Data mining for detecting Bitcoin Ponzi schemes

一、简介

本文尝试将数据挖掘的方法应用在比特币的区块链中，首先通过一系了启发式的手工方法，构造了一个“旁氏地址”的数据集，然后使用三种机器学习方法（基于规则，贝叶斯网络和随机森林）进行分类，最终可以在分类效果上取得比较好的结果。

相比较以太坊而言，在比特币的区块链中没有交易的概念，因此我们无法通过智能合约中的代码逻辑去标记旁氏骗局的产生。因此，作者首先通过启发式的方法人工去收集具备旁氏骗局的地址（人工去论坛与社区进行搜索），然后通过对这些旁氏骗局的地址的相关交易进行特征设计，最终完成分类任务，在本文中，作者仅收集到 32 个诈骗地址簇。

此处，值得借鉴的内容在于交易特征设计，主要有：

存在时间
活跃天数
最大交易/天
基尼系数
收到金额总数
转入交易数量
转入交易与转出交易比例
交易金额平均值
交易涉及账户
接收金额与返回金额的时间延迟
连续两天交易额的最大差额

交易特征目的是反映出旁氏骗局存在的典型特征（后来者无法获得任何回报）

二、结果

RIP: CM5

	P	nP
P	19	13
nP	7	6393

RIP: CM10

	P	nP
P	19	13
nP	7	6393

RIP: CM20

	P	nP
P	19	13
nP	7	6393

RIP: CM40

	P	nP
P	19	13
nP	7	6393

BN: CM5

	P	nP
P	24	8
nP	136	6264

BN: CM10

	P	nP
P	24	8
nP	155	6245

BN: CM20

	P	nP
P	24	8
nP	192	6203

BN: CM40

	P	nP
P	24	8
nP	213	6187

RF: CM5

	P	nP
P	25	7
nP	13	6387

RF: CM10

	P	nP
P	29	3
nP	26	6374

RF: CM20

	P	nP
P	31	1
nP	77	6323

RF: CM40

	P	nP
P	31	1
nP	132	6268

Figure 5: Confusion matrices of RIPPER, Bayes Net and Random Forest across different cost-matrices.

在本文中，随机森林的方法取得了最好的效果。需要注意的是，欺诈检测是一个典型的不平衡分类问题，因此作者采用了随机采样和代价函数修改的方法确保分类正确进行。从结果来看，召回率是一个不错的结果，随机森林可以找到 32 个中的 31 个诈骗地址。不过，还是有很大比例的误分类。

三、总结

本文是一个很好的开创性工作，主要的目的在于仅使用交易特征去尝试分析区块链中的旁氏骗局。目前，有关比特币交易中的相关交易的分析得到了比较多的研究，而以太坊引入智能合约之后，将会面临完全不一样的情况，这也是我们后续研究想要尝试一起考虑的内容。