(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 114511330 B
(45)授权公告日 2022.12.13

(21)申请号 202210401495.3
(22)申请日 2022.04.18
(65)同一申请的已公布的文献号
申请公布号 CN 114511330 A
(43)申请公布日 2022.05.17
(73)专利权人 山东省计算中心(国家超级计算济南中心)
地址 250014 山东省济南市历下区科院路19号
(72)发明人 张淑慧 兰田 王连海 徐淑奖 邵蔚
(74)专利代理机构 济南圣达知识产权代理有限公司 37221
专利代理师 王雪
(51)Int.Cl.
G06Q 20/40(2012.01)
G06K 9/62(2022.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

(56)对比文件
CN 113191775 A,2021.07.30
CN 113362071 A,2021.09.07
CN 113806746 A,2021.12.17
CN 107066553 A,2017.08.18
CN 113127933 A,2021.07.16
US 2022067738 A1,2022.03.03
汪谦生.浅谈互联网金融中“庞氏骗局”的识别与防范.《计算机产品与流通》.2017,(第09期),
Eko Mulyani 等.“Dropout Prediction Optimization through SMOTE and Ensemble Learning”.《2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)》.2020,

审查员 李平

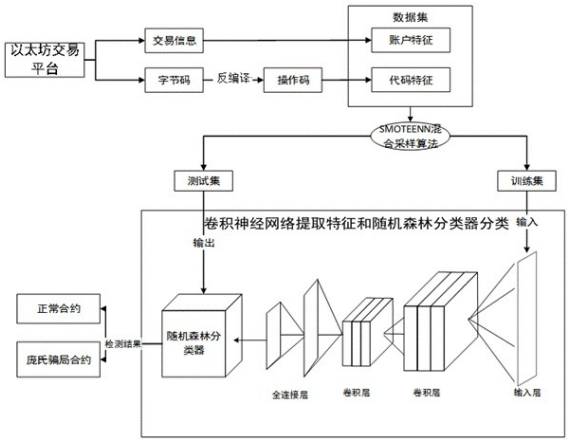
权利要求书2页 说明书5页 附图3页

(54)发明名称

一种基于改进的CNN-RF的以太坊庞氏骗局检测方法及系统

(57)摘要

本发明属于区块链异常行为检测领域,提供了一种基于改进的CNN-RF的以太坊庞氏骗局检测方法及系统。该方法包括,获取以太坊上的智能合约数据;提取智能合约数据的账户特征和操作码特征,将账户特征和操作码特征结合,得到混合特征;基于混合特征,采用CNN特征提取模型,提取得到庞氏骗局合约检测的关键特征;基于关键特征,采用RF分类模型,得到是否是庞氏骗局合约的检测结果。本发明使用卷积神经网络模型来进行关键特征数据的筛选,并融合了随机森林分类器的预测模型的训练和使用,提高了检测精确率。



CN 114511330 B

1.一种基于改进的CNN-RF的以太坊庞氏骗局检测方法,其特征在于,包括:

获取以太坊上的智能合约数据;

提取智能合约数据的账户特征和操作码特征,将账户特征和操作码特征结合,得到混合特征;

基于混合特征,采用CNN特征提取模型,提取得到庞氏骗局合约检测的关键特征;

基于关键特征,采用RF分类模型,得到是否是庞氏骗局合约的检测结果;

使用卷积神经网络模型来进行关键特征数据的筛选,并融合了随机森林分类器的预测模型的训练和使用,提高了检测精确率;

所述获取以太坊上的智能合约数据具体包括:依据正常合约的合约地址,获取正常合约的账户交易信息和合约运行字节码;依据庞氏骗局合约的合约地址,获取庞氏骗局合约的账户交易信息和合约运行字节码;

所述提取智能合约数据的账户特征具体包括:分别提取正常合约账户交易信息的正常账户特征和庞氏骗局合约账户交易信息的异常账户特征;

提取操作码特征具体包括:采用反编译工具分别对正常合约运行字节码和庞氏骗局合约运行字节码进行反编译,将反编译后的操作码进行调用频率统计,得到正常操作码特征和异常操作码特征;

所述提取智能合约数据的账户特征的具体过程包括:采用合约交易时间和交易金额计算处理得到投资者的数量、投资金额、投资者的收益和投资金额的比例、新投资者所占的比例和投资者支付的最大金额;

所述改进的CNN-RF中的CNN特征提取模型中设置参数自动优化器;在训练过程中,所述参数自动优化器用于不断进行模型的性能调优并把最优的训练模型保存;

所述训练过程包括:

基于智能合约数据的账户特征和操作码特征,构建特征数据集;

采用SMOTE+ENN算法对特征数据集进行混合采样,得到新的样本集;

采用新的样本集中的样本训练CNN-RF,得到训练好的CNN-RF;具体的,输入智能合约的账户特征和操作码特征至两个依次连接的卷积层,以自动提取关键特征,每个卷积层后连接一个线性整流函数;将最后一个卷积层的输出输入两个依次连接的全连接层,第一个全连接层的输入和输出维度分别为16和8,最后一个全连接层的输入和输出维度分别为8和2;最后一个全连接层的输出,通过交叉损失函数,计算得到损失值后,采用自适应动量估计算法,更新两个卷积层的权重;将提取的最优特征输入随机森林分类器,以实现随机森林分类器的自训练;

所述采用SMOTE+ENN算法对特征数据集进行混合采样的具体过程包括:采用SMOTE算法对特征数据集进行过采样处理,再采用ENN数据清洗算法对数据进行去重操作,最后生成新的样本集;

引入SMOTE+ENN混合采样算法对CNN-RF进行了改进,避免在经过SMOTE过采样处理后的数据重叠问题,实现关键特征的自动提取,并提高了检测精确率。

2.根据权利要求1所述的基于改进的CNN-RF的以太坊庞氏骗局检测方法,其特征在于,所述智能合约数据基于爬虫技术从以太坊区块链浏览器网站进行获取。

3.一种基于改进的CNN-RF的以太坊庞氏骗局检测系统,其特征在于,包括:

数据获取模块,其被配置为:获取以太坊上的智能合约数据;

特征提取模块,其被配置为:提取智能合约数据的账户特征和操作码特征,将账户特征和操作码特征结合,得到混合特征;

模型处理模块,其被配置为:基于混合特征,采用CNN特征提取模型,提取得到庞氏骗局合约检测的关键特征;

检测模块,其被配置为:基于关键特征,采用RF分类模型,得到是否是庞氏骗局合约的检测结果;

使用卷积神经网络模型来进行关键特征数据的筛选,并融合了随机森林分类器的预测模型的训练和使用,提高了检测精确率;

所述获取以太坊上的智能合约数据具体包括:依据正常合约的合约地址,获取正常合约的账户交易信息和合约运行字节码;依据庞氏骗局合约的合约地址,获取庞氏骗局合约的账户交易信息和合约运行字节码;

所述提取智能合约数据的账户特征具体包括:分别提取正常合约账户交易信息的正常账户特征和庞氏骗局合约账户交易信息的异常账户特征;

提取操作码特征具体包括:采用反编译工具分别对正常合约运行字节码和庞氏骗局合约运行字节码进行反编译,将反编译后的操作码进行调用频率统计,得到正常操作码特征和异常操作码特征;

所述改进的CNN-RF中的CNN特征提取模型中设置参数自动优化器;在训练过程中,所述参数自动优化器用于不断进行模型的性能调优并把最优的训练模型保存;

所述训练过程包括:

基于智能合约数据的账户特征和操作码特征,构建特征数据集;

采用SMOTE+ENN算法对特征数据集进行混合采样,得到新的样本集;

采用新的样本集中的样本训练CNN-RF,得到训练好的CNN-RF;具体的,输入智能合约的账户特征和操作码特征至两个依次连接的卷积层,以自动提取关键特征,每个卷积层后连接一个线性整流函数;将最后一个卷积层的输出输入两个依次连接的全连接层,第一个全连接层的输入和输出维度分别为16和8,最后一个全连接层的输入和输出维度分别为8和2;最后一个全连接层的输出,通过交叉损失函数,计算得到损失值后,采用自适应动量估计算法,更新两个卷积层的权重;将提取的最优特征输入随机森林分类器,以实现随机森林分类器的自训练;

所述采用SMOTE+ENN算法对特征数据集进行混合采样的具体过程包括:采用SMOTE算法对特征数据集进行过采样处理,再采用ENN数据清洗算法对数据进行去重操作,最后生成新的样本集;

所述提取智能合约数据的账户特征的具体过程包括:采用合约交易时间和交易金额计算处理得到投资者的数量、投资金额、投资者的收益和投资金额的比例、新投资者所占的比例和投资者支付的最大金额;

引入SMOTE+ENN混合采样算法对CNN-RF进行了改进,避免在经过SMOTE过采样处理后的数据重叠问题,实现关键特征的自动提取,并提高了检测精确率。

一种基于改进的CNN-RF的以太坊庞氏骗局检测方法及系统

技术领域

[0001] 本发明属于区块链异常行为检测领域，尤其涉及一种基于改进的CNN-RF的以太坊庞氏骗局检测方法及系统。

背景技术

[0002] 本部分的陈述仅仅是提供了与本发明相关的背景技术信息，不必然构成在先技术。

[0003] 数字货币的发行需要较高的技术需求，因此，拥有去中心化分布式数据库、智能合约和共识算法等技术优势的区块链技术，便成为了银行发行数字货币技术基础的重点备选技术。虽然各国对区块链技术的研究进展不断，但是区块链技术在实施上不仅存在着法律和监管方面的问题，而数字货币的集中管理需求与区块链技术去中心化的特性还存在着一些矛盾。尤其是附着智能合约功能的以太坊平台，因为智能合约在以太坊上部署成功后，只要满足运行条件就会自动执行且外界无法干预终止程序。这也就使得非法投机分子趁机而入，利用合约的恶意代码嵌入进行钱财的聚敛，最典型的代表是庞氏骗局的合约诈骗。所以以太坊上的庞氏骗局检测方法研究迫在眉睫。

[0004] 目前，已有众多的学者将研究的目光投向了区块链的异常行为检测。其中，研究的热点便是对区块链的以太坊庞氏骗局的合约检测。前期研究者都是通过人工对合约代码和账户交易信息进行分析判断该合约是否是庞氏骗局合约。直到部分研究学者将机器学习和数据挖掘算法引入才使得区块链的异常检测方法得到简化，这也使得以太坊庞氏骗局检测成为区块链异常检测的焦点。但是无论是数据挖掘技术还是机器学习算法，在以太坊庞氏骗局的检测上对于特征数据的不平衡处理和检测方法的检测性能上依旧存在不足之处。

发明内容

[0005] 为了解决上述背景技术中存在的技术问题，本发明提供一种基于改进的CNN-RF的以太坊庞氏骗局检测方法及系统，其采用SMOTE+ENN混合采样算法对样本不平衡的特征数据集进行处理，使用卷积神经网络模型来进行关键特征数据的筛选，并融合了随机森林分类器的预测模型的训练和使用，提高了检测精确率。

[0006] 为了实现上述目的，本发明采用如下技术方案：

[0007] 本发明的第一个方面提供一种基于改进的CNN-RF的以太坊庞氏骗局检测方法。

[0008] 一种基于改进的CNN-RF的以太坊庞氏骗局检测方法，包括：

[0009] 获取以太坊上的智能合约数据；

[0010] 提取智能合约数据的账户特征和操作码特征，将账户特征和操作码特征结合，得到混合特征；

[0011] 基于混合特征，采用CNN特征提取模型，提取得到庞氏骗局合约检测的关键特征；

[0012] 基于关键特征，采用RF分类模型，得到是否是庞氏骗局合约的检测结果。

[0013] 进一步地，所述获取以太坊上的智能合约数据具体包括：依据正常合约的合约地

址,获取正常合约的账户交易信息和合约运行字节码;依据庞氏骗局合约的合约地址,获取庞氏骗局合约的账户交易信息和合约运行字节码。

[0014] 进一步地,所述提取智能合约数据的账户特征具体包括:分别提取正常合约账户交易信息的正常账户特征和庞氏骗局合约账户交易信息的异常账户特征。

[0015] 进一步地,所述提取操作码特征具体包括:采用反编译工具分别对正常合约运行字节码和庞氏骗局合约运行字节码进行反编译,得到正常操作码特征和异常操作码特征。

[0016] 进一步地,所述改进的CNN-RF中的CNN特征提取模型中设置参数自动优化器;在训练过程中,所述参数自动优化器用于不断进行模型的性能调优并把最优的训练模型保存。

[0017] 进一步地,所述训练过程包括:

[0018] 基于智能合约数据的账户特征和操作码特征,构建特征数据集;

[0019] 采用SMOTE+ENN算法对特征数据集进行混合采样,得到新的样本集;

[0020] 采用新的样本集中的样本训练CNN-RF,得到训练好的CNN-RF。

[0021] 进一步地,所述采用SMOTE+ENN算法对特征数据集进行混合采样的具体过程包括:采用SMOTE算法对特征数据集进行过采样处理,再采用ENN数据清洗算法对数据进行去重操作,最后生成新的样本集。

[0022] 进一步地,所述提取智能合约数据的账户特征的具体过程包括:采用合约交易时间和交易金额计算处理得到投资者的数量、投资金额、投资者的收益和投资金额的比例、新投资者所占的比例和投资者支付的最大金额。

[0023] 进一步地,所述智能合约数据基于爬虫技术从以太坊区块链浏览器网站进行获取。

[0024] 本发明的第二个方面提供一种基于改进的CNN-RF的以太坊庞氏骗局检测系统。

[0025] 一种基于改进的CNN-RF的以太坊庞氏骗局检测系统,包括:

[0026] 数据获取模块,其被配置为:获取以太坊上的智能合约数据;

[0027] 特征提取模块,其被配置为:提取智能合约数据的账户特征和操作码特征,将账户特征和操作码特征结合,得到混合特征;

[0028] 模型处理模块,其被配置为:基于混合特征,采用CNN特征提取模型,提取得到庞氏骗局合约检测的关键特征;

[0029] 检测模块,其被配置为:基于关键特征,采用RF分类模型,得到是否是庞氏骗局合约的检测结果。

[0030] 与现有技术相比,本发明的有益效果是:

[0031] 1、本发明引入SMOTE+ENN混合采样算法对CNN-RF进行了改进,可以避免在经过SMOTE过采样处理后的数据重叠问题,实现关键特征的自动提取,并提高了检测精确率。

[0032] 2、本实施例的一种基于改进的CNN-RF的以太坊庞氏骗局检测方法不但解决了数据过采样后的数据重复问题,还简化了关键特征提取过程。并通过实验证明,本实施例的检测方法在任何层面上都更适合以太坊庞氏骗局的检测。

附图说明

[0033] 构成本发明的一部分的说明书附图用来提供对本发明的进一步理解,本发明的示意性实施例及其说明用于解释本发明,并不构成对本发明的不当限定。

[0034] 图1为本发明基于改进的CNN-RF的以太坊庞氏骗局检测方法流程图；

[0035] 图2为本发明CNN-RF训练模型的架构图；

[0036] 图3为本发明有SMOTE+ENN算法处理的CNN-RF检测混淆矩阵图；

[0037] 图4为本发明无SMOTE+ENN算法处理的CNN-RF检测混淆矩阵图。

具体实施方式

[0038] 下面结合附图与实施例对本发明作进一步说明。

[0039] 应该指出，以下详细说明都是例示性的，旨在对本发明提供进一步的说明。除非另有指明，本文使用的所有技术和科学术语具有与本发明所属技术领域的普通技术人员通常理解的相同含义。

[0040] 需要注意的是，这里所使用的术语仅是为了描述具体实施方式，而非意图限制根据本发明的示例性实施方式。如在这里所使用的，除非上下文另外明确指出，否则单数形式也意图包括复数形式，此外，还应当理解的是，当在本说明书中使用术语“包含”和/或“包括”时，其指明存在特征、步骤、操作、器件、组件和/或它们的组合。

[0041] 需要注意的是，附图中的流程图和框图示出了根据本公开的各种实施例的方法和系统的可能实现的体系架构、功能和操作。应当注意，流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分，所述模块、程序段、或代码的一部分可以包括一个或多个用于实现各个实施例中所规定的逻辑功能的可执行指令。也应当注意，在有些作为备选的实现中，方框中所标注的功能也可以按照不同于附图中所标注的顺序发生。例如，两个接连地表示的方框实际上可以基本并行地执行，或者它们有时也可以按照相反的顺序执行，这取决于所涉及的功能。同样应当注意的是，流程图和/或框图中的每个方框、以及流程图和/或框图中的方框的组合，可以使用执行规定的功能或操作的专用的基于硬件的系统来实现，或者可以使用专用硬件与计算机指令的组合来实现。

[0042] 术语解释：

[0043] CNN-RF，Convolutional Neural Network- Random Forest，是指一种卷积神经网络和随机森林模型融合的一种检测方法，其中CNN是卷积神经网络是一类包含卷积计算且具有深度结构的前馈神经网络，是深度学习的代表算法之一，这里我们用来进行关键特征提取。RF是一种是利用多棵树对样本进行训练并预测的一种分类器。这里我们将CNN提取的特征输入到RF中进行训练预测分类。

[0044] SMOTE+ENN，Synthetic Minority Oversampling Technique+Edited nearest neighborhood，是指一种集成的混合采样算法。其中SMOTE是对少数类样本进行分析并根据少数类样本人工合成新样本添加到数据集中形成新的样本集。ENN是一种应用最近近邻算法(KNN)来编辑数据集，对于每一个要进行下采样的样本，那些绝大多数近邻样本不属于该类的样本会被移除，而绝大多数的近邻样本属于同一类的样本会被保留，对新生成的样本集进行数据清洗。保证我们经过处理的数据不重叠。

[0045] 实施例一

[0046] 如图1所示，本实施例提供了一种基于改进的CNN-RF的以太坊庞氏骗局检测方法，本实施例以该方法应用于服务器进行举例说明，可以理解的是，该方法也可以应用于终端，还可以应用于包括终端和服务器和系统，并通过终端和服务器的交互实现。服务器可以是

独立的物理服务器，也可以是多个物理服务器构成的服务器集群或者分布式系统，还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务器、云通信、中间件服务、域名服务、安全服务CDN、以及大数据和人工智能平台等基础云计算服务的云服务器。终端可以是智能手机、平板电脑、笔记本电脑、台式计算机、智能音箱、智能手表等，但并不局限于此。终端以及服务器可以通过有线或无线通信方式进行直接或间接地连接，本申请在此不做限制。本实施例中，该方法包括以下步骤：

[0047] 获取以太坊上的智能合约数据；

[0048] 提取智能合约数据的账户特征和操作码特征，将账户特征和操作码特征结合，得到混合特征；

[0049] 基于混合特征，采用CNN特征提取模型，提取得到庞氏骗局合约检测的关键特征；

[0050] 基于关键特征，采用RF分类模型，得到是否是庞氏骗局合约的检测结果。

[0051] 具体地，本实施例的具体方案可以采用以下方案实现，如图2所示：

[0052] Step 1:根据公开标记好的合约账户地址，从以太坊交易平台(Etherscan.io)上爬取相关的智能合约的账户内部交易信息和智能合约的字节码信息。其中账户交易包含区块数、时间戳、哈希值、交易转账和交易收账、合约地址、交易金额、交易余额和交易花费的Gas值等信息，使用python编程最常用的反编译工具Easy Python Decompiler进行合约字节码的反编译。

[0053] Step 2:将获取到的粗略数据进行处理，对众多的账户交易特征进行特征分析和选择，然后使用反编译工具对合约的字节码进行反编译处理，将反编译后的操作码进行调用频率统计。

[0054] 其中，包括特征选取:由合约交易时间和交易金额计算处理得到投资者的数量、投资金额、投资者的收益和投资金额的比例、新投资者所占的比例、投资者支付的最大金额。对反编译后的操作码进行调用频率的统计。

[0055] Step 3:账户和操作码特征数据被处理后将存入features.csv文件中，记为数据集Q。Q中的样本数据比例为12:1，存在数据不平衡问题。针对该问题本发明对Q进行SMOTE+ENN算法处理，并生成新数据集T。

[0056] 具体地，首先，使用SMOTE算法对数据集Q进行过采样处理，处理后的数据会存在多数据的重复。然后，使用ENN数据清洗算法对数据进行去重操作，最后生成新的样本集T。

[0057] 将数据集T按照8:2的比例将数据集T划分为训练集T_{train}和测试集T_{test}。

[0058] Step 4:划分后的训练集T_{train}作为数据输入源，输入到CNN中进行关键特征提取。

[0059] Step 5:CNN特征提取模型可以实现自动提取以太坊庞氏骗局检测的关键特征，无需人工进行数据集中有效特征的占比分析，节省了时间，减轻了人工操作负担。在CNN特征提取模型中我们设置了参数自动优化器，不断进行模型的性能调优并把最优的训练模型保存。将T_{train}和最佳的CNN模型加载到RF模型中，使用CNN对T_{train}进行特征提取，将提取的特征对RF模型进行训练。RF模型训练完，使用T_{test}对模型进行检测评估。为了方便进行模型的比较，使用精确率、召回率、F1值三个常用的性能指标对模型进行评估。

[0060] Step 6:模型训练完成后，使用我们划分的测试集进行模型性能的检测，经过实验结果验证本发明方法的优越性。

[0061] 对比分析:针对实验验证本实施例的检测方法的高效性,我们也选用了一些经典的机器学习模型进行了复现。经过实验验证,本实施例改进的CNN-RF检测方法相对于其他的机器学习模型更实用于以太坊庞氏骗局的检测。实验对比结果如表1所示。

[0062] 表1:各种方法的检测性能比较

[0063]

方法模型	精确率	召回率	F1 值
XGBoost	83%	69%	75%
RF	89%	62%	73%
LightGBM	83%	70%	73%
LinearSVC	64%	58%	60%
DT	64%	67%	65%
PD-SECR	97%	98%	98%

[0064] 针对有无SMOTE+ENN算法处理的检测方法,我们也做了对比实验,并绘制了混淆矩阵图。如图3和图4所示,经过SMOTE+ENN算法处理的检测方法对样本数据的检测准确率更高。

[0065] 本实施例的一种基于改进的CNN-RF的以太坊庞氏骗局检测方法不但解决了数据过采样后的数据重复问题,还简化了关键特征提取过程。并通过实验证明,本实施例的检测方法在任何层面上都更适合以太坊庞氏骗局的检测。

[0066] 实施例二

[0067] 本实施例提供了一种基于改进的CNN-RF的以太坊庞氏骗局检测系统。

[0068] 一种基于改进的CNN-RF的以太坊庞氏骗局检测系统,包括:

[0069] 数据获取模块,其被配置为:获取以太坊上的智能合约数据;

[0070] 特征提取模块,其被配置为:提取智能合约数据的账户特征和操作码特征,将账户特征和操作码特征结合,得到混合特征;

[0071] 模型处理模块,其被配置为:基于混合特征,采用CNN特征提取模型,提取得到庞氏骗局合约检测的关键特征;

[0072] 检测模块,其被配置为:基于关键特征,采用RF分类模型,得到是否是庞氏骗局合约的检测结果。

[0073] 此处需要说明的是,上述数据获取模块、特征提取模块、模型处理模块和检测模块与实施例一中的步骤所实现的示例和应用场景相同,但不限于上述实施例一所公开的内容。需要说明的是,上述模块作为系统的一部分可以在诸如一组计算机可执行指令的计算机系统中执行。

[0074] 以上所述仅为本发明的优选实施例而已,并不用于限制本发明,对于本领域的技术人员来说,本发明可以有各种更改和变化。凡在本发明的精神和原则之内,所作的任何修改、等同替换、改进等,均应包含在本发明的保护范围之内。

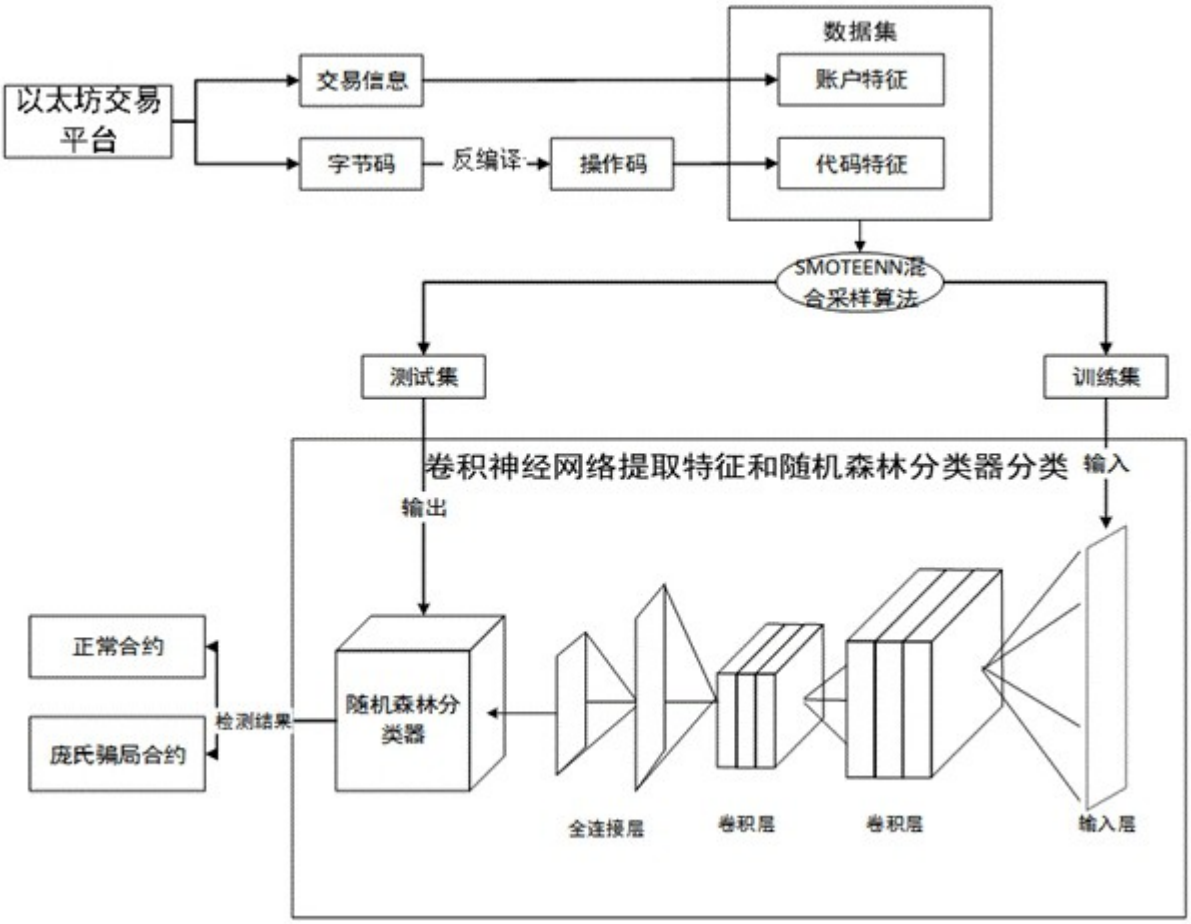

以太坊交易平台
交易信息
字节码
反编译
操作码
数据集
账户特征
代码特征
SMOTEENN混合采样算法
测试集
训练集
卷积神经网络提取特征和随机森林分类器分类
输入
输出
正常合约
庞氏骗局合约
检测结果
随机森林分类器
全连接层
卷积层
卷积层
输入层

图1

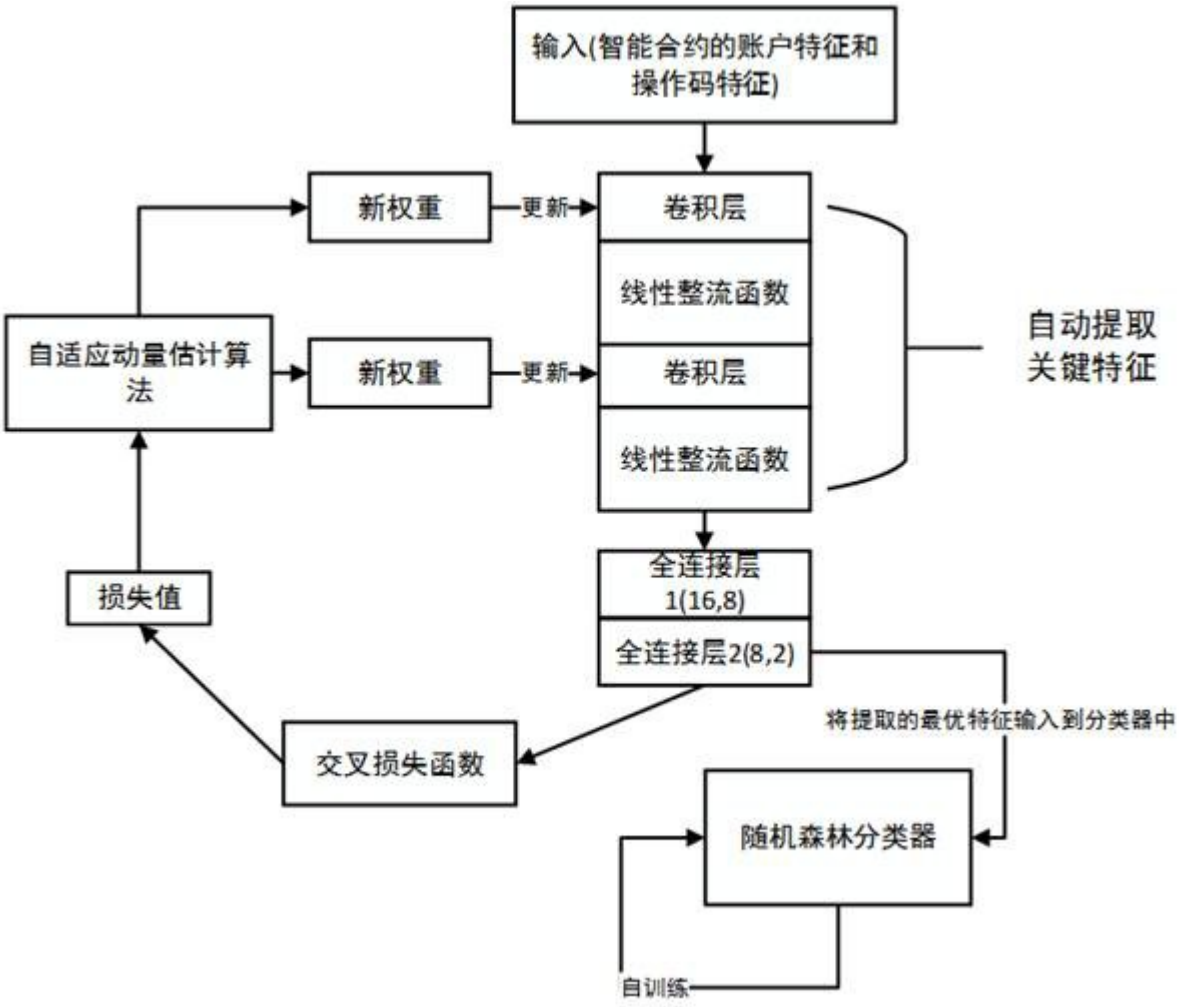

输入(智能合约的账户特征和操作码特征)
新权重
更新
卷积层
线性整流函数
卷积层
线性整流函数
自动提取关键特征
自适应动量估计算法
新权重
更新
全连接层1(16,8)
全连接层2(8,2)
损失值
将提取的最优特征输入到分类器中
交叉损失函数
随机森林分类器
自训练

图2

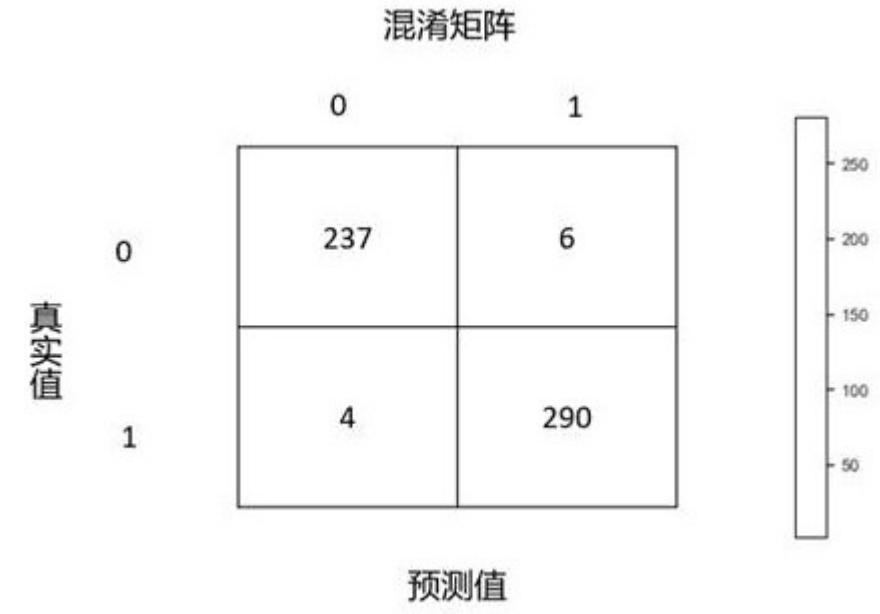

混淆矩阵
0
1
237
6
4
290
0
1
真实值
250
200
150
100
50
预测值

图3

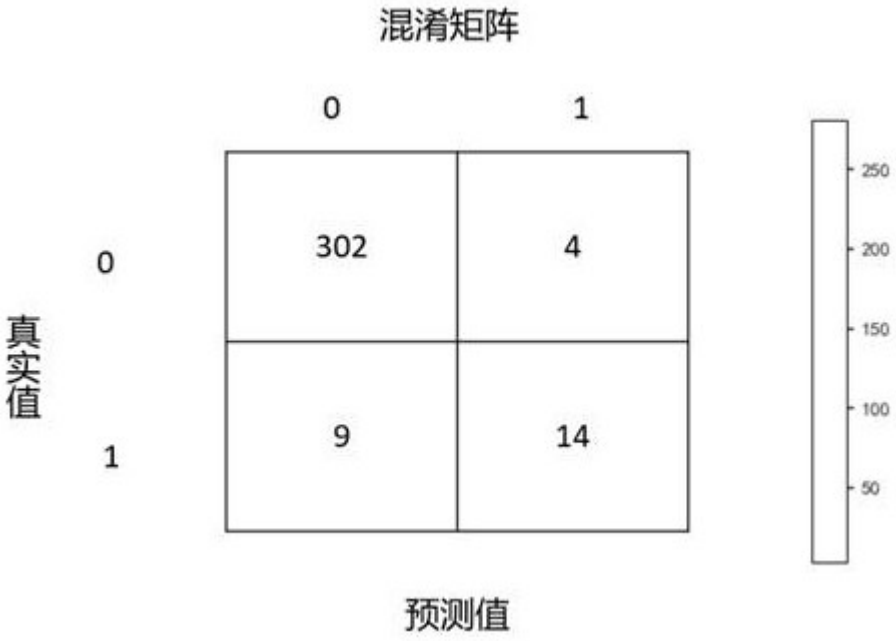

混淆矩阵
0
1
302
4
9
14
0
1
真实值
250
200
150
100
50
预测值

图4