Inline Text Wrapping Picture

北京邮电大学

硕士研究生学位论文开题报告

学 号: 2018111583

姓 名: 孙小文

学 院: 计算机学院

专业(领域): 计算机科学与技术

研究方向: 大数据技术与智能信息处理

导师姓名: 胡博

攻 读 学 位: 工学硕士

2019年12月25日

|  |  |  |  |
| --- | --- | --- | --- |
| 论文题目 | 面向交易维度的比特币地址关系分析算法的研究与实现 | | |
| 选题来源 | 企、事业单位委托项目 | 论文类型 | 应用研究 |
| 开题日期 | 2019-12-16 | 开题地点 | 北京邮电大学 |
| **一、立题依据（包括研究目的、意义、国内外研究现状和发展趋势，需结合科学研究发展趋势来论述科学意义；或结合国民经济和社会发展中迫切需要解决的关键科技问题来论述其应用前景。附主要参考文献目录）（不少于800字）**  **1、研究的目的与意义**  2008 年中本聪提出了比特币的理论[1]，2009 年比特币主网上的第一个区块被记录。2019 年是比特币诞生第 10 年，在各种新加密货币层出不穷的同时比特币市值遥遥领先，总市值最高曾达到3265.02亿美元[2]（其次的以太坊市值在百亿美元规模）。这说明比特币被广泛使用，并有较高认可度。  加密货币可以降低支付成本，提高金融系统运转效率和透明度、通过纯技术手段保障人们的信息安全（隐私）和财产安全[3][10]。但比特币仍是一个实验性的系统，也存在许多不足，区块链技术也是方兴未艾，有待更多的探索[10]。  比特币系统受人关注的一个特点是匿名性（但匿名性不是比特币系统的首要目标[17]），虽然所有交易信息都是公开的，但只要无法把交易和交易牵扯到的地址与现实中的人对应，那么系统还是匿名的。这个特点是一把双刃剑，它既为人们提供了保护自己隐私的方法，也给不法分子提供了躲避监管的手段。近年来加密货币被盗的事件时有发生，也有人利用加密货币进行传销、诈骗、敲诈勒索等犯罪活动，犯罪分子希望利用加密货币的匿名性保护自己的非法资金不受监管[2]。  许多研究表明，现在比特币用户身份会因容易受到某些攻击而被泄露[4][5]，尤其是现在比特币等加密货币支付还未普及，拥有比特币的人常需要（比如，通过交易所）把比特币换成人民币等法定货币，这既增加了个人隐私泄露的风险却也为监管机构追踪非法资金去向提供了可能[6]。  本论文希望系统通过比特币链上数据和一些可以公开获取的外部数据利用机器学习的方法设计和实现一种面向交易的分析比特币地址间关系的算法（下文中提到的交易混币鉴别也可以认为是广义的地址间关系或者说是地址间关系算法的扩展）。本论文的研究一方面可以为今后提升加密货币系统匿名性提供思路，另一方面也为监管机构追踪某些资金流向提供方法。  **2、研究现状和发展趋势**  比特币匿名化技术和去匿名技术都在不断发展。比特币交易信息完全公开，所以保障比特币用户隐私的重点是匿名技术。比特币地址由用户生成，且本身不包含身份标识，理论上无法与真实身份对应，但可以根据交易数据对交易进行追溯，而获得许多额外信息。根据交易之间的关系可以获得：   1. 任何一笔资金的流转情况。比特币通过每个区块的 coinbase （铸币）交易完成初始分发。任何人都可以追踪一笔资金从被矿工挖出以后的所有转账活动。 2. 任何地址的相关地址，因为发起交易需要使用地址的私钥签名。所以在同一个交易中作为输入的不同地址往往属于同一个实体。（但在混币交易中却不是这样） 3. 任何地址相关的交易。可以找到一个地址发起过多少笔转账，接收过多少笔资金。   在这种条件下混币是保证匿名性的重要手段，广义的混币就是把不在同一个人控制下的地址中的资金放到一个交易的输入进行混合，最终又分开，扣除手续费后回到了原来的控制者手中。而混币的类型主要分为中心化混币（依赖于第三方可信赖机构）例如：基于可信监管方的比特币强匿名混淆策略[5]、去中心化混币[4]，这两种混币技术各有优劣。  去匿名技术也有很多研究，例如，着眼于分析比特币地址和 IP 地址的关系，从而把匿名的比特币地址和实际的人和机构对应起来[8]；也有着重分析交易所内部地址和交易所与用户间关系的[6]。  有的研究把区块链上的隐私分为身份隐私和交易隐私——身份隐私指的是用户身份和地址间关系，交易隐私指交易记录和交易背后的知识[10]。  这些研究使用的方法主要是规则匹配和聚类。有的研究给出了根据链上数据得到一个交易所内部地址的步骤[6]。 有的研究则利用聚类方法，结合链上数据和在互联网上收集到的外部数据给比特币地址做分类[14] 。也有研究根据比特币交易可视化图形的模式对比特币交易进行鉴别[11]。  但目前利用机器学习尤其是监督学习的方法分析链上数据的研究仍比较少（有一些通过无监督学习分类比特币地址的研究，也有一个利用监督学习分类合法、非法交易的研究[32]），并得出比特币地址和交易相关信息的研究。这一方面是由于比特币链上数据不容易量化，而更容易应用定性分析得出结论。但也有一些比特币价格预测方面的研究，已经将监督学习应用到了比特币数据分析。  受到对于比特币交易可视化模式的研究[11]以及一些其他的比特币链上数据分析研究的启发，本论文希望能将监督学习应用在比特币地址和交易分析上。  **参考文献：**  [1] Nakamoto S. Bitcoin: A peer-to-peer electronic cash system[J]. 2008.  [2] 刘壮,袁磊.比特币研究新进展:属性、乱象与监管[J].金融与经济,2019(09):44-49+10.  [3] 李燕,马海英,王占君.区块链关键技术的研究进展[J].计算机工程与应用,2019,55(20):13-23+100.  [4] Ziegeldorf J H , Grossmann F , Henze M , et al. CoinParty: Secure Multi-Party Mixing of Bitcoins.[J]. 2015.  [5] 费天龙,郭静,鲁宁,史闻博.基于可信监管方的比特币强匿名混淆策略[J].中国电子科学研究院学报,2019,14(09):960-966.  [6] Hong S, Kim H. Analysis of Bitcoin Exchange Using Relationship of Transactions and Addresses[C]//2019 21st International Conference on Advanced Communication Technology (ICACT). IEEE, 2019: 67-70.  [7] https://www.kaggle.com/ellipticco/  [8] Koshy P , Koshy D , Mcdaniel P . An Analysis of Anonymity in Bitcoin Using P2P Network Traffic[J]. 2014.  [9] Mauro C , Kumar E S , Chhagan L , et al. A Survey on Security and Privacy Issues of Bitcoin[J]. IEEE Communications Surveys & Tutorials, 2018:1-1.  [10] 祝烈煌,高峰,沈蒙,李艳东,郑宝昆,毛洪亮,吴震.区块链隐私保护研究综述[J].计算机研究与发展,2017,54(10):2170-2186.  [11] McGinn, Dan, Birch, David, Akroyd, David,等. Visualizing Dynamic Bitcoin Transaction Patterns[J]. Big Data, 4(2):109-119.  [12] 孙坚. [基于经验模态分解方法的比特币价格波动分析及预测方法研究][D].湖南大学,2018.  [13] 艾青. 比特币交易趋势预测的研究[D].北京邮电大学,2016.  [14] Spagnuolo M . BitIodine: Extracting Intelligence from the Bitcoin Network[M]// Financial Cryptography and Data Security. Springer Berlin Heidelberg, 2014.  [15] https://www.walletexplorer.com/  [16] 毛洪亮,吴震,贺敏,唐积强,沈蒙.基于启发式的比特币地址聚类方法[J].北京邮电大学学报,2018,41(02):27-31.  [17] Reid F , Harrigan M . An Analysis of Anonymity in the Bitcoin System[J]. 2011.  [18] 李雪莲, 王海玉, 高军涛, et al. 一种匿名可撤销的比特币混淆方案[J]. 电子与信息学报, 2019, 41(8).  [19] https://en.bitcoin.it/wiki/CoinJoin  [20] http://zerocoin.org/  [21] Valenta L , Rowan B . Blindcoin: Blinded, Accountable Mixes for Bitcoin[C]// International Conference on Financial Cryptography and Data Security. Springer, Berlin, Heidelberg, 2015.  [22] Morenosanchez P, Ruffing T, Kate A. CoinShuffle: Practical Decentralized Coin Mixing for Bitcoin[M]// Computer Security - ESORICS 2014. 2014.  [23] https://en.bitcoin.it/wiki/PayJoin  [24] https://github.com/bitcoin/bips/tree/master/bip-0032  [25] https://en.bitcoin.it/wiki/Common-input-ownership\_heuristic  [26] Perozzi, Bryan, Al-Rfou, Rami, Skiena, Steven. DeepWalk: Online Learning of Social Representations[J].  [27] Zhou J , Cui G , Zhang Z , et al. Graph Neural Networks: A Review of Methods and Applications[J]. 2018.  [28] Wu Z , Pan S , Chen F , et al. A Comprehensive Survey on Graph Neural Networks[J]. 2019.  [29] https://research.huobi.cn/detail/23  [30] https://www.blockchain.com/btc/tree/492274005  [31] https://github.com/williamleif/GraphSAGE  [32] Weber M , Domeniconi G , Chen J , et al. Anti-Money Laundering in Bitcoin: Experimenting with Graph Convolutional Networks for Financial Forensics[J]. 2019. | | | |

|  |
| --- |
| **二、研究内容和目标（说明课题的具体研究内容，研究目标和效果，以及拟解决的关键科学问题。此部分为重点阐述内容）（不少于2500字）**  **1、研究内容**  本论文计划面向比特币交易，通过比特币地址向量化的方法表示比特币地址间关系密切程度，通过监督学习的方法进行比特币交易混币判断。使用到的数据包括比特币链上数据，以及互联网上的一些相关的公开数据。  要完成的功能是交易可视化、地址相关度判断和混币交易的鉴别。比特币系统中目前有近 6 亿个地址，它们之间的关系错综复杂。由于比特币地址是用户自主创建的，所以没有方法能够确切地知道一个地址对应的实体。虽然可以通过一些启发式规则，例如[25]，分析地址之间的关系，但是混币可以通过某种手段把不属于同一个所有者的地址下的资金放到同一个交易的输入中，这样会让外界误以为这些资金属于同一个人。本论文希望完成三个工作：   1. 一种比特币交易数据可视化方法，该部分贯穿整个研究过程，既帮助筛选特征，鉴别地址和交易关系也辅助解释模型输出结果和排查错误。其中追踪溯源算法也将在第三点的模型中帮助获得交易序列输入。 2. 一种比特币地址间关系判断方法，把比特币地址转化为一个固定维度的向量，这样可以通过不同地址向量间的夹角判断两地址关系的密切程度。 3. 一种基于监督学习的比特币交易混币判断模型，可以判断一个比特币交易序列是否存在混币行为。   未命名文件 (6)图 1 研究内容概览图  图 1底部的比特币链上数据和外部相关数据是原始数据，都是可以公开获取到的。中间两个数据集则是根据现有研究提供的方案（主要是规则匹配和聚类）筛选出来的。比特币链上数据可视化贯穿整个研究过程，为其他部分提供支持。   1. 比特币链上数据可视化   图 2 是来自McGinn, Dan, Birch, David, Akroyd, David 等人的研究中的比特币交易可视化的部分[11]，图中 A 部分是一个区块中的交易的展示，下面的 B、C、D、E 则是 4 种不同的交易类型，而其中 B 是普通交易，E 则被认为是混币交易，图中可以直观地看出混币交易和其他类型交易的差别。  截图_2019-11-19_15-29-27  图 2 McGinn 等人研究中的可视化效果  这一研究从宏观视角展示了不同类型交易的一些差异。本研究计划完成一个基于交易的比特币链上数据可视化程序。包括交易序列可视化程序和一个从单一交易出发，寻找该交易之前交易和后续交易序列的查找算法，也就是比特币交易溯源和追踪算法。  这部分工作是后续工作的基础。意义在于：1.帮助特征筛选，2. 辅助训练集数据的准确率的检验，3.帮助发现和纠正错误以及帮助解释模型结果，4. 交易追踪和溯源算法也提供了获取单个交易之前之后交易的数据，这将作为交易混币判断模型的输入。  这一部分有较多的可以参考的资料，例如 [11][30] 等。   1. 一种比特币地址间关系判断算法   有两种实现方案，可根据后续调研情况选择其中一种方案。  第一，通过监督学习的方法，根据已有研究创建数据集，筛选相关性很高的地址对或地址集合，再随机选择等量不相关的地址对或地址集合，得到数据集集。然后设计一个地址特征提取方案，训练模型，通过模型判断两地址的关系。  第二，使用图神经网络对比特币链上全部地址或选取部分地址，实现地址向量化，这样可以根据地址向量之间的夹角判断地址间关系是否密切，这就需要调查现有的图的向量化方法[26][27][28][31]。通过特征工程和数据预处理，从链上数据和其他相关数据中提取比特币地址相关数据。实现地址向量化。   1. 一种基于监督学习的比特币交易混币判断模型   需要调研现有技术方案，准备原始数据，再选取一部分数据得到一个比特币混币交易数据集，包括正样本（涉及混币交易）和负样本（非混币交易），然后设计训练和迭代模型。  关于原始数据，分为链上数据和其他公开数据，比特币的链上数据结构大致如图 3 所示：  图片1  图 3 比特币链上数据示意图  区块链是区块构成的链表，区块号从 0 开始，至今已有 60 多万块。每个区块中包含一到多个交易，每个交易中又包含了零到多个输入（coinbase 交易无输入），和一到多个输出。每个输入输出中包含了地址金额等字段。  除了原始数据外还可以计算一些统计值比如：序列中交易的数量、涉及地址数量、输入输出数量、输入输出平均金额、输入输出中金额的极差、输入输出金额的标准差等。  准备好数据以后，需要根据现有研究提供的方法筛选数据集，这里有比较多的研究可以参考，例如[4][18][19][20][22][23]。然后确认模型的输入输出。可能采用的输入是，交易向量或交易序列，输出则是一个 0 到 1 之间的浮点数，表示目标交易涉及混币的概率，0 表示不涉及混币，1 表示是混币。然后设计并训练模型。  **2、研究目标与效果**  针对上述三项具体研究内容，对应的研究目标和效果如下：  （1）比特币链上数据可视化  追踪溯源部分可以根据单个地址查找交易序列。链上数据可视化部分可以按照地址视角和交易视角展示比特币交易序列。能够体现出交易金额多少，地址间金额转移关系等。  具体来说，在交易视角下可以看到每一笔交易的后续交易或者看到这笔交易的输出尚未花费，可以看到每一笔交易的是否混合了交易树以外的交易的输出。  在地址视角可以看到每个交易输出中金额多少，其中包含了多少金额被认为是来自于起始交易的，又是否含有来自这个交易树外部的资金。   1. 一种比特币地址间关系判断算法   实现一种算法可以判断两个地址间关系是否密切。算法的输入是两个比特币地址，输出是这两个地址关系密切程度，比如如果这两个地址属于同一个所有者，那么这两个地址的密切程度为 1。如果这两个地址没有任何关系，它们的密切程度应该为 0。  计划根据现有研究，通过规则匹配方法，筛选出高相关度地址，再随机选择不相关的地址构成一个地址相关度数据集，可以在这个数据集上测试该地址间关系判断算法的效果。  由于比特币地址数量较多，本论文可能会选择一部分地址来实现该算法。   1. 一种基于监督学习的比特币交易混币判断模型   通过特征工程和数据预处理，把任意比特币交易特征转化为向量或者序列，以便于在模型中使用。  根据现有研究提供的方法，筛选比特币区块链上涉及混币的典型交易，再随机选择等量其他类型的交易，构成鉴别比特币交易混币模型的数据集。具体要求有：交易模型训练集要求正负样本数量不低于 2 千个（比特币总交易数量 5 亿多），要有足够准确率。  把训练数据集按比例划分为两部分，一部分用于模型训练，一部分用于模型评估。可能使用不同方法训练多个模型进行比较，也可能会使用多个模型进行模型融合。  模型需要有较好的性能，可以有效鉴别特定的混币交易（拟针对 CoinJoin 和 PayJoin 这两种应用普遍的混币交易）。可以使用准确率和召回率衡量模型的性能。   1. **拟解决的关键问题**   （1）比特币链上数据向量化的方法  既要尽可能多的提取比特币地址、交易的有效信息，又要尽量降低维度。而且需要考虑如何处理很多数量不确定的值，例如特定交易的后续交易序列如何表征。还有构造训练数据集合的时候如何根据交易可能涉及混币的程度合理的设置输出值。   1. 设计有效的模型用于鉴别混币交易   需要在数据量化的基础上选择合适的模型和恰当的参数。也可能需要多个模型进行融合。需要达到较高准确率。   1. 比特币链上数据可视化   就目前的研究来看比特币交易可视化可以有多种维度，例如交易维度、地址维度还有一些研究把交易和地址融入一张图中。但是由于比特币采用 UTXO 模型，总是有一些难以表现出来的东西。本论文的可视化部分将为特征选取、训练数据集评估已经模型评估做支撑。所以需要明显的体现两个维度的特征，第一是需要在地址维度展现两个地址间关系，第二个是在交易维度，展现一个交易以及其后续交易的情况。 |

|  |
| --- |
| **三、研究方案设计及可行性分析（包括：研究方法，技术路线，理论分析、计算、实验方法和步骤及其可行性等）（不少于800字）**  **1、研究方法**  本论文主要采用调查研究、算法设计与实验验证相结合的研究方法。具体的方法如下：  （1）调查研究  模型设计与训练之前需要了解比特币区块链上数据涉及的内容以及其含义和相互间关系。要了解比特币历史、重要变动、重大事件等内容。了解区块链技术发展历程。需要根据地址间关系和交易混币两个不同的场景选择合适的模型。调查现有模型适用程度和性能效果等。   1. 算法设计   地址向量化表示需要从比特币原始数据中提取地址间关系数据。根据地址间关系数据构建地址向量。  交易混币模型需要构建数据集。参照现有研究提供的方法，寻找混币交易。再选取非混币交易作为负样本。最后抽样调查并利用可视化工具计算数据集准确率。  因为目前没有找到已发表的用于比特币链上数据分析比特币混币判断算法，所以可以借鉴其他领域类似场景的模型设计方法， 例如[32]。  首先确定模型的输入输出，输入主要由原始数据预处理和向量化决定，输出则是概率值。然后选择合适的编程语言和框架。最后设计模型，训练模型，验证和评估模型，然后不断迭代，直到得到满足要求的版本。  （3）实验验证  首先把数据集按照比例划分为训练集和测试集。模型训练完成后使用测试集检验模型准确率和召回率。针对错误数据可以随机抽取少量判断错误的数据，利用可视化工具分析查找判断错误的原因。  **2、技术路线**  本论文拟采取图 4 所示的技术路线：  技术路线  图 4 本论文技术路线示意图    关于构建和训练模型，具体步骤为：  第一步，获得原始数据，主要需要使用两部分数据：  一部分是比特币链上数据，这部分数据是任何人都能公开获取到的。截至目前比特币链上数据的数据量大致是这样的：60 多万区块（平均每 10 分钟产生一个新区块），5 亿多笔交易（现在的每个区块大概上千条交易），近 6 亿个地址。  另一部分是外部数据，目前了解到的外部数据有：历史币价数据、比特币骨干节点 IP 地址及地理位置分布。比特币交易首先播报方 IP（这个数据不是完全准确的，只是一个参考值，是某些拥有大量骨干节点的机构统计发布的，但是这只是它们的节点第一次接收到这个交易的发送方地址）、比特币交易量，另外还有一些非结构化数据，比如比特币论坛上的讨论信息。  第二步，数据可视化。  实现地址视角和交易视角两种维度的比特币链上数据可视化。  第三步，构造数据集。  已经有研究使用规则匹配的方法寻找地址间关系，最广为人知的是共同输入法（Common-input-ownership heuristic）[25]，它可以根据交易信息发现某些地址属于同一个人（但是在混币的条件下这不准确）[16][25]。还有一些研究通过聚类等无监督学习做地址分类，例如 Spagnuolo M 的研究提到了一种利用聚类给比特币地址分类的方法[14]。而混币数据集则可以根据现有研究提供的方法，例如，Ziegeldorf J H , Grossmann F , Henze M 等人的研究指出了一些已经得到应用的混币方法的特征[4]，还有另外一些混币方面的研究提供了某些特定混币的特征，例如[4][18][19][20][22][23]。可以根据这些特征筛选出典型的混币交易作为正样本，然后选择一些其他类型的交易作为负样本，构成数据集。  第四步，特征向量化。  把非结构化数据和长度不定的数据转化成定长向量或者定长向量的序列。这里既要注意保留关键信息，又要尽量降低维度。  第五步，训练模型。  模型训练技术路线  图 5 模型训练技术路线示意图  如图 5 所示，首先需要调研相关领域或者类似场景下的主流方法。然后可以借鉴主流方法设计和训练模型，然后使用筛选得到的数据集评估模型效果，当模型达不到要求时需要进行有针对性的迭代。   1. **可行性分析**   本论文的工作第一是比特币可视化，这方面已有较多研究可供参考[11]，也有一些公开的系统可以学习[30]。第二是地址向量化表示。可以把比特币数据看成是地址为顶点，交易为边的网络。这样可以使用一些图神经网络算法[26][31]完成节点向量化工作。第三使用监督学习的方法训练模型鉴别混币交易，有利条件有：  第一，训练数据充足，比特币的链上数据是公开的结构化数据集，再加上互联网上有很多比特币相关的分析数据，例如地址标签信息，交易播报 IP 信息，比特币相关论坛，公开的交易信息和公开的地址，已经识别的混币行为。而且使用规则匹配的混币检测方法已经有了较多的研究。而地址关系方面，类似于 [15] 这样的根据共同输入法（Common-input-ownership heuristic）[25] 归纳地址间关系的方法已经非常成熟，该网站目前已经公开了它们分析的几千万个属于同一实体的地址集合（称为钱包）。而且比特币区块链上目前有丰富的地址和交易（均超过 5 亿），而且还在不断增加。所以可以得到充足的训练数据。  第二，混币交易有比较明显的特征，与其他多种类型的交易有比较明显的差异，可以参考McGinn, Dan, Birch, David, Akroyd, David 等人的研究[11]，在它们发表的文章中可以看到，典型的混币交易和普通交易以及中心支付交易（如薪酬发放），交易所日常交易等其他交易在多个方面有显著差异，比如，交易涉及的地址数量、交易的频率等，而且还有很多其他研究揭示了一些特定混币行为的特点和固定模式。而地址之间的相关性也有很多规律可循。  第三，可以使用量化数据表示交易和地址，原始数据中本身已经有很多字段是数量特征，其他一些非数量特征也可以转化为数量特征。 |

|  |
| --- |
| **四、本研究课题可能的创新之处（不少于500字）**  1、设计和实现一种比特币地址向量化表示方法  比特币系统中地址数量非常庞大，有近 6 亿条，比特币地址还在不断增加。目前判断比特币地址间关系的方法主要是共同输入法（Common-input-ownership heuristic）[25] 等一些启发式方法，可以把地址归纳为钱包，另外还有一些聚类方法。2013 年提出的BIP 32 [24] 建议使用一次性地址提高比特币系统匿名性，遵从这一建议的钱包会为每一次交易创造一次性地址，这些一次性地址会给比特币链上数据分析带来困难。地址聚类的准确率不太高[29]，启发式规则不能考虑混币的情况，遇到混币就会出现错误。本论文将设计和实现一种比特币地址向量化表示方法，根据地址间关系把所有比特币地址映射到一个较低维度的向量空间中。既能考虑混币的影响，又能定量的描述地址间相关度，而且可以把方便在机器学习中框架中使用地址特征。   1. 设计和实现比特币交易混币判断模型   模型的功能是鉴别交易是否涉及混币。目前的混币鉴别主要是根据已知混币类型的模式进行特征匹配。实际上，得到应用的混币技术有多种类型，例如[4][18][19][20][21][22][23]，且常有新的方法出现。已经有一些相关研究给出了混币的行为模式和鉴别方法，但这些方法难以适应混币方法的不断升级，往往无法对新出现的混币方法做出有效鉴别，因为这些新的混币方法通常是针对已有鉴别方法做了特别改进。而且有某些混币交易与普通交易从交易本身的特征来看是几乎没有任何区别的，例如[23]，仅仅从一个交易本身的信息无法判断其具有混币行为。  通过监督学习的方法，第一，不仅仅局限单个交易本身的特征，还会考虑交易之前之后的一个交易序列的特征，以及参与交易的地址的特征。第二，可以从大量确定的混币案例中学习到混币交易，以及相关交易序列、参与交易的地址的特征，对新类型的混币方法能有更好的识别效果。 |
| **五、研究基础与工作条件（1.与本项目相关的研究工作积累基础 2.包括已具备的实验条件，尚缺少的实验条件和拟解决途径）（不少于500字）**  1、与本项目相关的研究工作积累  （1）在项目开展前期通过阅读一些机器学习和深度学习方面的资料，了解了一些机器学习算法的工作原理和各自的优劣，在前期调研工作中，对机器学习的工作流程以及评价指标也有了一定的认识。  （2）通过阅读文献和源代码了解了以比特币为代表的加密货币和区块链的基础理论知识和运行原理。对比特币链上数据的结构有了一定的了解。  （3）已开始开发比特币交易序列可视化程序（交易、地址视角）  2、已具备的实验条件  本课题依托于北京邮电大学网络技术研究院网络与交换技术国家重点实验室宽带网中心。实验室科研实力雄厚，可以为本项目提供必要的软硬件支持和实验环境，包括提供计算功能强大的多台PC机和服务器集群，以及研究所需的实验数据获取平台等。  已具备的条件：  （1）实验所需数据集：比特币链上数据、2011 年以来的比特币币价数据、比特币公网主要节点的 IP 地址和地理位置  （2）数据存储软件：比特币链上数据存储软件MongoDB  （3）算法运行平台：多台Windows/Linux操作系统PC机，Linux操作系统服务器集群  3、尚缺少的实验条件和拟解决途径  尚缺少的实验条件：  第一，MongoDB 作为一种现代的、被广泛应用的 NoSQL 数据库，对大数据存储有良好的支持，但它的功能侧重于 OLTP，是作为事务型应用后端数据库的良好选择，而本论文的研究涉及到许多 OLAP 任务，这并不是 MongoDB 所擅长的领域。所以需要更高效的专用的计算框架做数据预处理工作。  第二，目前掌握的外部数据仅有2011 年以来的比特币币价数据、比特币公网主要节点的 IP 地址和地理位置等数据集，相对于需求仍不够充足。  拟解决途径：   1. 开发专用的程序做数据预处理工作。 2. 计划需要的数据，编写爬虫程序获取相应数据 |

**学位论文工作计划**

|  |  |  |
| --- | --- | --- |
| 时间 | 研究内容 | 预期效果 |
| 2019.09-2019.11 | 相关算法调研工作 | 了解相似场景常用方案 |
| 2019.12 | 论文开题 | 撰写开题报告，完成开题答辩 |
| 2019.12-2020.1 | 获取外部数据，数据预处理，得到数据集，可视化程序 | 得到数据特征提取方案，和两个数据集（地址和交易），完成可视化程序 |
| 2020.2-2020.2 | 设计和实现地址向量化方案 | 完成地址向量化方案 |
| 2020.3-2020.4 | 设计交易混币判断模型 | 完成模型的设计和训练 |
| 2020.5-2020.5 | 模型验证 | 模型性能符合预期 |
| 2020.6-2020.8 | 完成小论文 | 完成小论文 |
| 2020.9-2021.9 | 撰写阶段报告 | 完成阶段报告 |
| 2020.10-2020.12 | 撰写毕业论文 | 提交初稿、修改稿、定稿 |
| 2021.1-2021.5 | 论文送审、修改及答辩 | 论文送审、修改及答辩 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 评  定  小  组  成  员 | 姓 名 | 职 称 | 单位名称 | 职务 |
| 金跃辉 | 教授 | 北京邮电大学 | 组长 |
| 崔毅东 | 副教授 | 北京邮电大学 | 成员 |
| 胡博 | 教授 | 北京邮电大学 | 成员 |
| 杨谈 | 副教授 | 北京邮电大学 | 成员 |
|  |  |  |  |
|  |  |  |  |
| 导师意见： | | | | |
| 同意论文开题 | | | | |
| 导师（签名）：  日期： 年 月 日 | | | | |
| 开题报告小组意见： | | | | |
| 组长（签名）：  日期： 年 月 日 | | | | |
| 学院意见（签章）： | | | | |
| 负责人：  日期： 年 月 日 | | | | |