

面向金融活动的复合区块链关联事件溯源方法

李 素 宋宝燕 李 冬 王俊陆

辽宁大学信息学院 沈阳 110036

(liisuu@163.com)

摘 要 现有区块链系统多采用平等挖矿模式,所有记账人(实体)将账本记录在单一主链上,数据存储具有随机性,且在复杂或分类金融场景下,主链数据难以关联或规律存储,导致存储及查询效率很低;同时,现有区块链系统中事件溯源大多只查询到源区块,不能判识实体间的隐含关联,查询具有局限性。针对这些问题,提出一种复合区块链关联事件溯源方法。该方法首先构建区块链复合链式存储结构模型,提出私有链和联盟链的概念,实现复杂或分类场景下的自适应数据关联存储;然后,追溯查询时,在获取到事件源实体区块的基础上,建立辅助存储空间,以便对相关数据进行转存,并提出基于 Apriori 算法的关联实体区块溯源方法,将所得的溯源实体区块用于构造源事件关联图,进而描述事件实体间的关联关系;最后,提出基于强化学习的风险评估体系,实现对溯源实体的风险评估。实验结果表明,复合区块链关联事件溯源方法可降低 60% 的存储开销,查询准确率和安全性分别提升 90% 和 50%。

关键词: 复合链式结构;私有链;联盟链;关联事件追溯;强化学习

中图法分类号 TP311

Composite Blockchain Associated Event Tracing Method for Financial Activities

LI Su, SONG Bao-yan, LI Dong and WANG Jun-lu

School of Information, Liaoning University, Shenyang 110036, China

Abstract The existing blockchain system mostly adopts the equal mining mode. All bookkeepers (entities) record the books on a single main chain, and the data storage is random. Moreover, in complex or classified financial scenarios, the data of the main chain is difficult to realize association or regular storage, leading to low storage and query efficiency. At the same time, most event traceability in the existing blockchain system can only query the source block, and the implied association between entities cannot be identified, so the query has limitations. To solve these problems, composite blockchain associated event tracing method is proposed. Firstly, the composite chain storage structure model of blockchain is constructed, and the concepts of private chain and alliance chain are proposed to realize the adaptive data association storage under complex or classified scenarios. Then, in the traceability query, on the basis of obtaining the event source entity block, the auxiliary storage space is set up to transfer the relevant data. A tracing method of the associated entity block based on the Apriori algorithm is proposed, and then the obtained traceability entity block is constructed as the source event correlation graph to describe the correlation relationship between the event entities. Finally, the risk assessment system based on reinforcement learning is proposed to realize the traceability entity risk assessment. Experiments show that the composite blockchain associated event tracing method can reduce the storage overhead by 60%, improve the query accuracy by 90% and improve security by 50%.

Keywords Composite chain structure, Private chain, Alliance chain, Associated event tracing, Reinforcement learning

1 引言

区块链^[1-2]是一种在不可信的竞争环境中低成本建立信任的新型计算范式和协作模式,由于其具有存储密度大^[3]、防

篡改^[4]、可追溯等特性,基于区块链技术的应用越来越广泛。区块链通过增加区块的方式进行数据存储^[5],且数据一般都是存储在单链上,但随着时间的推移和交易数据的扩大,会引发数据膨胀^[6],并且可能导致存储和查询效率降低。同时,

到稿日期:2021-07-07 返修日期:2021-10-16

基金项目:国家自然科学基金(62072220, 61502215);中国博士后基金面上项目(2020M672134);辽宁省教育厅科学研究项目(LJC201913, LJKZ0094)

This work was supported by the National Natural Science Foundation of China(62072220, 61502215), China Postdoctoral Science Foundation Funded Project(2020M672134) and Scientific Research Project of Liaoning Province Education Department(LJC201913, LJKZ0094).

通信作者:王俊陆(wangjunlu@lnu.edu.cn)

单链存储模式也不能实现复杂或分类场景下的关联或规律存储^[7]。在溯源查询时,现有区块链系统多以时间戳和哈希指针为线索^[8-9],通过查找区块中的交易来实现溯源到某一区块上的某一交易数据,难以实现溯源查询出区块链中各实体之间的隐含关系,对后续的事件分析支持不足。

例如,在金融活动区块链系统中,各金融企业(实体)多采用平等单链模式^[10-12]进行数据存储,这会导致金融企业(实体)的交易数据无序且具有随机性^[13],查询效率低^[14],不利于后期维护。同时,单链模式存储各金融企业(实体)的交易数据,当有关联关系的实体间(如企业的总部和分部)进行坏账及不良资产转移操作时,现有金融区块链的溯源查询只能查到分实体信息,无法关联追溯到总实体,导致查询准确性很低。因此,如何建立高效的区块链存储模式并实现关联事件溯源,一直是金融区块链领域研究的热点和难点^[15-17]。

针对上述问题,本文提出了一种面向金融活动的复合区块链关联事件溯源方法,主要贡献如下:

(1)针对目前区块链存储结构单一所造成的存储及查询效率低的问题,提出了区块链复合链式存储结构模型,结合区块链不同应用模式以及存储金融实体特性,在不同实体间构建联盟链,在实体内部构建私有链。

(2)针对目前区块链单链追溯模式造成溯源准确率很低的问题,提出了基于 Apriori 算法的关联实体区块溯源方法,建立辅助存储空间,将联盟链上与事件相关的实体交易数据进行转存,引入 Apriori^[18-19]关联规则算法,获取与事件源实体存在关联关系的其他实体区块数据。

(3)在此基础上,对所有溯源出的实体区块数据建立源事件关联图,表示事件实体间的关联关系,将关联图中实体的区块数据转换成数值特征向量,并构建基于强化学习的风险评估体系,实现对溯源实体的风险评估。

2 相关工作

目前,许多专家学者对区块链存储及溯源查询进行了深入研究,并取得了一定的研究成果。

在区块链存储方面,文献[20]提出了一种利用区块链技术存储数据的方法,该方法验证了区块链存储数据密度大、可追溯、防篡改的特征,但没有提出具体的存储机制;文献[21]对现有区块链系统中的数据存储限制进行了阐述,指出系统存在查询功能简单、查询性能较低等问题,但没有给出具体的解决方法;文献[22]提出了一种单链模式区块链结构存储数据的方法,该方法存储效率高,但在复杂金融场景下不能准确获得实体间的关联关系;文献[23]提出了一种建立多叉链式结构存储数据的方法,该方法可以存储复杂庞大的数据,但多叉链式区块链结构存储数据会使存储和查询效率降低。

在区块链溯源查询方面,文献[24]提出了一种高效检索区块链的方法,包括范围查询和 Top-*k* 查询,有较好的灵活性,但是其解决方法是通过把区块链数据以及 key-value 数据库中的数据同步到另一个数据库,无法从根本上解决区块链查询效率低的问题;文献[25]提出了一种单链模式的区块链

追溯方法,可以高效实现基础查询应用,但在有关联和隐含关系的复杂场景下,无法实现相关实体的追溯查询;文献[26]提出了一种多叉链模式的区块链追溯方法,该方法可以实现复杂的追溯查询,但由于多叉链存储结构复杂,导致查询效率低;文献[27]提出了一种基于数据挖掘技术的数据关联查询方法,可以实现复杂场景下的关联查询,但该方法在区块链数据模型中存在不一致的问题;文献[28]提出了将带有唯一标识的数据添加到区块链上解决数据溯源的问题,但标识数据的准确性无法得到保证。

综上所述,现有方法在区块链存储和溯源查询方面存在存储效率低、存储结构复杂及溯源查询不准确等问题。因此,本文综合考虑存储和溯源查询的效率和准确性问题,提出了一种面向金融活动的复合区块链关联事件溯源方法。

3 复合区块链模型构建

针对复杂或分类金融模式下实体数据来源广泛、数量巨大的特征,建立基于 Merkle 树的私有链区块结构和基于 Merkle Patricia 树的联盟链区块结构,构建复合区块链模型,实现复杂或分类场景下的自适应数据关联存储。

复合链式结构模型由私有链和联盟链组成,私有链构建于实体内部,用于表示实体的交易信息,在私有链的基础上,进行实体间联盟链的构建,形成复合链式结构模型。实体的用户与所在的私有链进行交互,其他私有链对用户透明,实体用户的需求由私有链进行相应操作实现。复合链式结构模型示意图如图 1 所示。

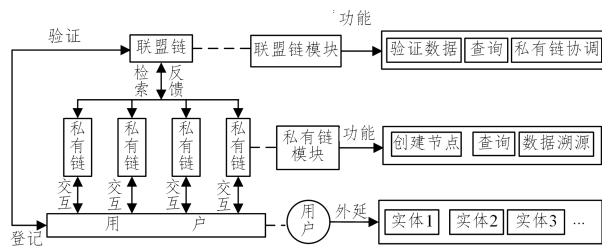


图1 复合链式结构模型

Fig. 1 Composite chain structure model

首先,构建每个实体的内部私有链。在该私有链网络上,根据实体的交易频率设定一个周期,每个周期内该实体发生的交易为一个独立节点,把从实体采集、认证及更新的交易信息按日期进行分类,将所有的交易数据通过哈希算法进行哈希运算生成相应的、唯一的 Hash 值,通过密码学算法对数据进行加密生成密钥对,盖上时间戳,存储在每一个交易节点的本地账本中。通过网络广播和共识机制验证,单个实体内部私有链网络的每个节点将记录所有交易类型及其对应的交易信息,自适应地构建实体内部私有链。

然后,构建实体间的联盟链。所有的实体通过协议达成统一的区块链技术标准和统一的行业准则,以保证参与成员的行为及技术的规范性,并基于达成的协议和实体数据量,在不同的实体间自适应地构建联盟链。在该联盟链上,每一个实体作为一个独立的区块链大节点,其内部私有链网络互相

连接,共同组成联盟链网络。实体采集认证的交易信息,在内部私有区块链认证通过后,再通过联盟链分布存储在每一个实体节点的本地账本中,基于此,整个联盟链的所有实体节点中将存储每个实体采集的交易信息,最后实现实体间交易数据信息的互联共享。在该复合链式结构中,数据都存储于区块链底层区块中,区块不仅用于存储数据,还用于验证交易信息的有效性和生成下一个区块,结合不同的存储和查询需求,本文提出了私+联的复合链式结构进行数据存储。

3.1 基于 Merkle 树的私有链区块结构

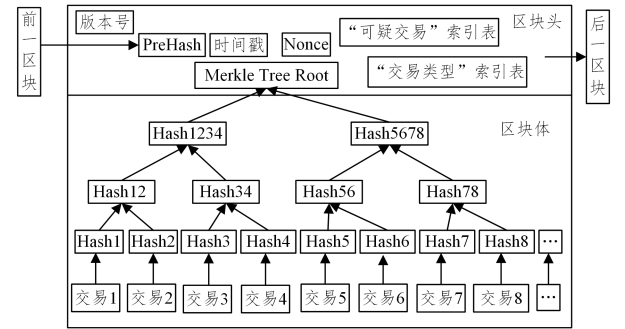
私有链采用基于 secp256k1 椭圆曲线数学的数字签名算法(Elliptic Curve Digital Signature Algorithm, ECDSA)生成两个不同的密钥(公钥和私钥),私钥将数据加密,在需要验证该交易数据时用公钥解密。每个区块由区块头和区块体两部分构成,通过将上个区块哈希、根哈希和随机数等信息进行哈希运算生成当前区块的哈希值,每个区块的前块哈希指针按照时间顺序链接构成整个私有链的链接关系,私有链区块头数据信息如表 1 所列。

表 1 私有链区块头数据信息

Table 1 Private chain block header data information

属性	含义
版本号	数据区块的版本号
前块哈希(Prev Hash)	对上一区块的 Merkle Root 和时间戳等数据进行 Hash 运算得到的 Hash 值
时间戳	该区块生成的近似时间
随机数(Nonce)	当前区块共识过程的解随机数
Merkle 树根(Root Hash)	区块中的所有交易通过哈希运算得到的 Merkle 树的根
交易类型索引表	记录该区块的交易所属的交易类型
可疑交易索引表	记录可疑交易的 Hash 值

区块体存储全部交易信息,每一笔交易的信息通过哈希函数把交易数据转换成一串唯一的哈希值存储在 Merkle 树的叶子节点上。通过哈希函数逐层生成上层节点的哈希值,每个区块数据集对应一个唯一的哈希根。如果底层交易记录被篡改,则 Merkle 树根的值也会发生变化,私有链区块结构示意图如图 2 所示。



在交易数据存入区块体前,首先需统计实体开展的全部交易类型并统一编号。在交易数据存入区块体的同时,在区块头的“交易类型”索引表中添加该交易的交易类型信息,发生非法金融活动需溯源查询时,“交易类型”索引表用于查询需要追溯的交易数据的交易类型。然后制定可疑交易规则,

在交易数据即将存入区块体时,根据可疑交易规则判断该交易是否属于可疑交易,若属于可疑交易,计算该交易数据的 Hash 值后,把 Hash 值存入 Merkle 树上的同时也把 Hash 值存入区块头的“可疑交易”索引表中,在数据溯源查询时,从“可疑交易”索引表开始检索,可提高查询效率。

3.2 基于 Merkle Patricia 树的联盟链区块结构

如 3.1 节所述,每一个实体私有链都将作为联盟链中的一个账户,实体之间通过签订合同的形式建立交易往来的联系,建立基于 Merkle Patricia 树的联盟链区块结构进行数据存储。联盟链区块结构示意图如图 3 所示。联盟链同样采用 ECDSA 算法对数据进行加密。区块头中的 3 个 Merkle Patricia 树的根哈希分别对应状态树、交易树和收据树,交易信息存储在区块体中。

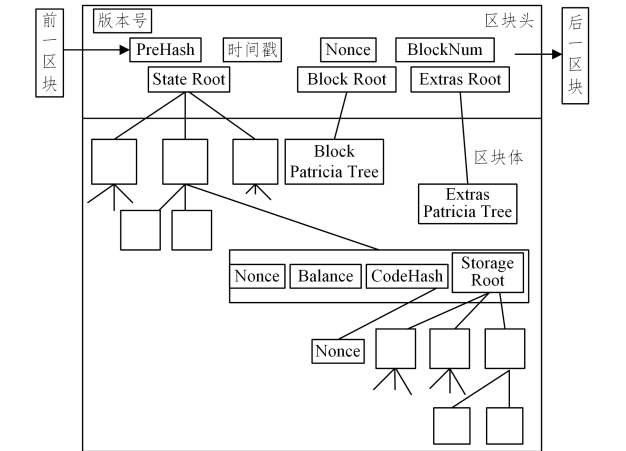


图 3 联盟链区块结构

Fig. 3 Alliance chain block structure

在联盟链中建立 3 个 levelDB 数据库,分别是 BlockDB, StateDB 和 ExtrasDB。其中,BlockDB 存储区块头和交易记录,StateDB 存储实体的状态数据,ExtrasDB 存储实体间签订的合同信息,并基于此构建联盟链的底层数据库。每个 Merkle Patricia 树的存储内容和功能如表 2 所列,每个区块都包含整个状态树的根哈希,其中状态树以周期 T 进行更新。

表 2 联盟链的 3 个 Merkle Patricia 树

Table 2 Three Merkle Patricia trees of alliance chain

标识符	状态树	交易树	收据树
Key	账号地址	交易编号	索引编号
Value	账户内容	交易内容	合同内容
存储数据库	StateDB	BlockDB	ExtrasDB
唯一性	区块链整体一棵	每个区块独立一棵	每个区块独立一棵
所支持的查询	实体的交易状态和实体是否存在	交易是否存在于区块中	地址的事件实例

4 关联实体溯源机制

在建立复合链式结构的基础上,当非法金融活动事件发生时,首先查询事件源问题实体区块交易数据,然后引入 Apriori 关联规则算法,获取关联实体区块交易数据,进而对所有溯源查询出的实体区块数据建立源事件关联图,最后构建基于强化学习的风险评价体系,实现对溯源实体的风险评估。

4.1 关联实体区块溯源

复合链式结构建立后,每个实体的全部交易信息都存储在实体内部私有链上。非法金融活动事件的源问题实体是由于其将交易数据存储在链上后,非法对账本进行修改,造成链上数据信息和链下自身账本不一致或在某一段时间内实体的交易出现异常等问题。在对事件源问题实体溯源时,以时间戳和前块哈希指针为线索,对区块链中的交易进行公钥解密后,通过逆哈希算法转换为存入私有链前的数据格式,与实体本身交易账本对比,出现不一致的交易数据就是事件源实体的问题交易数据。具体分为以下3种情况:

(1)问题交易可能出现在实体私有链区块头“可疑交易”索引表的交易中;

(2)若在“可疑交易”索引表中没有发现问题,则问题交易可能出现在实体私有链区块体的交易中;

(3)若在实体私有链区块体的交易中没有发现问题,则问题交易就会出现在某个实体私有链某一段时间内的交易中,对实体私有链上的交易数据进行横向观察,如果在某一段时间内的交易比较频繁、异常,或者交易金额过大,则该时间段内的交易即为事件源问题实体出现的问题交易数据。

在找出事件源实体出现问题的交易数据的基础上,需进一步查询出与事件源实体区块问题交易数据有隐含关系的其他实体区块交易数据。关联实体区块溯源方法流程如下:

Step1 建立辅助存储空间存储联盟链中与事件相关的实体交易数据。在找到事件源实体区块问题交易数据后,从该实体私有链区块头的交易类型索引表中查询出问题交易的交易类型,将联盟链收据数据库中与该事件源实体有合同往来的其他实体全部找出,查找这些实体私有链上与该事件源实体区块问题交易数据交易类型相同的其他实体私有链的交易数据,将上述交易数据经过解密和逆哈希运算存入辅助存储空间中。

Step2 关联实体区块溯源方法的相关定义和规则。设 $I = \{i_1, i_2, \dots, i_m\}$ 是辅助存储空间中所有实体的交易项目集合,事务数据库 $D = \{t_1, t_2, \dots, t_n\}$ 是实体的 TID 标识组成的事务,每个事务 $t_i (i=1, 2, \dots, n)$ 是 I 中的一组项目集合,也就是该实体中的所有交易,即 $t_i \in I$,其所含项目个数定义为该事务的长度。

首先,找出频繁项目集,设 D 为事务数据库, I 为全局项目集,若项目集 $L \in I, L \neq \emptyset$ 并且 L 的支持度不小于最小支持度,即 $\text{support}(L) \geq \text{Minsupport}$,则 L 称为频繁项目集,以下约定 L_k 表示 k -项频繁集。

然后,确定 Apriori 关联规则,关联规则的蕴含式为 $X \Rightarrow Y$,其中, $X, Y \subseteq I$ 且 $X \cap Y = \emptyset$ 。 X 称为前因(antecedent), Y 称为后果(consequence),上述关联规则表示 X 会导致 Y 。 $\text{count}(X)$ 为项集 X 在 D 中出现的次数, $\|D\|$ 为事务数据库记录总数。若关联规则 $X \Rightarrow Y$ 成立,则有:支持度 $\text{support}(X \Rightarrow Y) = \text{count}(X \cup Y) / \|D\|$ 和置信度 $C\% = \text{confidence}(X \Rightarrow Y) = P(Y|X) = P(X \cup Y) / P(X)$,即数据库 D 中包含了 X 记录的同时至少有 $C\%$ 也包含 Y 。

Step3 利用关联实体区块溯源方法找出关联的实体区块数据。根据实体数据特性设定支持度的一个阈值为最小支持度(Minsupport)和置信度的一个阈值为最小置信度(Min-confidence)。通过逐层搜索迭代的方法,先产生候选集 C_k ,计算所有 k 项集的支持度,若支持度满足大于最小支持度阈值的要求,则成为频繁 k 项集 L_k ,然后在 L_k 基础上产生候选集 C_{k+1} ,通过判断最小支持度来确定频繁 $(k+1)$ 项集 L_{k+1} ,以此类推,直至找不到下一个频繁项集为止。由每个频繁项集中的项组成关联规则,分别计算出规则的置信度,若置信度大于最小置信度阈值,则生成关联规则,同时满足最小支持度和最小置信度要求的规则为强关联规则,说明该项目交易数据与事件源实体出现的问题区块交易数据关联程度高,即确定该交易所在的实体与事件源实体有隐含的关联关系。

在获取所有与事件源实体区块问题交易数据有隐含关系的其他实体区块交易数据的基础上,按交易数据与实体一一对应的关系,构建基于交易时间的源事件关联图。源事件关联图由多个事实组成,每个事实可表示为一个 SPO 三元组 (Subject-Predicate-Object),具体为(实体1,关系,实体2)、(实体、属性、属性值)两个三元组表示事件中实体和交易的关系。实体和实体之间采用(实体1,关系,实体2)表示,实体与其交易采用(实体、属性、属性值)表示。源事件关联图的示意图如图4所示。

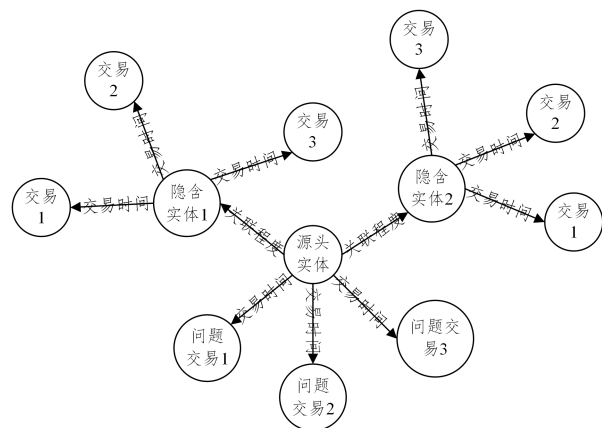


图4 源事件关联图

Fig. 4 Source event association diagram

在建立源事件关联图时,首先建立源实体及其问题交易数据信息,然后建立与源实体有隐含关系的其他实体。源实体和隐含实体间的连线表示两者的关联程度,关联程度由关联实体区块溯源方法计算出的支持度和置信度确定,支持度和置信度越高,关联程度就越高。在此基础上建立隐含实体中与事件有关的若干交易,通过该关联图可具体地表示出事件中实体间的关联关系。

4.2 基于强化学习的风险评价体系构建

如4.1节所述,在查找出与该事件有关的若干交易数据的基础上,将源事件关联图中所有实体区块数据表示成数值形式的特征向量,对源事件关联图中的实体构建基于强化学习的风险评价体系,实现对溯源实体的风险评估。风险评价

体系基线模型 (Risk Assessment System Baseline Model, RASBM)示意图如图 5 所示。

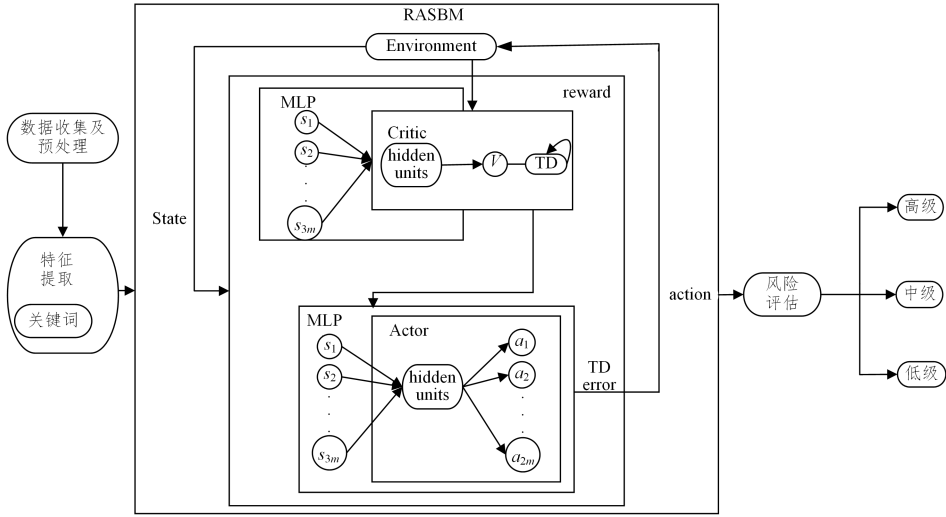


图 5 风险评价体系基线模型
Fig. 5 Baseline model of risk assessment system

由图 5 可知,数据收集和预处理包括从复合链式结构中获取关联图中实体的原始数据以及对文本信息的相关处理。实体交易信息包括交易对象、交易对象简介、交易合同等文本信息及交易金额、交易时间等数值信息。对数据进行初步的处理,为后续特征提取做准备。在预处理之后,对文本信息进行特征提取,将其转化成数值向量,从语义层面对文本信息提取特征,选择关键词方法进行文本特征的表示。本文采用信息增益法提取关键词。

在得到上述多文本特征后,引入 Max-min ACLA(Actor-Critic Learning Automaton)算法构建 RASBM。设数据集 $D=\{(x^1,y^1),(x^2,y^2),\cdots,(x^i,y^i),\cdots,(x^n,y^n)\}$, n 为样本总数,其中, x^i 为第 i 个样本的特征向量, y^i 是它的目标类别, $y^i \in \{0,1,\cdots,N-1\}$, N 是类别总数,这里设为 3。该算法为每个目标类别构建一个智能体,对于每个训练样本,与其类别相同的智能体会选择动作来使它获得的奖励最大化,而其他类别的智能体会选择动作来使它获得的奖励最小化。该算法的马尔可夫决策流程如下:

Step1 定义状态集 S 。 S 通常是连续的,对于长度为 m 的输入向量 x^i ,状态 $s^i \in S$,其包含 $3m$ 个元素。这些元素被分成 3 个桶, s_t 表示单次迭代中 t 时刻的状态向量,对于输入向量 x^i ,设初始状态 $s_0=(x^i,\vec{0},x^i)$,3 个桶的大小均为 m 。

Step2 定义动作集 A 。共有 $2m$ 个动作,每个动作设置其对应桶元素的值, a_t 表示 t 时刻选择的动作。

Step3 定义转移函数集合 T 。下一时刻状态 $s_{t+1}=O(s_t,a_t)$,操作 O 的执行规则为:如果动作满足 $0 \leq a_t < m$,则将第 $(m+a_t)$ 个桶元素设置为输入向量第 a_t 个元素的值;如果动作满足 $m \leq a_t < 2m$,则将第 $(m+a_t)$ 个桶元素设置为 0。

Step4 定义即时奖励 R 。 $r_t=1-\frac{z}{m}$,其中, z 表示状态向量中 0 的个数。

Step5 定义折扣因子 γ 。
在训练过程中,设单次迭代执行动作的次数为 h ,智能体

与训练样本进行交互,每个智能体执行 h 个动作,并且从观察到的状态转移和获得的即时奖励中学习。用两个不同的多层感知器 (Multilayer Perceptron, MLP) 来表示状态值函数 (Critic) 和挑选动作的功能函数 (Actor)。本文设两个 MLP 的隐含层节点个数和学习速率相同,以减少参数的数量。假设 $V_i(\cdot)$ 表示类别 i 的智能体 AC_i 的值函数,样本误差用 TD error 表示, AC_i 在状态 s_t 执行完动作后,将会收到 (s_t,a_t,r_t,s_{t+1}) ,使用 TD error δ_t 来更新值函数。

$$\delta_t = r_t + \gamma V_i(s_{t+1}) - V_i(s_t) \tag{1}$$

$$V_i(s_t) = V_i(s_t) + \alpha \delta_t \tag{2}$$

其中, α 是 critic 的学习速率,若样本类别 $y=i$,则 AC_i 选择动作的 MLP 目标值为:

$$\begin{cases} G=1, & \text{if } \delta_t \geq 0 \\ G=0, & \text{if } \delta_t < 0 \end{cases} \tag{3}$$

AC_i 可通过最大化即时奖励来学习更高的状态值函数。若样本类别 $y \neq i$,则 AC_i 选择动作的 MLP 目标值为:

$$\begin{cases} G=0, & \text{if } \delta_t \geq 0 \\ G=1, & \text{if } \delta_t < 0 \end{cases} \tag{4}$$

对于类别不同的样本, AC_i 会得到负的即时奖励,这些奖励将会通过 TD 学习传递给初始状态的值函数。在测试阶段,智能体不需要选择动作。首先对所有类别 i 的 AC_i 计算值函数 $V_i(s_0)$,输入样本将会被预测为具有最大值函数的智能体代表的类别 y_p 。

$$y_p = \arg \max_i V_i(s_0) \tag{5}$$

按照上述过程构建包含 3 个智能体的 RASBM,对实体进行风险评估。

5 实验与分析

实验数据集为中山大学区块链与智能金融研究中心 In-PlusLab 开发的 DApps(基于区块链的去中心化应用程序)活跃度数据集,数据集大小为 10 万条,数据集详细信息如表 3 所列。

表 3 实验数据集
Table 3 Experimental dataset

名称	属性
link	DApp 的链接
name	DApp 的名称
Txs_24h	该 DApp 最近 24 小时的交易量
Txs_7d	该 DApp 最近 7 天的交易量
balance	账户在 2019 年 9 月 13 日的余额
category	DApp 的类别
protocol	DApp 的协议
rank	DApp 在 2019 年 9 月 13 日的排名
users_24h	最近 24 小时内的活跃用户数量
volume_24h	最近 24 小时的交易额
volume_7d	最近 7 天的交易额
long_intro	完整介绍
other_link	其他链接
submitted	DApp 提交到市场的时间
contract	包含的合约

分别从存储效率、存储开销和关联溯源准确率等几个方面来验证复合区块链关联事件溯源方法的有效性,对比实验为相关工作中提到的单链结构和多叉链结构的存储和溯源。

进一步对实体进行风险评估分析实验,将源事件关联图中实体的数据集按照 7:3 的比例分为训练集和测试集,每类数据大体上保持了在总数据中所占的比例。在分类中,0 代表高级风险,1 代表中级风险,2 代表低级风险。对于多种文本特征,首先使用 RASBM 对相关实体进行分类,然后针对关键词特征,本文通过支持向量机 (Support Vector Machine, SVM)、逻辑回归 (Logistic Regression, LR)、多层感知机 (MLP)、朴素贝叶斯 (Naive Bayes Model, NBM) 等有监督机器学习方法进行实验对比。各模型参数设置如表 4 所列。

表 4 模型参数设置
Table 4 Model parameter setting

模型名称	参数
RASBM	隐含层节点个数: l_1 ; 学习速率 α : 0.03; 执行动作次数 h : 7; 折扣因子 γ : 0.9
SVM	惩罚因子 C : 1.0; 核函数: rbf
LR	惩罚项: l_2 ; 优化拟合参数算法: liblinear
MLP	优化器: lbfgs; 正则化项参数 α : 1×10^{-5} ; 隐含层节点个数: 5; 学习速率: 0.001
NBM	方式: Bernoulli; 平滑系数: 1.0

本文使用的效果衡量指标定义如下:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100\%$$
 (6)

$$precision = \frac{TP}{TP + FP} \times 100\%$$
 (7)

其中, TP 表示被正确分类的正例数量, FP 表示被错误分类的负例数量, TN 表示被正确分类的负例数量, FN 表示被错误分类的正例数量。 $accuracy$ 表示正确率,即模型的整体性能,不区分具体类别,通过 $accuracy$ 得到模型的整体性能和收敛情况; $precision$ 表示准确率,每一类别均可得到其准确率,通过 $precision$ 来评价模型对每个类别及整体的分类能力。

5.1 复合链式存储效率分析

实验模拟了采用复合链式结构与采用单链、多叉链的区块链存储结构的效率对比。横坐标表示存储实体数据集,纵坐标表示存储所需的时间,实验结果如图 6 所示。

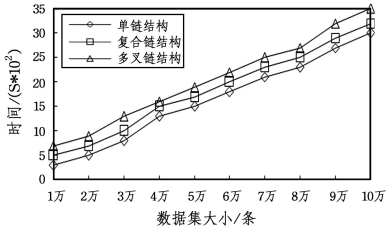


图 6 复合链式存储效率对比

Fig. 6 Composite chain storage efficiency comparison

由图 6 可得,本文提出的复合链式存储结构的存储效率介于单链结构和多叉链结构之间,同时采用区块链存储数据,相比传统数据库,安全性提高了 50%。

5.2 复合链式存储开销分析

实验模拟了复合链式结构与采用单链、多叉链的区块链存储结构的开销对比。横坐标表示存储实体数据集,纵坐标表示存储需要占用的空间大小,实验结果如图 7 所示。

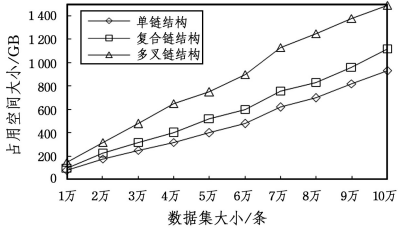


图 7 复合链式存储开销对比

Fig. 7 Composite chain storage overhead comparison

由图 7 可得,本文提出的复合链式存储结构的存储开销介于单链结构和多叉链结构之间。相比多叉链结构,其存储开销可降低 60%,原因在于多叉链冗余存储实体数据,因此其比单链和复合链存储结构具有更大的空间开销。

5.3 关联溯源准确率分析

图 8 为采用复合区块链关联事件溯源方法和单链、多叉链存储模式溯源的准确率对比图。其中横坐标表示存储的实体数据集,纵坐标表示溯源准确率,实验结果如图 8 所示。

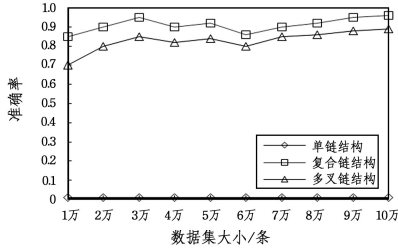


图 8 关联追溯查询准确率对比

Fig. 8 Associated traceability query precision comparison

由图 8 可得,相比单链模式溯源查询,复合链式和多叉链溯源查询准确率显著提高,可提高 90%,但多叉链存储模式结构复杂,其查询效率显著低于复合链式存储结构。

5.4 风险评估正确率分析

不同关键词个数实验的训练集 $accuracy$ 结果如图 11 所示,其中横轴表示迭代次数,纵轴表示 $accuracy$,不同标记的曲线表示不同的关键词个数的模型 $accuracy$,实验结果如图 9 所示。

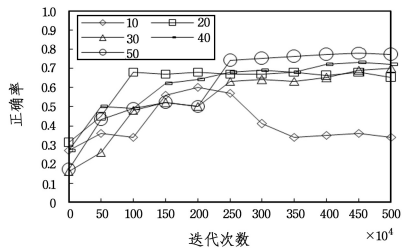


图9 风险评估正确率对比

Fig. 9 Risk assessment accuracy

从图9中可以看出,模型未收敛时,accuracy呈上升趋势,当模型收敛后,accuracy趋于稳定。

5.5 风险评估准确率分析

本文使用4种传统的机器学习算法做相同的实验进行对比,验证RASBM的有效性。选择precision作为主要的衡量标准找出有风险的机构,其中横轴表示关键词个数,纵轴表示precision,不同标记的曲线表示不同算法的关键词特征precision,实验结果如图10所示。

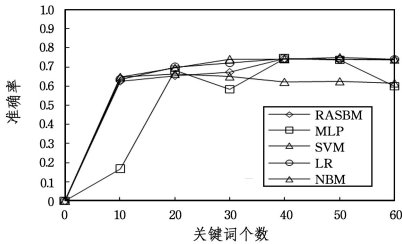


图10 风险评估准确率对比

Fig. 10 Risk assessment precision comparison

图11给出了各算法效果最优时,各个子类别的precision。

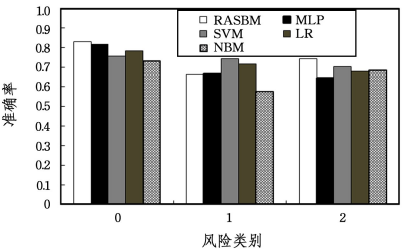


图11 每个风险类别的准确率对比

Fig. 11 Precision comparison of each risk category

从图10可以看出,在特征方面,大多数模型在关键词取40个时,模型性能达到最优;在模型方面,就总体的precision而言,RASBM优于大多数传统机器学习模型。同时,从图11可以看出,RASBM受到数据不平衡的影响较小,而大多数传统机器学习模型都会受到数据不平衡的影响,使得数据较少一类的precision较低。

结束语 区块链是数据安全存储和溯源查询的有效手段。区块链通过增加区块的方式来进行数据存储,数据通过单链模式进行存储,随着时间的推移和交易数据的扩大,会引发数据膨胀,可能出现存储和查询效率降低的问题。同时,单链存储模式也不能实现复杂或分类场景下的自适应数据关联存储。在溯源查询时,现有区块链以时间戳和哈希指针为

线索,通过查找区块体中的交易只能实现溯源到某一区块上的某一交易数据,无法实现溯源查询出区块链中各实体之间隐含的联系,对后续的事件分析支持不足。针对这些问题,本文对区块链存储结构和溯源方法进行了深入的研究,提出了一种复合区块链关联事件溯源方法。该方法首先构建区块链复合链式存储结构模型,实现复杂或分类场景下的自适应数据关联存储;然后,在追溯查询时,在获取到事件源实体区块的基础上,建立辅助存储空间对相关数据进行转存,提出基于Apriori算法的关联实体区块溯源方法,并将所得的溯源实体区块用于构造源事件关联图,进而描述事件实体间的关联关系;最后,提出基于强化学习的风险评价体系,实现对溯源实体的风险评估。

参考文献

[1] AHMAD R W,HASAN H,YAQOOB I,et al. Blockchain for Aerospace and Defense:Opportunities and Open Research Challenges[J]. Computers & Industrial Engineering, 2021, 151: 106982.

[2] TIAN G H,HU Y H,CHEN X F. Research progress of blockchain system attack and defense technology[J]. Journal of Software, 2021, 32(5): 1495-1525.

[3] WAN P K,HUANG L,HOLTSKOG H. Blockchain-enabled Information Sharing within a Supply Chain: A Systematic Literature Review[J]. IEEE Access, 2020, 8: 49645-49656.

[4] CAI T,LIN H,CHEN W H, et al. Blockchain-enabled Efficient Data Sharing Scheme for Internet of Things[J]. Journal of Software, 2021, 32(4): 953-972.

[5] WEI L F,ZHU J Y,HENG X R, et al. Design and implementation of intelligent traceability system for aquatic product quality and safety based on block chain technology and HACCP management[J]. Fishery Modernization, 2020, 47(4): 8.

[6] BARTOLETTI M,BRACCIALI A,LANDE S, et al. A general framework for blockchain analytics[C]// Proceedings of the 1st Workshop on Scalable and Resilient Infrastructures for Distributed Ledgers. Las Vegas, Nevada, USA, 2017: 11-15.

[7] HE P, YU G,ZHANG Y F, et al. Review of Blockchain Technology and Application[J]. Computer Science, 2017, 44(4): 1-7.

[8] IEMIEUX V L. Trusting records:is blockchain technology the answer? [J]. Records Management Journal, 2016, 26(2): 110-139.

[9] WANG S,DINH T A,LIN Q, et al. ForkBase: An Efficient Storage Engine for Blockchain and Forkable Applications[J]. arXiv:1802. 04949, 2018.

[10] HALPIN H,PIEKARSKA M. Introduction to Security and Privacy on the Blockchain[C]// IEEE European Symposium on Security & Privacy Workshops. IEEE, 2017: 1-3.

[11] KRAFT D. Difficulty control for blockchain-based consensus systems[J]. Peer-to-Peer Networking and Applications, 2016, 9(2): 397-413.

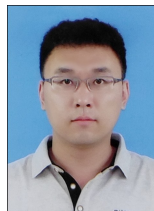
[12] DANNEN C. Introducing Ethereum and Solidity: Foundations of Cryptocurrency and Blockchain Programming for Beginners [M]. Apress, 2017.

[13] DINH T T A,WANG J,CHEN G, et al. BLOCKBENCH: a

- framework for analyzing private blockchains[C]// International Conference on Management of Data, 2017:1085-1100.
- [14] CAO Y, MIAO Z G. An Improved Apriori Algorithm for Frequent Terms Optimization Based on Vector Matrix[J]. Journal of Jilin University (Natural Science Edition), 2016, 54(2): 349-353.
- [15] JIN H, DAI X, XIAO J. Towards a Novel Architecture for Enabling Interoperability amongst Multiple Blockchains[C]// International Conference on Distributed Computing Systems. IEEE Computer Society, 2018: 1203-1211.
- [16] SHAO Q F, JIN C Q, ZHANG Z, et al. Blockchain Technology: Architecture and Progress[J]. Chinese Journal of Computers, 2018, 41(5): 969-988.
- [17] WANG Q G, HE P, NIE T Z, et al. Overview of Blockchain System Data Storage and Query Technology[J]. Computer Science, 2018, 45(12): 12-18.
- [18] LIU L J. Research and application of improved Apriori algorithm [J]. Computer Engineering and Design, 2017(12): 142-146.
- [19] BIAN G Q, WANG Y. An Improved Apriori Algorithm Based on Matrix and Weight[J]. Microelectronics & Computer, 2017, 34(1): 137-140.
- [20] LI Y, ZHENG K, YAN Y, et al. EtherQL: a query layer for blockchain system[C]// International Conference on Database Systems for Advanced Applications, 2017.
- [21] ENRICO C, SERENA N, ANTONINO N, et al. A two-tier Blockchain framework to increase protection and autonomy of smart objects in the IoT[J]. Computer Communications, 2022, 181: 338-356.
- [22] LUU L, NARAYANAN V, ZHENG C, et al. A secure sharding protocol for open blockchains[C]// 2016 ACM SIGSAC Conference. ACM, 2016.
- [23] WANG Z H, LIU P Z, SONG C B, et al. Research on Flexible Trustable Traceability System of Agricultural Products Based on Blockchain [J]. Computer Engineering, 2020, 46(12): 319-326.
- [24] HAN H. Blockchain infiltration into data transaction to solve traceability and authorization "pain points"[J]. Communications World, 2017(19): 1.
- [25] HUANG Z Z, ZHANG X D, ZHAO J H, et al. Design of knowledge sharing mechanism based on blockchain [J]. Journal of Chongqing University of Technology (Natural Science), 2021, 35(9): 143-151.
- [26] QIAO R, CAO Y, WANG Q X. Dynamic data traceability mechanism of internet of things based on alliance chain[J]. Journal of Software, 2019, 30(6): 1614-1631.
- [27] SONG S, PENG W. BLOCCE+: An Improved Blockchain-Based Covert Communication Approach [J]. Journal of Chongqing University of Technology (Natural Science), 2020, 34(9): 238-244.
- [28] XI L, WANG R D, FAN H Y, et al. Unsupervised Anomaly Detection Model based on Sample Association Perception[J]. Chinese Journal of Computers, 2021, 44(11): 2317-2331.



LI Su, born in 1997, postgraduate, is a student member of China Computer Federation. Her main research interests include large scale map processing techniques big data processing techniques, and stream data processing techniques.



WANG Jun-lu, born in 1988, Ph.D candidate, lecturer, is a member of China Computer Federation. His main research interests include large scale map processing techniques, big data processing techniques and stream data processing techniques.

(责任编辑:李亚辉)