算机研究与发展,2018,55(9):1853-1870.
[5] FANUSIE Y,ROBINSON T. Bitcoin laundering:an analysis of illicit flows into digital currency services[EB/OL]. 2018. https://www. fdd. org/wp - content/uploads/2018/01/MEMO_Bitcoin.
[6] 陈　鹏,李勇志,余肖生. 基于特征选择模型的钓鱼网站快速识别方法[J]. 计算机技术与发展,2021,31(4):40-45.
[7] WU J,YUAN Q,LIN D,et al. Who Are the phishers? phishing scam detection on ethereum via network embedding[J]. arXiv:1911. 09259,2019.
[8] 喻文强,张艳梅,李梓宇,等. 以太坊庞氏骗局的类型分析与识别方法[J]. 重庆大学学报,2020,43(11):111-120.
[9] CHAINALYSIS. 2020 年加密犯罪报告 2020. [EB. OL]. 2020. http://www. 199it. com/archives/tag.
[10] SZABO N. Smart contracts:building blocks for digital markets[EB/OL]. 1996. http://www. fon. hum. uva. nl/rob/Courses/InformationInSpeech/CDROM/Literture/LOTwinterschool2006/szabo. best. vwh. net/smart_contracts_2. html.
[11] 欧阳丽炜,王　帅,袁　勇,等. 智能合约:架构及进展[J]. 自动化学报,2019,45(3):445-457.
[12] 徐继伟,杨　云. 集成学习方法:研究综述[J]. 云南大学学报:自然科学版,2018,40(6):1082-1092.
[13] YAMOUT F,LAKKIS R. Improved TFIDF weighting techniques in document retrieval[C]//2018 thirteenth international conference on digital information management (ICDIM). Berlin,Germany:[s. n.],2018:69-73.
[14] LI Y,MAI Y,LIN Z,et al. Online transaction detection method using catboost model[C]//2020 international conference on communications, information system and computer engineering (CISCE). Kuala Lumpur,Malaysia:[s. n.],2020:236-240.
[15] 王子玥,谢维波,李　斌. 采用口袋算法构造的多类别决策树模型[J]. 华侨大学学报:自然科学版,2019,40(1):121-127.
[16] BARTOLETTI M,PES B,SERUSI S. Data mining for detecting bitcoin ponzi schemes[C]//2018 crypto valley conference on blockchain technology (CVCBT). Zug, Switzerland:[s. n.],2018:75-84.
[17] CHEN W,ZHENG Z,CUI J,et al. Detecting ponzi schemes on ethereum:towards healthier blockchain technology[C]//WWW 2018:the 2018 web conference. Lyon,France:ACM,2018.
[18] CHEN W,ZHENG Z,NGAI E C H,et al. Exploiting blockchain data to detect smart ponzi schemes on ethereum[J]. IEEE Access,2019,7:37575-37586.

(上接第 152 页)

[8] 孔　琪,姚菁怡,戴宗昊,等. 针灸治疗重症肌无力如何取穴——一项数据挖掘研究[J]. 世界科学技术-中医药现代化,2021,23(2):647-654.
[9] 马　文,田　园. 基于聚类方法的工业电气设备大数据特征识别[J]. 计算机技术与发展,2020,30(11):190-194.
[10] 王雪姣,叶　枫. 基于关联规则算法的工业生产班组运行质量分析[J]. 计算机应用,2005,25(S1):211-212.
[11] 董轩萌,郭立稳,董宪伟,等. 基于 Apriori 算法的煤自燃影响因素关联挖掘[J]. 华北理工大学学报:自然科学版,2021,43(1):21-25.
[12] SONG Yunfeng. A correlation analysis model of human factors in mine accidents based on Apriori algorithm[J]. International Journal of Safety and Security Engineering,2020,10(3):22-28.
[13] BEKAR E T,NYQVIST P,SKOOGH A. An intelligent approach for data preprocessing and analysis in predictive maintenance with an industrial case study[J]. Advances in Mechanical Engineering,2020,12(5):168781402091920.
[14] 燕荣杰,王国庆,戴汝泉,等. 车联网数据预处理[J]. 物联网技术,2017,7(1):81-82.
[15] 戴新建. 基于大数据挖掘的广播电视客户价值分析[J]. 科技视界,2019(34):223.
[16] 张　莉. 基于聚类和决策树算法的成绩影响因素分析[J]. 中国石油大学胜利学院学报,2013,27(2):33-35.
[17] 陈　治,吴娟娟. 基于关联规则的医疗数据挖掘研究[J]. 统计与决策,2020,36(6):174-177.
[18] 吴　斌,马　超. 一种旅行数据约束关联规则挖掘算法[J]. 计算机工程与应用,2010,46(20):129-132.
[19] 肖　斌,肖亚飞. 时序关联规则在钻井事故中的应用[J]. 计算机应用,2017,37(S1):308-311.
[20] 马占欣,黄维通,陆玉昌. 相关度计算方法存在的问题及修正[J]. 计算机工程,2007,33(11):67-69.