Detecting Ponzi Schemes on Ethereum: Towards Healthier Blockchain Technology

一、 简介

区块链账户的匿名性，交易的不可篡改性以及智能合约执行的强制性，在成功解决了互不信任节点之间执行交易的同时，也使得以太坊中的欺诈行为更加难以分辨与防范。

作为一种典型的欺诈模式，“庞氏骗局”将后来者的投资金额，作为回报返还给最先的投资者，而并不通过实际的生产经营产生利润。而智能合约所具备的“强制执行性”，使得现如今的以太坊平台中，有许多掩藏在“智能合约”之下的旁氏骗局。由于智能合约的防篡改性和强制执行，会使得投资者坚定的认为他们的投资可以获得较好的汇报，从而身陷骗局。而发起人的匿名性也使得投资者根本无法追溯回自己的损失。

庞氏骗局的危害性使得检测“旁氏合约”的任务迫在眉睫，但是，不同于传统的旁氏骗局模式，检测区块链上的旁氏骗局仍然有以下问题：

1. 用户和投资者可能并不明白区块链的运行机理，更加难以理解隐藏在其后的旁氏骗局。
2. 区块链技术正处在不受监管的“灰色”地带，并没有相关的法律法规对其进行相应的约束
3. 智能合约的源代码往往难以获取得到

鉴于此，作者通过分析现有的标记数据，通过对智能合约的特征抽取，使用 XGBoost 模型来完成“旁氏合约”的检测工作。

二、 解决思路

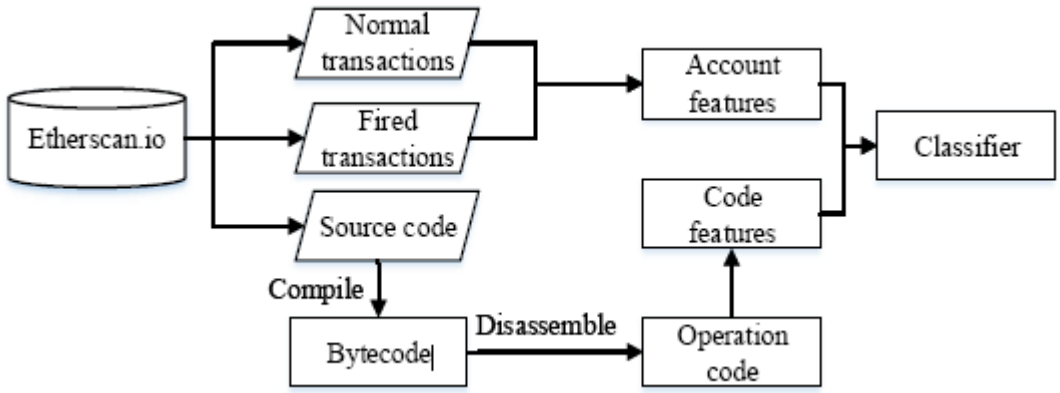


图 1 智能合约检测框架

如图 1 所示，作者首先通过公开途径，下载可以得到源码的智能合约和与其相关的交易、字节码数据；通过从交易和字节码构造特征，再使用分类模型以达到“旁氏合约”检测的目的。在数据方面，作者继承前人的工作，收集了 1382 个可以获得源码的认证合约，人工对其标注，并将其相关的交易和字节码从区块链网络中收集到，从而得到了 131 个“旁氏合约”和 1251 个“非旁氏合约”。

（一）交易特征

在以太坊中，作者将交易的数据分为两个类型，正常的以太币转移交易和通过智能合约触发的交易，而“旁氏合约”的主要特征之一，便是通过智能合约来

完成以太币的转移，除此之外，对于旁氏骗局，还有以下特征：

1. 支付的交易通常发生在投资交易之后，合约将会支付给一个已知的账户
2. 大多数的投资者将不会获得任何回报
3. 少数投资者将会获得绝大多数的回报

从这三个观测出发，作者启发式的构造了 7 个特征，并且从统计特征上表明，在这些特征在“旁氏合约”和“非旁氏合约”上具备明显的不同，例如收获回报的投资者比例等，详细介绍可以参看论文。

（二）字节码特征

通过公开的途径，我们只能获得智能合约编译过后的字节码，因此，作者将智能合约的字节码根据频率进行统计。如下图所示，左边的字节码云图是用一个庞氏骗局所做，能够明显发现 JUMPI 这样的字节码出现频率较高，这个和庞氏骗局需要更多的条件出发相吻合。

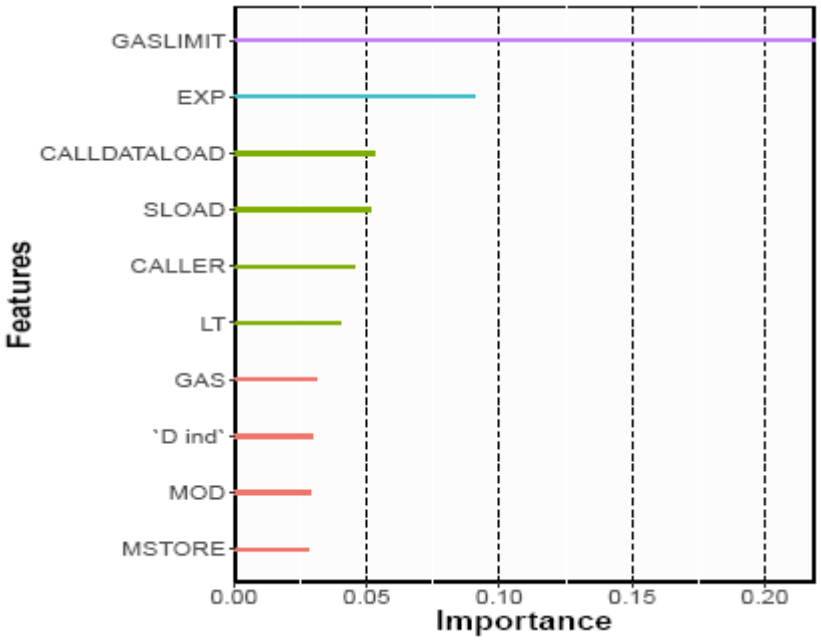
在构造出基本特征后，作者使用 XGBoost 作为分类模型。XGBoost 是一个基于树的集成模型，基本的原理是训练多个基分类器，每个基分类器都会将上一轮的结果一起考虑到目标函数之中，在很多分类任务上取得了非常好的效果。

三、 实验

在实验中，作者没有选择其他分类方法作为 Baseline,只是分别对比了使用不同模型的分类效果，结果如下：

Features	Precision	Recall	F-score
Account	0.74	0.32	0.44
Opcode	0.90	0.80	0.84
Account + Opcode	0.94	0.81	0.86

从这个结果中，我们可以看出，常用的分类模型，已经可以很好的在实验数据集上区分出“旁氏合约”，并且字节码的特征，具有更好的分类效果。如下图所示，在前十个重要特征中，只有一个 D_ind 属于交易特征，它是用来记录一个合约的所有参与者投资和回报次数的不同的特征向量。



四、 总结与展望

总结而言，这篇文章的核心创新点，是在一个新的应用场景下完成特定的欺诈检测。作者文章中的主要内容，在于特征的分析与构造，模型上选择了现成的结果，也达到了很好的效果。

由于可以说是一个开创性的工作，作者为基于以太坊的数据分析定义了很好的框架。可以预见的是，除了典型的旁氏骗局，以太坊中充斥着大量其他类型的欺诈骗局，我们可以从特征抽取和模型建立两个方面，着手数据分析工作。