基于双层孪生神经网络的区块链智能合约分类方法

郭加树* 王 琪 李择亚 武梦德 张红霞
(中国石油大学(华东)青岛软件学院 青岛 266580)
(中国石油大学(华东)计算机科学与技术学院 青岛 266580)

摘 要：当前通过深度学习方法进行区块链智能合约分类的方法越来越流行，但基于深度学习的方法往往需要大量的样本标签数据去进行有监督的模型训练，才能达到较高的分类性能。该文针对当前可用智能合约数据集存在数据类别不均衡以及标注数据量过少会导致模型训练困难，分类性能不佳的问题，提出基于双层孪生神经网络的小样本场景下的区块链智能合约分类方法：首先，通过分析智能合约数据特征，构建了可以捕获较长合约数据特征的双层孪生神经网络模型；然后，基于该模型设计了小样本场景下的智能合约训练策略和分类方法。最后，实验结果表明，该文所提方法在小样本场景下的分类性能优于目前最先进的智能合约分类方法，分类准确率达到94.7%，F1值达到94.6%，同时该方法对标签数据的需求更低，仅需同类型其他方法约20%数据量。

关键词：智能合约；区块链；孪生网络；以太坊
中图分类号：TN918; TP391 **文献标识码**：A **文章编号**：1009-5896(2024)03-1060-09
DOI: 10.11999/JEIT230185

Blockchain Smart Contract Classification Method Based on Double Siamese Neural Network

GUO Jiashu WANG Qi LI Zeya WU Mengde ZHANG Hongxia
(*Qingdao Institute of Software*, *China University of Petroleum*, *Qingdao* 266580, *China*)
(*College of Computer Science and Technology*, *China University of Petroleum*, *Qingdao* 266580, *China*)

Abstract: At present, methods for classifying blockchain smart contracts using deep learning methods are becoming increasingly popular. However, methods based on deep learning often require a large amount of sample label data for supervised model training to achieve high classification performance. A blockchain smart contract classification method based on a two-level twin neural network in a small sample scenario is proposed to address the problem that currently available smart contract datasets have uneven data categories and insufficient labeled data volumes, which can lead to difficulty in model training and poor classification performance. Firstly, by analyzing the characteristics of smart contract data, a two-level twin neural network model that can capture the characteristics of longer contract data is constructed; Then, based on this model, a training strategy and classification method for smart contracts in small sample scenarios are designed. Finally, experimental results show that the classification performance of the proposed method in this paper is superior to the most advanced smart contract classification methods in small sample scenarios, with a classification accuracy of 94.7% and an F1 value of 94.6%. At the same time, this method requires less tag data, requiring only about 20% data from other methods of the same type.

Key words: Smart contract; Blockchain; Siamese network; Ethereum

1 引言

智能合约[1]是以以太坊为代表的区块链2.0时代的核心技术，是一种部署在以太坊上，保障互不信任的参与节点之间进行通信交互的计算机协议，通常具有不可篡改和自动执行的特性。随着区块链技术的应用领域不断加深，智能合约的数据量也日益庞大，研究统计[2]，以太坊上平均每月发布的智能合约可以达到六位数。数量巨大的智能合约带来了机遇和挑战，一方面智能合约的应用类别已超过数百种，使得许多与智能合约和区块链相结合的业务变得流行起来，比如：区块链与云计算[3]、云存

收稿日期：2023-03-22；改回日期：2023-09-20；网络出版：2023-10-07
*通信作者： 郭加树 396105856@qq.com
基金项目：中石油重大科技项目(ZD2019-183-004)，中央高校基本科研业务费专项资金(20CX05019A)
Foundation Items: The Major Scientific and Technological Projects of CNPC (ZD2019-183-004), The Fundamental Research Funds for the Central Universities (20CX05019A)

储[4]，智能合约与物联网[5]、智慧医疗[6]等相结合；而另一方面，基于智能合约的骗局陷阱正变得日益猖獗[7]，包括庞氏骗局[8]、蜜罐陷阱[9]等恶意欺诈型骗局合约。对于用户来说，如何正确地检索和管理庞大的智能合约数据变得越来越困难。智能合约的分类问题已成为一项新兴的研究热点。

近年来，学者大多基于机器学习或者深度学习设计智能合约分类方法。例如，在智能合约的公共数据网站上爬取可用于模型训练的带有标签的数据集，通过引入智能合约语义信息以及相关交易信息作为数据特征，采用深度学习长短期记忆(Long-Short Term Memory, LSTM)模型或机器学习支持向量机(Support Vector Machine, SVM)算法等进行模型训练，合约分类。但目前的研究仍存在以下问题：(1)对合约类别不均衡以及数据重复问题考虑不充分，导致模型出现过拟合问题。(2)高精度的深度学习分类模型的前提是足够的标签训练数据量，但目前可用的标签数据量过少，无法满足较为复杂的深度神经网络的训练要求，导致训练的模型鲁棒性较差。(3)当前所用智能合约数据长度过长，未进行有效处理，在模型训练和分类时无法捕获合约自身的全局特征，导致分类效果不佳。

本文通过分析智能合约的数据结构特征，针对当前方法存在的问题，引入基于孪生神经网络[10]的度量学习方法，提出了小样本场景下基于双层孪生神经网络的区块链智能合约分类方法。首先，对数据类别不均衡的智能合约数据集进行了欠采样处理，每个类别仅保留相同数目且少量的合约数据，旨在以此克服数据类别不均衡问题。其次，针对智能合约数据长度过长的特点，设计了可以分别获取前后段合约特征的底层孪生网络。最后，在底层孪生网络的基础上构建了可以捕获全局数据特征并进行合约对相似度判断的双层孪生神经网络，提出了小样本场景下基于双层孪生神经网络的智能合约分类方法。对于智能合约分类问题，本文提出的分类方法借鉴了小样本学习[11]和度量学习[12]的思想，训练环境对标注数据量的需求更低，仅需要少量的数据量便可以完成模型训练，并达到较高的分类性能。

本文主要具有以下贡献：

(1)构建用于智能合约分类的双层孪生神经网络模型，可以有效处理较长的智能合约数据，并准确判别输入合约对的相似程度。

(2)提出小样本场景下基于双层孪生神经网络的模型训练策略和智能合约分类方法，以克服数据类别不均衡和标签数据量过少问题。

(3)真实数据集的实验结果表明，本文所提方法在小样本场景下的分类效果优于当前最先进的分类方法，同时训练过程对标签数据的需求更低，仅需同类型其他方法约20%数据量。

2　相关工作

得益于区块链技术的快速发展，区块链以及智能合约技术与其他领域的结合应用正变得越来越流行，部署在区块链上的智能合约的数量和规模也正变得越来越庞大。由于智能合约数量的爆发式增长，一方面为了方便用户检索，另一方面为了缓解平台信息泛滥问题，对于智能合约的分类已成为当前研究的热点问题。

Bartoletti等人[13]是智能合约分类领域的首批研究者之一，他们采用人工分析的方法研究了以太坊上811个智能合约和比特币网络23个智能合约的公开源代码，并依照合约的功能特征对它们进行了分类，包括以下5个类别，金融、公证、游戏、钱包和图书馆。黄步添等人[14]提出了一种基于语义嵌入与交易信息的智能合约分类方法，为了更好地捕获智能合约代码中隐藏的语义信息，该方法引入了Word2Vec[15]词向量嵌入模型进行向量转换，之后将包含全局语义信息的词向量表示输入到长短期记忆神经网络LSTM模型中进行反复训练以获取分类模型。此外，为了提高分类模型的准确性，作者还在模型训练过程中加入了智能合约中的账户信息特征。高飞[16]设计并实现了一种基于语义信息和相似性的智能合约分类系统，该系统可以对智能合约数据进行特征提取，并选取其中具有代表性的特征词条用以计算合约相似度，最后根据合约内容输出合约类别，实现合约分类。不同于之前的基于智能合约源代码或是字节码的研究方法，Sun等人[17]认为智能合约程序二进制接口(Application Binary Interface, ABI)含有包括智能合约功能和行为的关键信息，提出了一种基于ABI粒度的智能合约自动分类方法。该方法以ABI的词频和逆向文本频率(Term Frequency–Inverse Document Frequency, TF-IDF)向量作为输入，训练了一个基于传统机器学习方法的智能合约分类模型。

部分学者除关注到智能合约本身的数据信息，引入了注意力机制[18]到模型训练过程中以提高分类准确性。吴雨芯等人[19]提出了一种基于层级注意力和双向长短期记忆神经网络(Bidirectional Long Short Term Memory, BiLSTM) [20]的智能合约分类模型基于源代码和账户的分层注意力神经网络(Hierarchical Attention Neural Network with Source Code and Account, HANN-SCA)。该模型利用Bi-LSTM网络同时捕获智能合约源代码和账

户信息特征，在特征提取和模型训练过程中，分别从词层面和句层面引入注意力机制，重点关注对分类模型建立有重要意义的句子和词语。Tian等人[21]综合考虑了智能合约数据的全局信息特征和数据中存在的语义稀疏问题，提出了一种基于Bi-LSTM模型和高斯隐含狄利克雷分布(Latent Dirichlet Allocation, LDA)的智能合约分类方法。该方法充分利用了智能合约中的辅助数据，可以将多种信息作为模型特征输入，包括源代码、账户信息、标签和注释等。并且还引入了注意力机制和语义增强方法来分别关注数据中的重要信息特征和克服合约注释存在的语义稀疏问题，以提高分类模型的可用性。

目前国内外的研究侧重于采用基于文本数据的机器学习和深度学习进行智能合约分类方法设计，包括引入合约数据的上下文信息、语义信息和账户特征等辅助数据，以及加入注意力机制、语义增强方法等来提高合约分类模型的准确性。但当前的大多数研究都仅是将传统的文本分类方法简单应用到智能合约领域，并未考虑到区块链领域和智能合约数据的特殊性，导致训练出的模型存在应用场景限制和分类性能不佳等问题。针对当前存在的问题，本文通过分析智能合约数据以及数据集的特点，提出了小样本场景下基于双层孪生神经网络的智能合约分类方法。

3 智能合约分类数据分析

智能合约数据区别于传统的文本数据，是以太坊平台上该合约用户下所有信息的集合，包括合约主体、合约注释、关键字信息、账户信息和历史交易数据等，具有特殊性和复杂性的特征。智能合约的主体指的是采用高级编程语言 Solidity编写的智能合约源代码，源代码中包含了丰富的语义信息以及注释信息。但由于每个合约创建者的代码编程习惯不同以及注释信息的存在，若要深度发掘合约主体中包含的语义信息用以建立特征，对于合约主体的预处理过程极为重要。

3.1 数据获取及预处理

由于不是每一个智能合约都包含有足够的账户信息用于特征提取，考虑到模型应用过程中数据特征的可用性，本文仅采用智能合约源代码进行数据分析和模型训练。本文通过以太坊以及区块链去中心化应用平台获取了带有类别标签的智能合约数据3 458份。进行数据整合时发现数据集存在严重的合约复制[22]现象，如果不进行处理，大量的重复数据会影响最终的分类结果。在进行去重以及删除自毁数据后得到了2 957份具有唯一精准字节码匹配的合约数据。仅包含有源代码数据和注释信息的原始合约主体无法直接用于模型训练，需要进行反编译处理为操作码格式。本文通过以太坊虚拟机数据包完成这一转换，图1展示了反编译后的操作码文件，其中包含两种不同类型的数据信息，包括操作码和指令地址，模型训练仅需保留操作码。

3.2 数据集分析

通过检查经过预处理后的原始数据集，本文发现数据集存在严重的类别不均衡以及数据量过少问题，某些类别包含近千条数据，比如游戏类别的智能合约具有864条数据；而有些类仅包含几条数据，比如保险和存储类别的智能合约都仅有个位数的数据。每个类别包含的数据量比例，如图2所示。

本文从以太坊以及去中心化应用平台所获取的原始数据集共包含21个类别，涵盖了当前智能合约应用的大多数领域，具体的类别名称和所含数据量，如表1所示。其中DEFI为去中心化金融类别，NFT为非同质化通证类别，Farm为挖矿类别，Tools为工具类别。对于深度学习模型的训练来说，过少的标签数据量无法满足最优化权值参数的训练过程，容易出现模型过拟合现象。同时，数据类别不均衡问题也极易导致分类结果出现偏差，严重影响模型的整体分类性能。

本文对反编译后的智能合约数据长度(智能合约包含的操作码数量)进行检查时发现，大部分的智能合约是较长的，而过长的数据对传统的算法模型来说通常是难以处理的，因为模型无法精准地学习到序列数据中的长期依赖关系，甚至会在模型训练过程中出现梯度爆炸问题，导致最终的分类效果不理想。

4 基于双层孪生神经网络的分类方法设计

针对以上数据分析中存在的问题，本文提出并设计了小样本场景下基于双层孪生神经网络的区块链智能合约分类方法。下面将详细介绍本文针对数据类别不均衡问题的数据欠采样处理，针对数据长度过长以及数据量过少问题设计的基于双层孪生神经网络的模型训练策略和智能合约分类方法的实现细节。

```
ISZERO
PUSH2
0×0010
JUMPI
PUSH1
0×00
DUP1
REVERT
JUMPDEST
POP
PUSH1 0×04
```

图 1 操作码文件

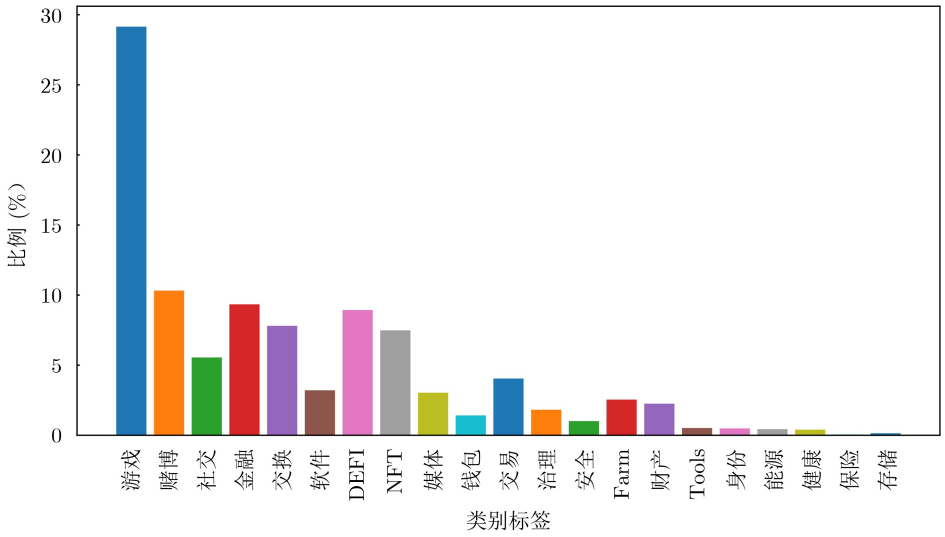


图 2 智能合约数据类别比例

表 1　智能合约类别名称及数据量

游戏	赌博	社交	金融	交换	软件	DEFI	NFT	媒体	钱包	交易
864	299	159	256	232	100	261	236	100	45	115
治理	安全	Farm	财产	Tools	身份	能源	健康	保险	存储	
52	33	72	67	15	17	15	12	2	5	

4.1　数据集欠采样

为了克服数据类别不均衡问题，本文进行了数据集欠采样处理。首先重新进行了数据整合，去除了因为标签网站长时间不更新而不具有参考价值的合约数据以及合并了一些数据量过少的类别，最后共有15个类别的智能合约，每个类别仅保留50条数据。具体的数据集类别名称，如表2所示。

4.2　双层孪生神经网络模型

为了能够有效处理长度过长的智能合约数据，本文构建了用于智能合约分类的双层孪生神经网络模型。孪生神经网络是当前深度学习领域常用的框架结构，多用于处理两个及两个以上类似输入的相似度判别问题。考虑到智能合约的数据特点以及可用数据的稀缺性，本文参考了小样本学习和度量学习的思想，引入了孪生网络来处理合约分类问题。通过双层的孪生网络充分利用较长智能合约数据的全局上下文信息，以实现智能合约精准匹配分类。模型的总体设计，如图3所示。

总体来看，模型可以分为5层、两部分。自底向上来看，分别是最底层的合约输入层、向量表示层、向量拼接层、距离计算层以及最上层的概率输出层。两部分包括处理合约的底层孪生神经网络以及类别判断的上层孪生神经网络，其中底层孪生神经网络部分包括底部的合约输入层和向量表示层，其余3层为上层孪生神经网络部分。

表 2　数据集类别

序号	类别	序号	类别
1	游戏	9	财产
2	赌博	10	媒体
3	社交	11	钱包、存储
4	金融	12	交易
5	交换	13	治理
6	软件	14	Farm
7	DEFI	15	NFT
8	Tools、能源、健康、保险		

最底层是合约输入层，输入数据来自一对智能合约数据，包括ContractA和ContractB，根据数据分析发现大部分的智能合约是较长的，普遍包含大量的操作码，为了更好地捕获输入智能合约的全局序列特征，本文参考了文献[23]中的模型构造方法在输入层将输入的智能合约数据进行了对半切割处理，对半切割后的4个前后智能合约段，分别为ContractA1, ContractA2, ContractB1, ContractB2。

之后是向量表示层，为了精准获取合约操作码的语义信息，本文选用了智能合约领域语料预训练后的Bert模型用于词向量嵌入。该模型的限定最长输入长度是512个字符，而本文所使用数据集中大部分数据长度是大于512个字符的，在经过合约输

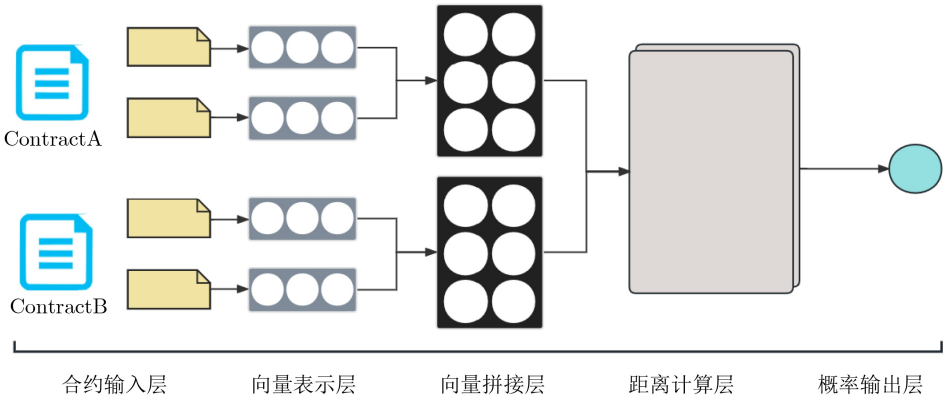


图 3 双层孪生神经网络模型

入层对半切割处理后的智能合约段符合Bert模型的输入长度限制。输入到Bert模型中进行词向量嵌入表示，在该层可以获取4个合约段ContractA1, ContractA2, ContractB1, ContractB2的向量表示$\boldsymbol{u}_1$,$\boldsymbol{v}_1$,$\boldsymbol{u}_2$,$\boldsymbol{v}_2$。

然后是向量拼接层，在该层将之前获取的来自同一合约数据的前后段向量表示进行拼接处理，分别得到一对合约数据的全局向量表示 $\boldsymbol{E}_1$和$\boldsymbol{E}_2$，可以表示为

$$\boldsymbol{E} = (\boldsymbol{u}, \boldsymbol{v}, |\boldsymbol{u} - \boldsymbol{v}|) \tag{1}$$

拼接层之上是距离计算层，本文使用L1距离即曼哈顿距离来计算一对合约数据的向量距离，希望通过模型训练使得属于同一类别合约的向量距离尽可能小，分属不同类别的向量距离尽可能大，其中L1距离及该层的向量距离公式可以表示为

$$d(x, y) = \sum_{i=1}^{n} |\boldsymbol{x}_i - \boldsymbol{y}_i| \tag{2}$$

$$D = d(\boldsymbol{E}_1, \boldsymbol{E}_2) \tag{3}$$

最顶层是概率输出层，在该层通过多个线性层以及Sigmoid激活函数输出合约数据对属于同一类别的概率，输出范围为[0,1]，其中0表示不属于同一类别，1表示属于同一类别，可以表示为

$$\text{Linear}(x) = \boldsymbol{W} \times \boldsymbol{x} + \boldsymbol{b} \tag{4}$$

$$P = \text{Sigmoid}(\text{Linear}(D)) \tag{5}$$

4.3 模型训练及分类方法设计

针对智能合约数据量过少的问题，本文提出小样本场景下基于孪生神经网络的数据对形式的模型训练策略，可以对原始数据集实现平方级的数据扩充，并基于此设计了智能合约分类方法。

首先，本文选取数据集中每个类别80%的数据用于组成数据对进行模型训练，其中同一类别的合约数据所组成的数据对视为正类数据对，标签置为1；而分属不同类别的合约数据所组成的数据对视为负类数据对，标签置为0，同时将训练数据的正负数据对数量之比控制为1:1。通过匹配组合正负数据对形式的训练策略可以对原有数据集进行有效扩充增强。

然后，在基于双层孪生神经网络的模型训练过程中，Bert词向量嵌入模型的参数也不固定，同样参与模型的反向传播，进行参数更新。目的是希望通过数据对形式的模型训练策略，最终属于同一类别的智能合约输出的向量表示在空间上是相邻的，属于不同类别的合约向量表示是彼此远离的。

最后，剩余的20%数据用于合约分类，匹配测试。在进行合约分类测试时，待测试的合约数据通过训练好的智能合约分类模型与每个类别带有标签的支撑数据进行类别相似度计算，通过相似概率叠加的方式得到相似度得分，其中相似度得分最高的类别即作为测试合约的类别。训练及分类过程，如图4所示。

在使用基于度量学习的方法进行分类时需遵循一个假设：即相同类别的数据应保持相同的数据分布情况。词云图可以直观地通过图片的方式展示文本词汇的频率情况，通过比较数据集中相同类别智能合约的操作码词云图，如图5所示，本文发现相同类别的合约具有相似的操作码类型和频率。两幅图中的操作码词云图中均包含大量高频的表示跳转类型的操作码指令，如SWAP和JUMP等。因此可以认为相同类别的智能合约保持相同的数据分布情况，数据集遵循假设，符合度量学习的使用条件。

5 实验与分析

本节所有实验均处于相同的硬件环境下，模型的训练和测试工作是在开源的深度学习框架Keras下进行的。实验的推荐参数设置如下：学习率设置为0.001，批量大小设置为16，训练轮次设置为80，Bert模型的输入最大句长设置为510(其中还需要包括[CLS]和[SEP]字符)。对于实验的评价标

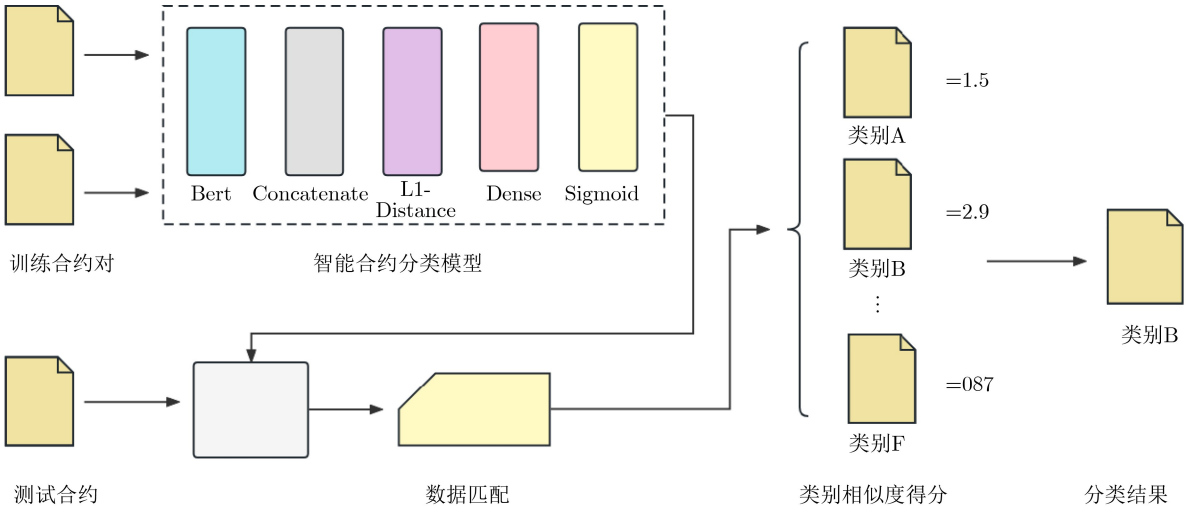
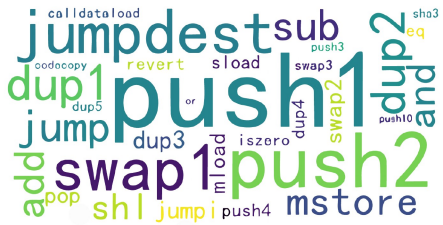
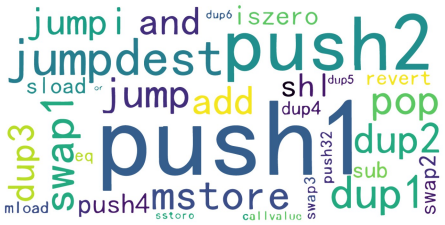


图 4 模型训练及分类流程图



(a) 游戏类别合约操作码1



(b) 游戏类别合约操作码2

图 5 操作码词云图

准，本文采用了3种较为常见的评价指标来定量评估每类实验中的分类性能，包括F1值、精确率(Precision)和召回率(Recall)。

5.1 Bert模型微调预训练分析

在双层孪生神经网络模型的向量表示层，本文引入了Bert模型[24]进行词向量嵌入转换。本文选取的是轻量级英文语料预训练的Bert模型。为了得到合适的词向量嵌入表示，本文参考了文献[25]关于Bert模型微调预训练的建议，通过获取海量的智能合约数据，重新在基于通用英文语料训练的Bert模型进行了基于智能合约领域语料的微调预训练。为了验证对Bert模型进行领域语料微调预训练是否有效，即验证加入先验知识是否会使得Bert模型更加适应于智能合约领域的训练环境，在本节设计了Bert模型是否进行领域语料微调预训练的实验，实验结果如图6所示，其中Finetune-Bert为经过智能合约领域语料微调预训练后的Bert模型，Bert为通用语料训练的Bert模型。

通过对通用语料训练的Bert模型以及加入智能合约领域语料微调预训练后的Finetune-Bert模型进行损失值下降速率对比，发现经过微调预训练后的Finetune-Bert模型在训练过程中损失值下降速率要更快一些，因此可以认为基于智能合约操作码微调预训练后输出的向量表示在词向量空间上要更加准确、贴切，更加适用于智能合约领域的分类任务，本文在后续实验部分均采用了基于智能合约操作码微调预训练后的Finetune-Bert模型。

5.2 单层/双层孪生神经网络模型分析

为了验证本文所提出的双层孪生神经网络模型是否在智能合约分类时发挥作用，在本节设计了单层/双层孪生神经网络模型的对比实验。其中Single-Siamese模型指的是未加入处理合约的底层孪生神经网络部分，对大于输入限制的合约数据进

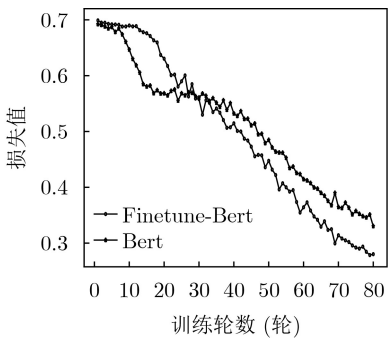


图 6 模型预训练损失值对比

行舍弃处理的单层孪生神经网络智能合约分类模型，而Double-Siamese即本文提出的完整的双层孪生神经网络智能合约分类模型。实验结果如表3和图7所示。

通过对比两类模型训练时的收敛速度以及训练完成后用于合约分类时的实验性能，验证了本文所提出的完整的双层孪生神经网络模型是有效的，在进行合约分类时性能有所提升。双层孪生神经网络模型在进行数据对训练时达到高精度的速度要明显快于单层孪生神经网络模型，而且合约分类时的准确率也要明显高于单层孪生神经网络模型，证明完整的双层孪生神经网络模型获取到的合约特征信息要更加全面，可以有效捕获智能合约数据的全局特征，提升分类的准确性。

5.3 向量拼接方式分析

为了探究如何才能更好地捕获智能合约数据的全局信息特征，本文设置了向量拼接层的向量拼接方式的性能比较实验。本实验比较的向量拼接方式包括$(\boldsymbol{u},\boldsymbol{v})$,$(|\boldsymbol{u}-\boldsymbol{v}|)$,$(\boldsymbol{u}\times\boldsymbol{v})$,$(\boldsymbol{u},\boldsymbol{v},\boldsymbol{u}\times\boldsymbol{v})$,$(|\boldsymbol{u}-\boldsymbol{v}|,\boldsymbol{u}\times\boldsymbol{v})$以及$(\boldsymbol{u},\boldsymbol{v},|\boldsymbol{u}-\boldsymbol{v}|)$，其中$\boldsymbol{u}$,$\boldsymbol{v}$指的是来自同一合约数据的前后段向量表示。实验结果如表4所示。

通过比较以上向量拼接方式的分类性能，本文发现$(\boldsymbol{u},\boldsymbol{v},|\boldsymbol{u}-\boldsymbol{v}|)$的向量拼接方式具有最好的实验效果，相比于其他的向量拼接方式，该拼接方式的设计能够更为有效地捕获并输出前后段合约的关键信息，在之后用于合约分类的效果也要更好。

5.4 数据量参数分析

本节主要实验数据量参数对于分类模型的性能影响。实验中均采用了相同数量数据对训练好的模型进行合约分类实验，但改变了每个类别用于合约分类测试的数据量，分别设置为10(8/2), 20(16/4), 30(24/6), 40(32/8), 50(40/10)条数据，同时保持相同的划分比例8:2，即当类别数据量设置为50时，每个类别会划分出10条测试数据与40条带有标签的支撑数据进行合约分类测试。实验结果，如表5所示。

本文发现随着类别数据量的增加，合约分类的准确率也在提升，数据量参数设置为50时具有最佳的分类性能，分类准确率达到了94.7%，召回率达到了94.6%，F1值也达到了94.6%。本文通过分析认为，在数据量参数设置为50时由于需要匹配计算的每个类别的支撑数据在增加，可以有效避免与错误类别匹配成功的偶然性，直接提升了总体的分类性能。但同时因为数据量参数设置的提升，需要遍历匹配的数据量在增加，合约分类所需要花费的时间也在增长。可以根据合约分类对于准确性与消耗时间需求的衡量，灵活选择用于分类的类别数据量。为了达到最优的分类性能，本文实验部分选择的类别数据量为50。

5.5 分类性能分析

最后为了验证本文所提方法的实验性能，与其他论文所提方法以及基线方法进行了对比实验。其中，基于智能合约分类的双向长短时记忆网络 (Bi-LSTM for Smart Contract Classification, SCC-BiLSTM)为文献[21]提出的方法，该方法采用了语义增强以及引入了注意力机制的Bi-LSTM模型来实现智能合约分类；支持向量机+交易信息以及神经网络+交易信息为文献[14]提出的方法，这两种方法采用了传统的机器学习算法并引入了智能合约中的交易信息特征来构建智能合约分类模型；

表 3 单双层孪生网络模型实验结果对比

模型	Precision	Recall	F1
Single-Siamese	0.880	0.873	0.874
Double-Siamese	0.947	0.946	0.946

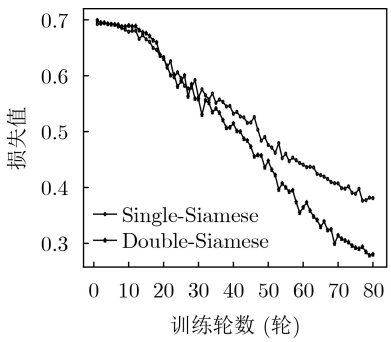


图 7 单双层孪生网络模型损失值变化

表 4 向量拼接方式实验结果对比

向量拼接方式	Precision	Recall	F1
$(\boldsymbol{u},\boldsymbol{v})$	0.848	0.844	0.845
$(\lvert\boldsymbol{u}-\boldsymbol{v}\rvert)$	0.889	0.833	0.860
$(\boldsymbol{u}\times\boldsymbol{v})$	0.894	0.880	0.886
$(\boldsymbol{u},\boldsymbol{v},\boldsymbol{u}\times\boldsymbol{v})$	0.927	0.920	0.923
$(\lvert\boldsymbol{u}-\boldsymbol{v}\rvert,\boldsymbol{u}\times\boldsymbol{v})$	0.916	0.924	0.920
$(\boldsymbol{u},\boldsymbol{v},\lvert\boldsymbol{u}-\boldsymbol{v}\rvert)$	0.947	0.946	0.946

表 5 类别数据量实验结果

类别数据量	Precision	Recall	F1
10(8/2)	0.867	0.833	0.849
20(16/4)	0.921	0.900	0.910
30(24/6)	0.928	0.922	0.924
40(32/8)	0.930	0.925	0.927
50(40/10)	0.947	0.946	0.946

HANN-SCA为文献[19]提出的一种基于层级注意力机制与Bi-LSTM神经网络的智能合约自动分类模型。Double-Siamese为本文提出的基于双层孪生神经网络的智能合约分类方法。其中对比方法遵循原论文的实验环境设置，所用数据是在以太坊或者区块链去中心化应用平台获取的原始的类别不均衡的智能合约数据集，本文所提方法是在小样本数据环境下进行的，对数据进行了过采样处理，每个类别都仅保留50条合约数据。与其他方法的分类性能比较如表6所示。

表 6　合约分类对比实验结果

模型	Precision	Recall	F1
支持向量机+交易信息	0.852	0.854	0.852
神经网络+交易信息	0.889	0.881	0.849
HANN-SCA	0.931	0.920	0.926
SCC-BiLSTM	0.917	0.906	0.911
Double-Siamese	0.947	0.946	0.946

实验结果表明，相比于其他方法，本文所提方法具有最优的分类性能。但值得注意的是，为了达到最佳的分类效果，本文所提方法对内存的要求较高，因为该方法需要存储所有的带有标签的支撑数据，同时在合约分类预测阶段也需要花费较多时间，因为该方法需要遍历所有的支撑数据，逐一匹配计算相似性。但本文所提方法可以根据分类性能高低与耗费时间长短的不同任务需要，通过灵活调整用于分类匹配的数据量参数来实现不同场景下的不同需求。

6　结论

本文针对当前智能合约分类领域存在的问题，提出基于双层孪生神经网络的智能合约分类模型，并通过分析智能合约数据集的特点，设计了小样本场景下的智能合约分类方法。首先，进行数据欠采样处理，对每个类别的智能合约仅保留相同数目少量的合约数据。其次，针对智能合约数据长度过长的特点，设计了可以分别获取前后段合约特征的底层孪生网络部分。然后，为了捕获全局数据特征，构建了可以进行向量拼接、合约对相似度判断的上层孪生神经网络部分。最后，提出了小样本场景下基于双层孪生神经网络的模型训练策略和智能合约分类方法，并通过实验验证了本文所提方法的可行性和有效性。

参 考 文 献

[1] SZABO N. Formalizing and securing relationships on public networks[J]. *First Monday*, 1997, 2(9). doi: 10.5210/fm.v2i9.548.

[2] MOHANTA B K, PANDA S S, and JENA D. An overview of smart contract and use cases in blockchain technology[C]. The 9th International Conference on Computing, Communication and Networking Technologies, Bengaluru, India, 2018: 1–4. doi: 10.1109/ICCCNT.2018.8494045.

[3] GONG Jianhu and NAVIMIPOUR N J. An in-depth and systematic literature review on the blockchain-based approaches for cloud computing[J]. *Cluster Computing*, 2022, 25(1): 383–400. doi: 10.1007/s10586-021-03412-2.

[4] 牛淑芬, 杨平平, 谢亚亚, 等. 区块链上基于云辅助的密文策略属性基数据共享加密方案[J]. 电子与信息学报, 2021, 43(7): 1864–1871. doi: 10.11999/JEIT200124.
NIU Shufen, YANG Pingping, XIE Yaya, *et al*. Cloud-assisted ciphertext policy attribute based encryption data sharing encryption scheme based on BlockChain[J]. *Journal of Electronics & Information Technology*, 2021, 43(7): 1864–1871. doi: 10.11999/JEIT200124.

[5] ABDELMABOUD A, AHMED A I A, ABAKER M, *et al*. Blockchain for IoT applications: Taxonomy, platforms, recent advances, challenges and future research directions[J]. *Electronics*, 2022, 11(4): 630. doi: 10.3390/electronics11040630.

[6] JOHARI R, KUMAR V, GUPTA K, *et al*. BLOSOM: BLOckchain technology for security of medical records[J]. *ICT Express*, 2022, 8(1): 56–60. doi: 10.1016/j.icte.2021.06.002.

[7] ZHENG Zibin, XIE Shaoan, DAI Hongning, *et al*. An overview on smart contracts: Challenges, advances and platforms[J]. *Future Generation Computer Systems*, 2020, 105: 475–491. doi: 10.1016/j.future.2019.12.019.

[8] CHEN Weili, ZHENG Zibin, NGAI E C H, *et al*. Exploiting blockchain data to detect smart Ponzi schemes on Ethereum[J]. *IEEE Access*, 2019, 7: 37575–37586. doi: 10.1109/ACCESS.2019.2905769.

[9] TORRES C F, STEICHEN M, and STATE R. The art of the scam: Demystifying honeypots in Ethereum smart contracts[C]. The 28th USENIX Conference on Security Symposium, Santa Clara, USA, 2019: 1591–1607.

[10] LI Yikai, CHEN C L P, and ZHANG Tong. A survey on Siamese network: Methodologies, applications, and opportunities[J]. *IEEE Transactions on Artificial Intelligence*, 2022, 3(6): 994–1014. doi: 10.1109/TAI.2022.3207112.

[11] ZHANG Jianwei, ZHANG Xubin, LV Lei, *et al*. An applicative survey on few-shot learning[J]. *Recent Patents on Engineering*, 2022, 16(5): 104–124. doi: 10.2174/1872212115666210715121344.

[12] WANG Zhenzhi, WANG Limin, WU Tao, *et al*. Negative sample matters: A renaissance of metric learning for temporal grounding[C]. The 36th AAAI Conference on Artificial Intelligence, Vancouver, Canada, 2022: 2613–2623. doi: 10.1609/aaai.v36i3.20163.

[13] BARTOLETTI M and POMPIANU L. An empirical analysis of smart contracts: Platforms, applications, and design patterns[C]. The International Conference on Financial Cryptography and Data Security, Sliema, Malta, 2017: 494–509. doi: 10.1007/978-3-319-70278-0_31.

[14] 黄步添, 刘琦, 何钦铭, 等. 基于语义嵌入模型与交易信息的智能合约自动分类系统[J]. 自动化学报, 2017, 43(9): 1532–1543. doi: 10.16383/j.aas.2017.c160655.

HUANG Butian, LIU Qi, HE Qinming, *et al*. Towards automatic smart-contract codes classification by means of word embedding model and transaction information[J]. *Acta Automatica Sinica*, 2017, 43(9): 1532–1543. doi: 10.16383/j.aas.2017.c160655.

[15] MIKOLOV T, SUTSKEVER I, CHEN Kai, *et al*. Distributed representations of words and phrases and their compositionality[C]. The 26th International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, 2013: 3111–3119.

[16] 高飞. 基于区块链技术的智能合约自动分类系统设计[J]. 高原科学研究, 2018, 2(4): 51–59. doi: 10.16249/j.cnki.2096-4617.2018.04.007.

GAO Fei. Design of intelligent contract automatic classification system based on blockchain technology[J]. *Plateau Science Research*, 2018, 2(4): 51–59. doi: 10.16249/j.cnki.2096-4617.2018.04.007.

[17] SUN Xun, LIN Xingwei, and LIAO Zhou. An ABI-based classification approach for Ethereum smart contracts[C]. 2021 IEEE International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, International Conference on Cyber Science and Technology Congress, Calgary, Canada, 2021: 99–104. doi: 10.1109/DASC-PICom-CBDCom-CyberSciTech52372.2021.00029.

[18] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[C]. The 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 6000-6010.

[19] 吴雨芯, 蔡婷, 张大斌. 基于层级注意力机制与双向长短期记忆神经网络的智能合约自动分类模型[J]. 计算机应用, 2020, 40(4): 978–984. doi: 10.11772/j.issn.1001-9081.2019081327.

WU Yuxin, CAI Ting, and ZHANG Dabin. Automatic smart contract classification model based on hierarchical attention mechanism and bidirectional long short-term memory neural network[J]. *Journal of Computer Applications*, 2020, 40(4): 978–984. doi: 10.11772/j.issn.1001-9081.2019081327.

[20] ENAMOTO L, SANTOS A R A S, MAIA R, *et al*. Multi-label legal text classification with BiLSTM and attention[J]. *International Journal of Computer Applications in Technology*, 2022, 68(4): 369–378. doi: 10.1504/IJCAT.2022.125186.

[21] TIAN Gang, WANG Qibo, ZHAO Yi, *et al*. Smart contract classification with a Bi-LSTM based approach[J]. *IEEE Access*, 2020, 8: 43806–43816. doi: 10.1109/ACCESS.2020.2977362.

[22] LIU Han, YANG Zhiqiang, LIU Chao, *et al*. EClone: Detect semantic clones in Ethereum via symbolic transaction sketch[C]. The 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Lake Buena Vista, USA, 2018: 900–903. doi: 10.1145/3236024.3264596.

[23] REIMERS N and GUREVYCH I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks[C]. The 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 2019. doi: 10.18653/v1/D19-1410.

[24] DEVLIN J, CHANG Mingwei, LEE K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding[C]. The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, USA, 2019. doi: 10.18653/v1/N19-1423.

[25] SUN Chi, QIU Xipeng, XU Yige, *et al*. How to fine-tune BERT for text classification?[C]. The 18th China National Conference on Chinese Computational Linguistics, Kunming, China, 2019: 194–206. doi: 10.1007/978-3-030-32381-3_16.

郭加树：男，博士，副教授，研究方向为机器学习、人工智能等.
王 琪：男，硕士生，研究方向为区块链技术、数据挖掘.
李择亚：女，硕士生，研究方向为联邦学习.
武梦德：男，硕士生，研究方向为服务计算.
张红霞：女，博士，副教授，研究方向为边缘计算、区块链技术、服务计算等.

责任编辑：余 蓉