Journal of Computer Applications
计算机应用,2021,41(2):363 - 371
ISSN 1001-9081
CODEN JYIIDU
2021-02-10
http://www.joca.cn

文章编号:1001-9081(2021)02-0363-09
DOI:10.11772/j.issn.1001-9081.2020050609

面向区块链的在线联邦增量学习算法

罗长银[1,2,3],陈学斌[1,2,3*],马春地[1],王君宇[1,2,3]
(1.华北理工大学 理学院,河北 唐山 063210; 2.河北省数据科学与应用重点实验室(华北理工大学),河北 唐山 063210;
3.唐山市数据科学重点实验室(华北理工大学),河北 唐山 063210)
(*通信作者电子邮箱chxb@qq.com)

摘 要:针对传统数据处理技术存在模型过时、泛化能力减弱以及并未考虑多源数据安全性的问题,提出一种面向区块链的在线联邦增量学习算法。该算法将集成学习与增量学习应用到联邦学习的框架下,使用stacking集成算法来整合多方本地模型,且将模型训练阶段的模型参数上传至区块链并快速同步,使得在建立的全局模型准确率仅下降1%的情况下,模型在训练阶段与存储阶段的安全性均得到了提升,降低了数据存储与模型参数传输的成本,同时也降低了因模型梯度更新造成数据泄漏的风险。实验结果表明,在公开的数据集上进行训练,各时间段内模型的准确度均在91.5%以上,且方差均低于10^{-5};与传统整合数据训练模型相比,该模型在准确率上略有下降,但能够在保证模型准确率的同时提高数据与模型的安全性。

关键词:区块链;集成学习;联邦学习;增量学习
中图分类号:TP391 **文献标志码**:A

Online federated incremental learning algorithm for blockchain

LUO Changyin[1,2,3], CHEN Xuebin[1,2,3*], MA Chundi[1], WANG Junyu[1,2,3]
(1. *School of Science, North China University of Science and Technology, Tangshan Hebei* 063210, *China*;
2. *Hebei Key Laboratory of Data Science and Applications (North China University of Science and Technology), Tangshan Hebei* 063210, *China*;
3. *Tangshan Data Science Laboratory (North China University of Science and Technology), Tangshan Hebei* 063210, *China*)

Abstract: As generalization ability of the out-dated traditional data processing technology is weak, and the technology did not take into account the multi-source data security issues, a blockchain oriented online federated incremental learning algorithm was proposed. Ensemble learning and incremental learning were applied to the framework of federated learning, and stacking ensemble algorithm was used to integrate the local models and the model parameters in model training phase were uploaded to the blockchain with fast synchronization. This made the accuracy of the constructed global model only fall by 1%, while the safety in the stage of training and the stage of storage was improved, so that the costs of the data storage and the transmission of model parameters were reduced, and at the same time, the risk of data leakage caused by model gradient updating was reduced. Experimental results show that the accuracy of the model is over 91.5% and the variance of the model is lower than 10^{-5}, and compared with the traditional integrated data training model, the model has the accuracy slightly reduced, but has the security of data and model improved with the accuracy of the model guaranteed.

Key words: blockchain; ensemble learning; federated learning; incremental learning

0 引言

谷歌于2016年提出了一种新兴的隐私保护技术——联邦学习[1],因其具有保护隐私和本地数据安全的优势,被广泛应用于多个领域。

联邦学习可以应用至金融领域,如杨强教授团队将联邦学习应用在小额信贷的风险管理、反洗钱等案例中[2];联邦学习还可应用于语音识别,如:保险客服的语音识别与质量检测的语音识别,采用联邦学习的框架建立二者共享的语音识别(Automatic Speech Recognition, ASR)模型,并取得了很好的收益。

联邦学习的训练数据来源于不同数据源,导致训练数据的分布与数量成为影响联邦模型的条件。若数据源的训练数据分布不同,那么整合多方本地模型就成为难题。文献[3]使用Logistic回归模型作为初始全局模型对各数据源的数据进行训练,采用神经网络来整合本地模型;但神经网络模型的表现为非凸函数,很难使参数平均化后的模型损失函数[4]达到最优。针对此问题,文献[5]中提出了联邦平均算法FedAvg,采用权重或梯度的平均值来整合多方本地模型后获得整合的全局模型。但文献[6]针对联邦平均算法FedAvg提出了深度梯度泄漏算法,能够根据本地模型的梯度更新还原出大部分训练数据。同时,上述文献均没有考虑联邦模型时效性的

收稿日期:2020-05-16;**修回日期**:2020-07-22;**录用日期**:2020-07-31。
基金项目:国家自然科学基金资助项目(61572170,61170254);唐山市科技项目(18120203A)。
作者简介:罗长银(1994—),男,陕西安康人,硕士研究生,CCF会员,主要研究方向:数据安全; 陈学斌(1970—),男,河北唐山人,教授,博士,CCF高级会员,主要研究方向:数据安全、物联网安全、网络安全; 马春地(1999—),男,河北唐山人,主要研究方向:网络安全;王君宇(1996—),女,河北唐山人,硕士研究生,主要研究方向:网络安全。

问题。

针对数据安全性与模型时效性问题，本文提出了一种面向区块链的在线联邦增量学习算法。该算法利用集成学习的思想来整合多方本地模型，以训练出多方都满足的全局模型。训练结果表明，该算法的准确度比传统整合数据训练模型的方法略有降低，但数据与模型在训练模型阶段的安全性得到提升；同时相较一般联邦学习模型，该算法可以将每个时间段、每次迭代的参数与结果自动上传至对应的数据块中并快速同步，数据传输成本大大降低；而且因区块链数据不可篡改与不可删除的特点，使模型在存储阶段拥有双重保障敏感数据的安全性。

1 相关知识

1.1 区块链

区块链的概念在文献[7]中首次提出，是分布式数据存储、点对点传输、共识机制、加密算法、时间戳、哈希算法以及相关计算机技术组成互联网时代的创新应用模式。

区块链因其具有去中心化、数据不可篡改、数据安全可靠与可溯源以及集体维护的优势，所以被广泛应用。例如：文献[8]利用区块链去中心化的特点，提出了基于区块链的电子健康记录安全存储模型。

区块链由多个区块连接而成，它利用哈希算法对每个数据区块的头部进行运算可以得到一个哈希值，使用这个哈希值可以将区块之间连接起来构成一条链，此为区块链结构的本质。数据区块当中记录了当前时间下的交易信息，采用默克尔树信息进行保存。每一个数据区块由区块头与区块体组成，数据区块的结构如图1所示。

1.2 RSA加密算法

RSA加密算法是Ron Rivest、Adi Shamir和Leonard Adleman三人于1977年在文献中首次提出[9]，因其是一种非对称加密算法，所以在公开密钥加密和电子商业中被广泛应用。例如：文献[10]提出的基于强认证技术的会话初始协议安全认证模型使用RSA数字签名来保证消息传输的机密性、真实性、完整性和不可否认性。

1.3 椭圆曲线数字签名算法

椭圆曲线数字签名算法(Elliptic Curve Digital Signature Algorithm，ECDSA)[11]是使用椭圆曲线加密(Elliptic Curve Cryptography，ECC)算法对DSA(Digital Signature Algorithm)的模拟，在2000年成为IEEE和NIST的标准，具有生成的密钥长度短而且签名和验证速度更快，在具有相同安全性的情况下，所需的存储资源更少的优点，因此被广泛应用。

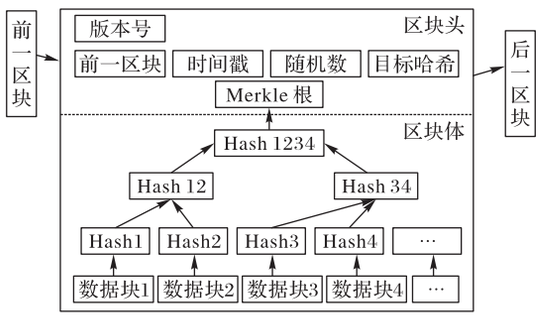

图1 区块链结构示意图

Fig. 1 Schematic diagram of blockchain structure

1.4 联邦学习

联邦学习是隐私保护下的算法优化可实现路径和保护数据安全的“数据孤岛”问题的解决方案[12]。具体实现过程为：对多个参与方在本地私有数据上进行模型训练，然后将不同的模型参数上传到云端进行整合和更新，之后将更新的参数发送至各参与方。整个过程私有数据不出本地，既保证了数据隐私，同时解决了各参与方“数据孤岛”的困境，根据其实现过程使得联邦同样具有联邦学习保护隐私和本地数据安全的优势。

1.5 集成学习

在有监督学习算法[13]中，训练出的模型在满足稳定性的同时还要求模型各方面的性能都较好，但实际情况是有时只能得到多个有所偏好的模型(弱监督模型[14])。集成学习[15]就是将多个弱监督模型组合成一个更好、更全面的强监督模型，集成学习的思想是即使某一个弱分类器得到了错误的预测，其他的弱分类器也可以将错误纠正回来。在集成学习中，stacking集成是目前提升机器学习性能最有效的方法[16-18]。从图2中可以看出，stacking集成分为两步：1)使用多个算法求出结果；2)将结果作为特征输入到下一个算法中训练出最终的预测结果[19-21]。

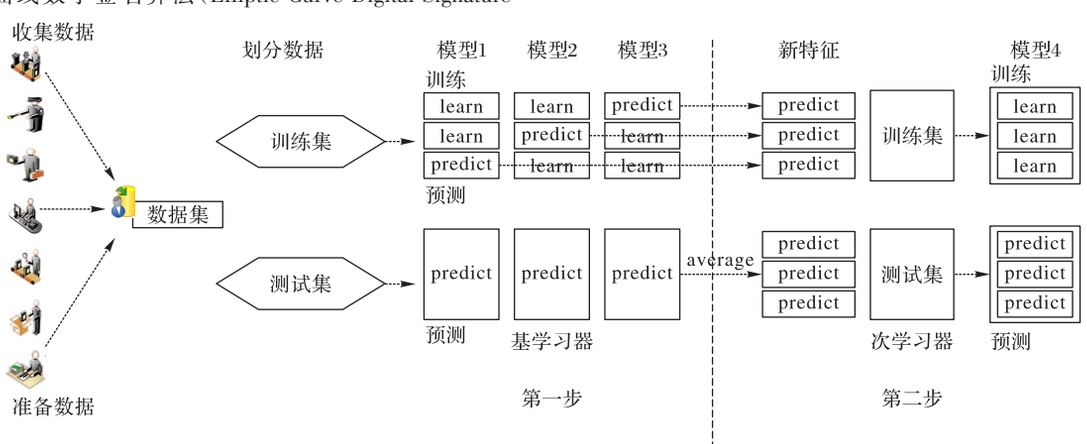

图2 stacking集成示意图

Fig. 2 Schematic diagram of stacking ensemble

2　在线联邦增量学习算法

2.1　算法描述

在线联邦增量学习算法是在联邦学习的框架与集成学习的思想下建立的，图3为该算法的整体框架。从图3中可以看出，该算法包括数据收集阶段、模型训练阶段和模型存储阶段三个阶段。在数据收集阶段使用数字签名来保证数据的安全性与完整性；在训练模型阶段采用联邦学习框架与增量学习算法，保证了数据的安全性与模型的时效性；在模型存储阶段，采用区块链来存储每个时间段内各模型的参数，使数据传输成本大大降低，同时使数据的安全性得到保障。

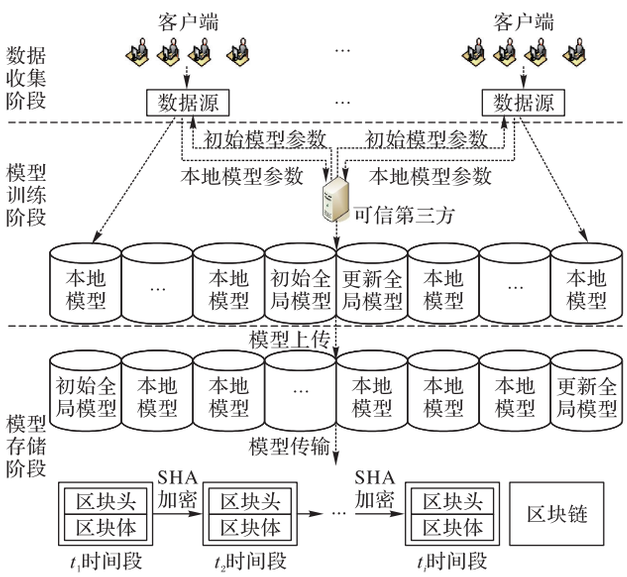

图3　基于区块链的联邦增量学习算法的整体框架

Fig. 3　Overall framework of federated incremental learning algorithm based on blockchain

数据收集阶段的算法如图4所示，具体流程如下：1）各客户端需要计算数据的hash值，并使用由RSA加密算法产生的公钥来加密hash值，再传输至各数据源。2）各数据源使用私钥解密，并重新计算数据的hash值。3）判断解密得到的hash值与重新计算的hash值是否相等：若相等，需要将数据存储到数据源中，等待模型训练；若不相等，表明此客户端的数据在传输过程中被篡改，以此可保证数据收集阶段数据的安全性与完整性。

模型训练阶段的算法如图5所示，具体流程如下：1）由可信第三方使用RSA加密算法将初始全局模型传输至各数据源，保证模型安全性；2）各数据源将解密的初始全局模型在历史数据与增量数据上进行训练，获得每个时间段内的本地模型；3）将本地模型传输可信第三方，可信第三方使用stacking集成算法来整合多个本地模型，获得每个时间段更新的全局模型，且不断迭代训练。其中，该阶段训练模型的公式如下：

初始全局模型分为三种情况：

1）$H_i = voting\{max\_score(s_1(D_{ij}), s_2(D_{ij}), s_3(D_{ij}))\}_{j=1}^{n+l}$，$l$ 表示新加入的数据源数量；

2）$H_i = voting\{max\_score(s_1(D_{ij}), s_2(D_{ij}), s_3(D_{ij}))\}_{j=1}^{n-k}$，$k$ 表示减少的数据源数量；

3）$H_i = h_{i-1}$，无新增数据源。

本地模型：$h_{ij} = training(H_i, D_{ij})$。

更新的全局模型：$h_i = stacking(\{h_{ij}\}_{j=1}^{n})$。

其中：i 表示当前为训练模型的次数；n 表示上一次训练模型时的数据源个数；j 表示当前训练的数据源；D_{ij} 表示增量数据集；$s_1(D_{ij})$、$s_2(D_{ij})$、$s_3(D_{ij})$ 表示初始化参数的随机森林 m_1、朴素贝叶斯 m_2、神经网络 m_3 在增量数据集上训练的模型。

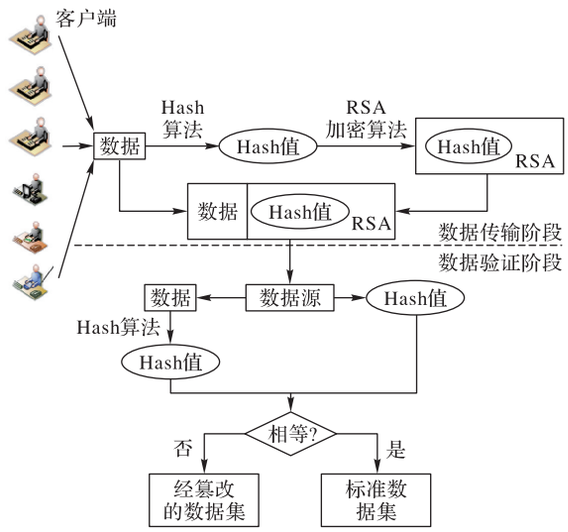

图4　数据收集阶段算法流程

Fig. 4　Flowchart of data collection stage

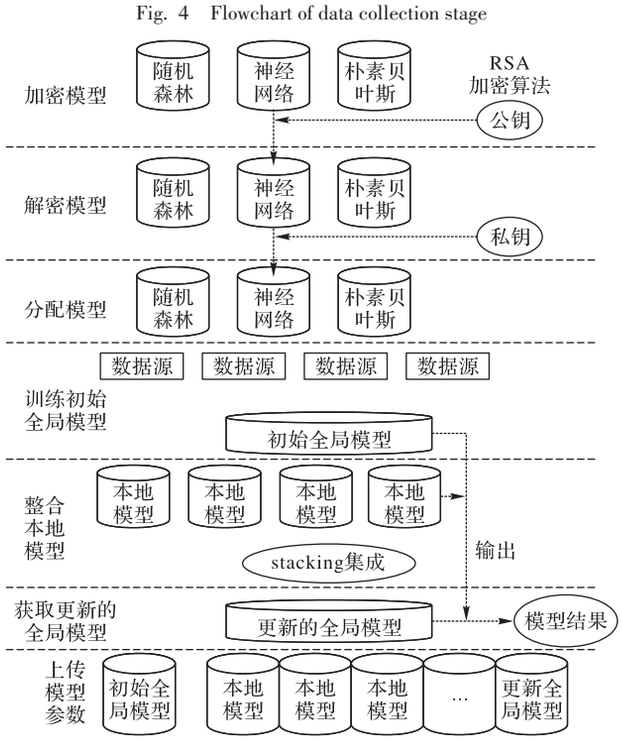

图5　模型训练阶段算法流程

Fig. 5　Flowchart of model training stage

模型存储阶段的算法如图6所示，可以看到将各个时间段内本地模型的参数使用ECDSA上传至数据块2至 $n-1$ 中，数据块1中存储的是本轮时间段内初始全局模型，数据块 n 存储的是本轮更新后的全局模型，以此来保证 t_i 时间段的本地模型以及全局模型的安全，利用区块链的不可逆和不可篡改以及可追溯的特点来保证模型层面上的安全。

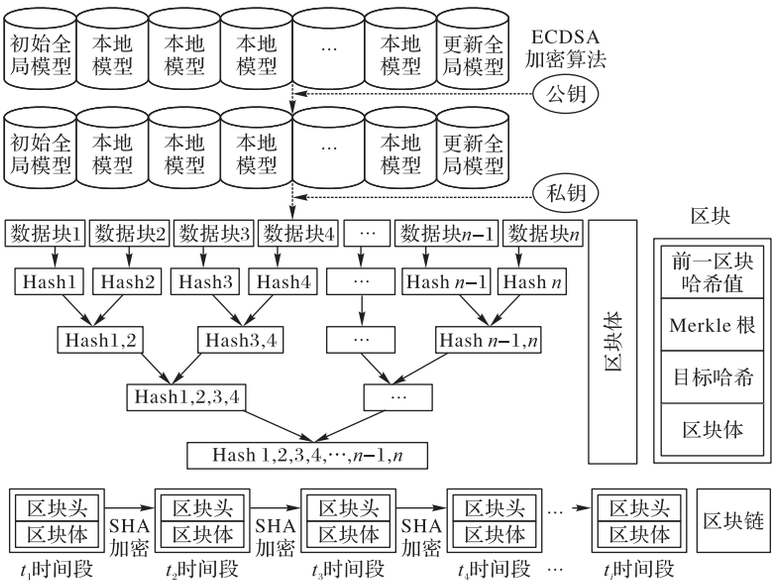

图6 模型存储阶段算法流程

Fig. 6 Flowchart of model storage stage

综上所述，基于区块链的在线联邦增量学习算法的流程如下：

1)数据收集阶段。

步骤1 由可信第三方使用RSA加密算法产生的公钥传输至各数据源与客户端；

步骤2 客户端计算数据的hash值，并将hash值加密与数据共同传输至数据源；

步骤3 数据源使用私钥解密，并重新计算数据的hash值，将传输前后计算的hash相等的数据存储至数据源，不相等的删除。

2)模型训练阶段。

步骤1 各数据源使用RSA加密算法产生的公钥传输至可信第三方；

步骤2 可信第三方将初始化参数的随机森林、朴素贝叶斯、神经网络使用公钥加密并传输至各数据源；

步骤3 各数据源使用初始化参数的3种模型在历史数据上进行训练，根据3种初始化参数模型的准确度计算其平均值与方差来衡量模型的性能，将性能最优的作为初始全局模型；

步骤4 从t_2时间段起，判断是否有新增加或减少的数据源；

步骤5 若出现新增加的数据源，将初始化参数的随机森林、朴素贝叶斯、神经网络传输至新的数据源，并重新计算并选择性能最优的作为t_i时间段的初始全局模型；

步骤6 若减少了数据源，需要根据上一次训练初始全局模型时的准确度重新计算并选择性能最优的作为t_i时间段的初始全局模型；

步骤7 若无新增或减少数据源，将t_{i-1}时间段上更新的全局模型作为t_i时间段上的初始全局模型；

步骤8 将初始全局模型在t_i时间段产生的增量数据上进行训练，获得t_i时间段的本地模型；

步骤9 各数据源将t_i时间段的本地模型传输至可信第三方；

步骤10 可信第三方使用stacking集成算法来整合t_i时间段的多方本地模型，获得t_i时间段上的更新的全局模型。

3)模型存储阶段。

步骤1 将可信第三方在t_i时间段上使用ECDSA算法产生的私钥传输至各数据源，公钥传输至区块i；

步骤2 各数据源使用私钥加密t_i时间段的初始全局模型参数、本地模型参数、更新的全局模型参数并传输至区块i；

步骤3 区块i使用公钥解密并将初始全局模型参数、本地模型参数、更新的全局模型参数依次存储至数据块1, 2, …, n中。

算法伪代码如下：

Input: historical data $S_{j0}=\{x_{j0},y_{j0}\}_{j=1}^{n}$, Incremental dataset $D_{ij}=\{\{x_{ij},y_{ij}\}_{j=1}^{n}\}_{i=1}^{n}$, i represents the number of incremental data, j represents the number of data sources, Random Forest m_1, Naive Bayes m_2, neural network m_3.

Output: Hash encrypted dataset S'_{i0}、E_{ij}; RSA generate key pair P_{ij}、p_{ij}、P'_{ij}、p'_{ij}; ECDSA generate key pair P''_{i+2j}、p''_{i+2j}; original global model H_i, local model h_{ij}, updatated global model h_i.

1) Data collection stage
2) for $i=0$ to n
3) for $j=1$ to n
4) $S'_{j0}, E_{ij}=E\_hash(S_{j0}, D_{ij})$
5) return S'_{j0}, E_{ij}
6) end for
7) end for
8) Generate secret key
9) for $i=0$ to n
10) for $j=1$ to n
11) Trusted third party: $P_{ij}=G_{\text{public}}\_RSA(x)$, $p_{ij}=G_{\text{private}}\_RSA(x)$, $P''_{i+2j}=G_{\text{public}}\_ECDSA(x)$, $p''_{i+2j}=G_{\text{private}}\_ECDSA(x)$, x represents a random number
12) Various data sources: $P'_{ij}=G_{\text{public}}\_RSA(x)$, $p'_{ij}=G_{\text{private}}\_RSA(x)$, x represents a random number
13) end for
14) end for
15) Validation dataset

16) for $i = 0$ to n
17) for $j = 1$ to n
18) Clients: $S''_{j0}, D'_{ij} = E\_P_{ij}(S'_{j0}, E_{ij})$
19) Various data sources $S'_{j0}, E_{ij} = D\_P'_{ij}(S''_{j0}, D'_{ij})$ and recalculate: $S''_{j0}, D''_{ij} = E\_hash(S_{j0}, D_{ij})$
20) if $S'_{j0}, E_{ij} \neq S''_{j0}, D''_{ij}$
21) delete S_{j0}, D_{ij}
22) exit
23) else
24) send S_{j0}, D_{ij} to various data sources
25) end for
26) end for
27) Original model training stage
28) Trusted third party: $Y_1, Y_2, Y_3 = E\_RSA(m_1, m_2, m_3)$ send to Various data sources
29) Various data sources: $m_1, m_2, m_3 = D\_RSA(Y_1, Y_2, Y_3)$
30) $s_1(S_{j0}), s_2(S_{j0}), s_3(S_{j0}) = m_1\_train(S_{j0}), m_2\_train(S_{j0}), m_3\_train(S_{j0})$
31) $H_0 = voting\{max\_score(s_1(S_{j0}), s_2(S_{j0}), s_3(S_{j0}))\}_{j=1}^{n}$
32) for $j = 1$ to n
33) $\{h_{j0}\}_{j=1}^{n} = \{H_{j0}(S_{j0})\}_{j=1}^{n}$
34) Trusted third party: $h_1 = stacking(\{h_{j0}\}_{j=1}^{n})$
35) end for
36) model updated stage
37) for $i = 1$ to n
38) if add data sources
39) $s_1(D_{li}), s_2(D_{li}), s_3(D_{li}) = m_1\_train(D_{li}), m_2\_train(D_{li}), m_3\_train(D_{li})$
40) $H_i = voting\{max\_score(s_1(D_{ij}), s_2(D_{ij}), s_3(D_{ij}))\}_{j=1}^{n+l}$, l represents the number of data sources increases
41) exit
42) if reduce data sources
43) $s_1(D_{ki}), s_2(D_{ki}), s_3(D_{ki}) = m_1\_train(D_{ki}), m_2\_train(D_{ki}), m_3\_train(D_{ki})$
44) $H_i = voting\{max\_score(s_1(D_{ij}), s_2(D_{ij}), s_3(D_{ij}))\}_{j=1}^{n-k}$, k represents the number of data sources to reduce
45) exit
46) else
47) $h_{i-1} = H_i$
48) $h_{ij} = train(H_i)$
49) Trusted third party: $h_i = stacking(\{h_{ij}\}_{j=1}^{n})$
50) end for
51) model saving stage
52) for $i = 0$ to n
53) for $j = 1$ to n
54) Various data sources: $H'_i, h'_{ij}, h'_i = E\_P''_{i+2j}(H_i, h_{ij}, h_i)$
55) Various data sources send H'_i, h'_{ij}, h'_i to block i
56) block i: $H_i, h_{ij}, h_i = D\_p''_{i+2j}(H'_i, h'_{ij}, h'_i)$ and save data block 1, data blocks 2 to $n-1$, data block n
57) end for
58) end for

2.2 性能分析

2.2.1 算法的复杂度分析

联邦增量学习算法的复杂度为Hash算法的复杂度、RSA加密算法的复杂度[15]、ECDSA数字签名算法的复杂度、首轮最优的初始全局模型的复杂度、模型传输的复杂度、首轮模型整合的复杂度、模型更新的复杂度、模型存储的复杂度之和，即时间复杂度为$O(((n*\log(n)*d*k) + N^3 + W^{k+2} + G^{k+2})*l)$，其中：$n$表示样本数，$d$表示特征维度总数，$k$示决策树数量，$N$表示加密算法的复杂度，$W$表示模型传输的复杂度，$G$表示模型传输的复杂度，$l$表示轮数。采用联邦学习与stacking集成算法将必然造成此算法的时间复杂度和空间复杂度均高于传统的数据融合算法。传统数据处理方法的时间复杂度为：$O((n*\log(n)*d*k) + N^3 + W + G)$，其中，$n$表示总体样本数，$d$表示特征维度，$k$表示决策树数量，$N$表示加密算法的复杂度，$W$表示模型传输的复杂度，$G$表示模型存储的复杂度。因本文算法采用增量上传，联邦增量算法的时间复杂度为：$O((n/l*\log(n/l)*d*k) + N^3 + W + G)$，通信开销节约的时间复杂度为$O(\log n^{(n-n/l)*l^{n/l}})$。采用联邦学习与stacking集成算法在增量数据上进行训练，在保证了所训练模型准确率的情况下，时间复杂度比传统数据处理技术要低。

2.2.2 算法的安全性分析

联邦增量学习算法使用了联邦学习的框架与区块链的性质，从数据层面上，使用RSA加密算法对每个客户端的数据进行hash计算，并将hash值与数据共同传输至各数据源，各数据源重新计算其hash值，可保证数据在收集阶段的安全性与完整性。从模型层面上可分为两部分：1)从模型传输的角度，使用RSA加密算法产生的公钥P_{ij}对初始全局模型H_i进行加密并传输至各数据源；各数据源使用私钥p_{ij}解密并进行训练，可保证模型传输过程中的安全性。2)从模型存储的角度，由可信第三方使用ECDSA产生密钥对P''_{i+2j}, p''_{i+2j}，使用私钥p''_{i+2j}对t_i时间段的初始全局模型H_i、本地模型$h_{i1}, h_{i2}, \cdots, h_{in}$和更新的全局模型$h_i$进行签名并传输至区块$i$，区块$i$使用公钥$P''_{i+2j}$进行验证并依次存储区块$i$的数据块中(其中，初始全局模型$H_i$存储至数据块1，本地模型$h_{i1}, h_{i2}, \cdots, h_{in}$存储至数据块$2, 3, \cdots, n-1$，更新的全局模型$h_i$存储至数据块$n$)，可保证模型存储过程中的安全性。

2.2.3 算法的时效性分析

联邦增量学习算法采用增量学习的思想：在数据层面上，将各数据源在$[t_{i-1}, t_i](i = 1, 2, \cdots, n)$时间段内所产生的数据表示为$t_{i-1}$时间段上训练模型时的数据，从而可保证数据层面的时效性；在模型层面上，将t_{i-1}时间段更新的全局模型h_{i-1}作为t_i时间段的初始全局模型h_i进行训练，可得t_i时间段的本地模型$h_{i1}, h_{i2}, \cdots, h_{in}$及更新的全局模型$h_i$，从而可保证模型层面的时效性。

综上可以看出，本文提出的在线联邦增量学习算法的准确性比传统的数据融合模型略有下降，但具有很高的时效性与安全性。

3 实验分析

3.1 实验参数设置

该算法由python语言和pycharm集成软件开发实现，实验硬件环境为：Inter Core i5-4200M CPU 2.50 GHz处理器，内存8 GB；操作系统为Windows 10。在实验数据方面，采用从http://sofasofa.io/competition.php?id=2下载的数据集，该数据集有15.6 MB。

3.2 实验数据分析

将数据集随机划分成17份表示不同数据源($k = 1, 2, \cdots, 17$)在不同时间段内所产生的数据，随机对数据集划分100次能更合理地表示各数据源中的数据。随机划分100

次的数据集反映出划分前后数据的变化关系，随机划分数据集可以满足数据源特征相同样本不同的需求，以及可以满足交叉验证模型的合理性。使用RSA加密算法产生的公钥来加密数据的hash值，并与数据共同传输至各数据源，各数据源使用私钥解密，且重新计算数据的hash值，判断数据经过传输的hash值与传输前的hash值是否相等，将hash值相等的数据存储至各数据源内，可保证数据在收集阶段的安全性与完整性。

3.3 实验模型分析

本文实验分为四个部分，第一部分：将随机森林、神经网络、朴素贝叶斯作为初始模型分发至各数据源，在历史数据集上进行训练，根据三种模型训练的结果计算准确度的平均值（mean，即准确率）及方差（variance），选择准确率最高且方差最小的作为初始全局模型。第二部分：将t_i时间段的初始全局模型分发至各数据源，并进行训练得t_i时间段的本地模型，再使用stacking集成算法集成本地模型，得t_i时间段最有效的更新的全局模型。第三部分：各数据源使用RSA加密算法产生256 B的密钥对，将其公钥传输至可信第三方，t_i时间段的初始全局模型使用公钥加密并传输至各数据源，各数据源使用私钥解密并进行训练。第四部分：可信的第三方使用ECDSA数字签名算法产生密钥对，将其私钥传输至各数据源，公钥传输至对应的区块中，各数据源使用私钥对t_i时间段的初始全局模型参数、本地模型参数、更新的全局模型参数进行签名并传输至对应的区块，区块使用公钥进行验证并存储相应的数据块中。因随机森林、朴素贝叶斯、神经网络三种模型具有随机性，所以本文将每种初始模型迭代100次，计算其平均值及方差来衡量初始模型的性能，表1反映了初始模型在4个数据源的历史数据上的性能。

表1 三种初始模型在历史数据上迭代100次的准确度均值与方差

Tab. 1 Means and variances of 100 iterations of three initial models on historical data

数据源	初始模型	准确度	
		均值	方差/10^{-6}
$k=1$	随机森林	0.918 785	6.691 04
	朴素贝叶斯	0.821 916	1 526.848
	神经网络	0.909 995	9.258 59
$k=2$	随机森林	0.920 328	8.699 53
	朴素贝叶斯	0.569 247	191.951
	神经网络	0.908 753	13.652 1
$k=3$	随机森林	0.918 170	8.153 68
	朴素贝叶斯	0.606 767	5 914.395
	神经网络	0.910 728	9.311 93
$k=4$	随机森林	0.914 854	10.047 9
	朴素贝叶斯	0.547 813	133.432
	神经网络	0.906 778	11.172 2

从表1可以看出，对于历史数据来说，平均值由大到小的顺序是随机森林、神经网络、朴素贝叶斯，而方差由小到大的顺序是随机森林、神经网络、朴素贝叶斯，综合考虑随机森林的性能最优，故本文选择随机森林作为首轮的初始全局模型，且使用stacking集成算法依次对首轮初始模型在数据源上建立的模型进行集成，得到首轮更新的全局模型的准确度。因随机森林作为初始全局模型具有随机性，且数据为随机划分而成，所以本文将初始模型在随机划分的数据集上迭代10次的平均值作为数据划分一次模型的准确度，将划分100次数据的平均值作为首轮模型的准确度。图7是在历史数据集的4个数据源上使用随机森林在随机划分一次数据上迭代过程的准确率变化情况。

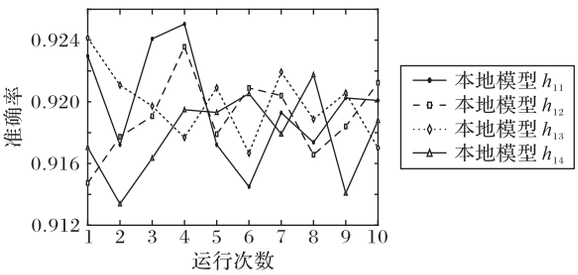

图7 初始全局模型H_0在随机划分一次数据上迭代过程的准确率变化情况

Fig. 7 Changes in accuracy of a iteration of initial global model H_0 on once randomly divided data

随机森林在首轮数据的准确度可表示为随机森林在划分100次的数据且在每次划分的数据上迭代10次训练准确度的平均值，可保证每次数据的准确性，同时每次均产生4个本地模型$h_{in}(n=1,2,3,4)$，为检验本地模型的性能，采用准确度平均值及方差来衡量。

从表2可以看出，初始全局模型H_0在t_1时间段上数据的准确度均值均在91.8%以上，且方差都很小，表明多个本地模型的性能都很好，且本地模型在数据上训练的准确度均值结果几乎相等。

表2 H_0在t_1时间段的增量数据上迭代100次的准确度均值与方差

Tab. 2 Means and variances of 100 iterations of H_0 on incremental data in t_1 period

数据源	准确度	
	均值	方差/10^{-6}
$k=1$	0.918 860 566	4.062 87
$k=2$	0.919 212 101	4.096 09
$k=3$	0.918 765 425	4.846 68
$k=4$	0.918 961 112	3.188 87

将训练的多个本地模型使用stacking集成算法来集成，得到更新后的全局模型h_1在t_1时间段数据上的准确度保持在91.858%，且方差为$4.889\,5\times10^{-6}$，说明更新的全局模型h_1有很好的稳定性。

为检验t_1时间段更新的全局模型h_1在t_2时间段上的训练结果，本文算法将会判断在t_i时间段上有无新增加或减少的数据源，实验中假设在t_2时间段和t_1时间段的数据源一致，t_3时间段将会减少1个数据源，t_4时间段将会增加1个数据源，当t_i时间段新增加或减少了数据源，那t_i时间段的初始全局模型H_i需要重新计算。因t_2时间段和t_1时间段的数据源一致，所以将t_1时间段更新的全局模型h_1作为t_2时间段的初始全局模型H_1，在t_2时间段所产生的数据上进行训练。

表3中的数据为t_2时间段的初始全局模型H_1在每次随机划分的数据上迭代10次训练的平均值，可保证每次数据的准确性，同时每次均产生4个本地模型$h_{in}(n=1,2,3,4)$，为检验t_2时间段本地模型的性能，仍采用平均值及方差来衡量。

从表3中可以看到，在t_2时间段的初始全局模型H_1的准确度均在91.8%以上，且方差都很小，表明初始全局模型H_1在t_2时间段内所产生的数据上训练的结果均很好，模型很稳定。

将t_2时间段训练的多个本地模型使用stacking集成算法来集成,得到更新后的全局模型h_2的准确度保持在91.86%以上,且方差为6.12×10^{-6},说明更新的全局模型h_2有很好的稳定性。

表3　H_1在t_2时间段的增量数据上迭代100次的准确度均值与方差

Tab. 3　Means and variances of 100 iterations of H_1 on incremental data in t_2 period

数据源	准确度	
	均值	方差/10^{-6}
$k=1$	0.919 022 744	4.986 88
$k=2$	0.919 126 724	4.089 55
$k=3$	0.918 519 442	5.775 57
$k=4$	0.918 735 183	4.955 69

为充分检验增加或减少数据源对算法的影响,将t_3时间段减少1个数据源(3个数据源)进行训练,因数据源发生变化,所以需要重新计算t_3时间段的初始全局模型H_2,同样地使用准确度均值及方差衡量3种初始模型性能。表4是三种初始全局模型在t_3时间段的增量数据上重新计算的准确度均值与方差的变化情况。从表4可以看出,随机森林的准确度均值与方差明显优于朴素贝叶斯和神经网络,而神经网络的准确度均值与方差均优于朴素贝叶斯。

表4　三种初始模型在t_3时间段的增量数据上迭代100次的准确度均值与方差

Tab. 4　Means and variances of 100 iterations of three initial models on incremental data in t_3 period

数据源	初始模型	准确度	
		均值	方差/10^{-6}
$k=1$	随机森林	0.918 454	8.22
	朴素贝叶斯	0.750 931	15 485
	神经网络	0.908 217	8.59
$k=2$	随机森林	0.918 820	8.54
	朴素贝叶斯	0.557 233	177
	神经网络	0.908 671	11.8
$k=3$	随机森林	0.919 633	8.50
	朴素贝叶斯	0.565 212	56.5
	神经网络	0.910 828	13.2

表5中的数据为t_3时间段的初始全局模型H_2在每次随机划分的数据上迭代100次的准确度均值,可保证每次数据的准确性,同时每次均产生4个本地模型$h_{in}(n=1,2,3,4)$。从表5可以看出减少数据源可以使初始全局模型H_2的准确度有所增加,且方差较小,满足实验性需求。

表5　H_2在t_3时间段的迭代100次的准确度均值与方差

Tab. 5　Means and variances of 100 iterations of H_2 on incremental data in t_3 period

数据源	准确度	
	均值	方差/10^{-6}
$k=1$	0.919 244 561	5.235 52
$k=2$	0.918 914 000	6
$k=3$	0.919 099 000	4.75

将t_3时间段训练的多个本地模型使用stacking集成算法集成,得到更新后的全局模型h_3的准确度保持在91.931%以上,且方差为5.7×10^{-6},说明更新的全局模型h_3有很好的稳定性。同时可以看到数据源的减少可以使更新的全局模型h_3的准确度有所提升,且方差在变小,说明减少数据源可以使模型更加稳定。

为检验增加数据源对联邦增量学习算法的影响,在t_3时间段的基础上增加2个数据源,表6反映的是初始模型在5个数据源上训练情况。从表6可以看出,增加数据源后初始全局模型H_3的准确度均有所下降,且方差也都有所增加,但随机森林的准确度依旧很高,模型的稳定性也较好。

表6　三种初始模型在t_4时间段的增量数据上迭代100次的准确度均值与方差

Tab. 6　Means and variances of 100 iterations of three initial models on incremental data in t_4 period

数据源	初始模型	准确度	
		均值	方差/10^{-6}
$k=1$	随机森林	0.919 433	9.362 08
	朴素贝叶斯	0.565 373	135.829
	神经网络	0.909 032	12.006 1
$k=2$	随机森林	0.914 845	8.972 61
	朴素贝叶斯	0.564 198	103.903
	神经网络	0.906 597	14.137 3
$k=3$	随机森林	0.919 003	10.460 9
	朴素贝叶斯	0.743 660	14 551.69
	神经网络	0.911 841	17.611 8
$k=4$	随机森林	0.918 572	8.181 98
	朴素贝叶斯	0.699 203	8 584.856
	神经网络	0.909 041	10.453 1
$k=5$	随机森林	0.913 854	15.032 1
	朴素贝叶斯	0.736 417	15 939.276
	神经网络	0.909 694	21.932 6

表7中的数据为t_4时间段的初始全局模型H_3在每次随机划分的数据上迭代100次的准确度均值,可保证每次数据的准确性,同时每次均产生4个本地模型$h_{in}(n=1,2,3,4)$。从表7可以看出,增加数据源使初始全局模型H_2的准确度有所减小,且新增加的数据源所对应的方差比原有数据源的方差要大,但还是能满足实验需求。

将t_4时间段训练的多个本地模型使用stacking集成算法来集成,得到更新后的全局模型h_4的准确度保持在91.588%以上,且方差为9.315×10^{-6},说明更新的全局模型h_4有很好的稳定性。同时可得到在增加数据源后,在t_4时间段的数据上,全局模型h_4的准确度达到91.59%,且方差很小。

表7　H_3在t_4时间段的增量数据上迭代100次的准确度均值与方差

Tab. 7　Means and variances of 100 iterations of H_3 on incremental data in t_4 period

数据源	准确度	
	均值	方差/10^{-6}
$k=1$	0.918 861	4.583 43
$k=2$	0.918 502	4.766 95
$k=3$	0.919 132	4.813 12
$k=4$	0.918 633	5.322 26
$k=5$	0.915 758	11.437 20

传统的多源数据源处理技术将多方数据整合后再进行训练,所训练的模型准确率为92.521%,方差为0.934×10^{-6}(为保证数据的准确度与真实性,该数据为随机森林在整合的数据上迭代100次的准确度均值与方差)。

为保证每次初始全局模型H_i能够安全传输至各数据源,需使用256 B的公钥加密模型进行传输。

图8反映出可信第三方与各数据源使用RSA加密算法在不同的时间段内产生的公钥与私钥的变化情况,可信第三方

使用不同的公钥来加密不同时间段的初始全局模型，进而提升不同时间段内初始全局模型传输的安全性；同时，各数据源使用不同的公钥来加密本地模型并传输至可信第三方，可以提升不同时间段内本地模型传输的安全性。

(a) 公钥变化图

(b) 私钥变化图

图8 RSA加密算法的公钥、私钥变化图

Fig. 8 Public and private key changes of RSA encryption algorithm

为保证模型在传输过程的安全性以及防止模型被篡改与攻击的风险，本文由可信的第三方使用ECDSA数字签名算法产生密钥长度为570的密钥对与RSA加密算法产生的256 B的密钥对，图8反映出RSA加密算法产生的密钥对的变化情况，各数据源将RSA加密算法产生的公钥传输至可信第三方，私钥保留在各数据源中。

各数据源使用图9(b)中的私钥对每个时间段的初始全局模型、本地模型、更新的全局模型使用由第三方提供的私钥进行签名，并传输至对应区块的数据块中。区块链中对应区块的数据块使用图9(a)中的公钥进行验证，验证过程能够验证模型参数是否完整地传输至数据块中，同时也能降低模型在传输过程中被篡改及攻击的风险。

(a) 公钥变化图

(b) 私钥变化图

图9 ECDSA的公钥、私钥变化图

Fig. 9 Public and private key changes of ECDSA

对于模型存储阶段，RayBaaS平台能够快捷、高效地构建基于区块链的服务和应用，硬件设备采用Inter Core i5-4200M CPU 2.50 GHz处理器进行实验，区块链底层基于CentOS 7.6操作系统进行部署，存储的数据包括t_i时间内的初始全局模型参数、本地模型参数、更新的全局模型参数，将t_1时间段内的初始全局模型参数存储至区块1的数据块1中，本地模型参数存储至区块1的数据块2、3、4、5中，更新的全局模型参数存储至区块1的数据块6中；将t_2时间段内的初始全局模型参数存储至区块2的数据块1中，本地模型参数存储至区块2的数据块2、3、4、5中，更新的全局模型参数存储至区块2的数据块6中；t_3时间段内的初始全局模型参数存储至区块3的数据块1中，本地模型参数存储至区块3的数据块2、3、4中，更新的全局模型参数存储至区块3的数据块5中；t_4时间段内的初始全局模型参数存储至区块4的数据块1中，本地模型参数存储至区块4的数据块2、3、4、5、6中，更新的全局模型参数存储至区块4的数据块7中。

3.4 实验小结

在线联邦增量学习算法将随机森林、神经网络、朴素贝叶斯分发至历史数据上进行训练并迭代100次，计算其准确度均值及方差作为衡量初始全局模型的标准，其中随机森林在历史数据上的准确度为91.4%，且方差均小于1.2×10^{-5}，所以将随机森林作为首轮初始全局模型并分发至数据源进行训练，且使用stacking集成算法集成多个本地模型，获得更新的全局模型h_1在t_1时间段内所产生的数据的准确度为91.858%。因t_2时间段并无新增数据源，则将更新的全局模型h_1作为t_2时间段的初始全局模型分发至数据源并进行训练，并使用stacking算法来集成多个本地模型，获得的全局模型h_2的准确率为91.86%，表明更新的全局模型h_2具有很强的泛化性。为研究增加与减少数据源对初始全局模型的影响，在t_3、t_4时间段分别减少1个数据源和增加1个数据源，重新计算其平均值及方差，得到在t_3时间段内所产生的数据上随机森林的准确度最高，且均在91.8%以上，使用stacking集成后的准确度为91.9%；在t_4时间段内所产生的数据上随机森林的准确度最高为91.5%，使用stacking集成后的准确度为91.58%，说明增加与减少数据源会对初始全局模型产生影响。使用传统的数据处理技术将多方数据整合后再进行训练，获得的模型的准确率为92.521%，与之相比，本文算法的准确度仅下降约1%，但数据及模型在训练过程中的安全性得到很大提升。

4 结语

本文提出了一种面向区块链的在线联邦增量学习算法，采用数字签名的方式保证数据在收集阶段的安全性与完整性；使用stacking集成算法来整合多方本地模型，虽然模型的准确率略有下降，但模型的安全性得到提升，同时模型在增量数据上进行训练，使模型具有时效性；将每个时间段的初始全局模型参数、本地模型参数、更新的全局模型上传至对应数据块中并快速同步，使数据的传输成本降低，同时模型参数的安全性得到了保障。接下来的工作中，我们会尝试将本文算法应用到其他隐私保护技术中，在保证模型准确率的基础上，进一步提升数据与模型的安全性。

参考文献（References）

[1] SUN C, SHRIVASTAVA A, SINGH S, et al. Revisiting unreasonable effectiveness of data in deep learning era[EB/OL]. [2019-08-04]. https://arxiv.org/pdf/1707.02968.pdf.

[2] YANG Q, LIU Y, CHEN T, et al. Federated machine learning: concept and applications[J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): No. 12.

[3] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[EB/OL]. (2017-02-28)[2019-12-02]. https://arxiv.org/pdf/1602.05629.pdf.

[4] KIM H, PARK J, BENNIS M, et al. On-device federated learning via blockchain and its latency analysis[EB/OL]. [2019-06-01]. https://arxiv.org/pdf/1808.03949v1.pdf.

[5] LI S, CHENG Y, LIU Y, et al. Abnormal client behavior detection in federated learning[EB/OL]. [2019-12-06]. https://arxiv.org/pdf/1910.09933.pdf.

[6] ZHU L, LIU Z, HAN S. Deep leakage from gradients[EB/OL]. [2019-12-19]. https://arxiv. org/pdf/1906. 08935. pdf.

[7] 章宁,钟珊. 基于区块链的个人隐私保护机制[J]. 计算机应用, 2017, 37(10):2787-2793. (ZHANG N, ZHONG S. Mechanism of personal privacy protection based on blockchain [J]. Journal of Computer Applications, 2017, 37(10):2787-2793.)

[8] 拜亚萌,满君丰,张宏. 基于区块链的电子健康记录安全存储模型[J]. 计算机应用, 2020, 40(4):961-965. (BAI Y M, MAN J F, ZHANG H. Secure storage model of electronic health records based on blockchain[J]. Journal of Computer Applications, 2020, 40(4):961-965.)

[9] 李云飞,柳青,郝林,等. 一种有效的RSA算法改进方案[J]. 计算机应用, 2010, 30(9):2393-2397. (LI Y F, LIU Q, HAO L, et al. Efficient variant of RSA cryptosystem[J]. Journal of Computer Applications, 2010, 30(9):2393-2397.)

[10] 娄悦,施荣华,曹龄兮. 基于强认证技术的会话初始协议安全认证模型[J]. 计算机应用, 2006, 26(10):2332-2335. (LOU Y, SHI R H, CAO L X. SIP secure authentication model based on strong authentication technology [J]. Journal of Computer Applications, 2006, 26(10):2332-2335.)

[11] 汪希仁,孙战辉,祝永霞. 一种新的封闭ECDSA签名阈下信道方案[J]. 计算机与网络, 2013, 39(6):71-73. (WANG X R, SUN Z H, ZHU Y X. A new subliminal-free protocol in ECDSA [J]. Computer and Network, 2013, 39(6):71-73.)

[12] 胡彬轩. 基于联邦学习的空气质量监测系统设计与实现[D]. 北京:北京邮电大学, 2019: 30-40. (HU B X. Design and implementation of air quality monitoring system based on federal learning [D]. Beijing: Beijing University of Posts and Telecommunications, 2019: 30-40.)

[13] GAO H, HUANG W, YANG X. Applying probabilistic model checking to path planning in an intelligent transportation system using mobility trajectories and their statistical data[J]. Intelligent Automation and Soft Computing, 2019, 25(3):547-559.

[14] GAO H, HUANG W, DUAN Y, et al. Research on cost-driven services composition in an uncertain environment[J]. Journal of Internet Technology, 2019, 20(3): 755-769.

[15] PREUVENEERS D, RIMMER V, TSINGENOPOULOS I, et al. Chained anomaly detection models for federated learning: an intrusion detection case study[J]. Applied Sciences, 2018, 8(12): No. 2663.

[16] BRISIMI T S, CHEN R, MELA T, et al. Federated learning of predictive models from federated electronic health records[J]. International Journal of Medical Informatics, 2018, 112: 59-67.

[17] ZHANG W, ZHANG Y, ZHAI J, et al. Multi-source data fusion using deep learning for smart refrigerators[J]. Computers in Industry, 2018, 95: 15-21.

[18] LEE J, SUN J, WANG F, et al. Privacy-preserving patient similarity learning in a federated environment: development and analysis[J]. JMIR Medical Informatics, 2018, 6(2): No. e20.

[19] SHEN G, HAN X, ZHOU J, et al. Research on intelligent analysis and depth fusion of multi-source traffic data[J]. IEEE Access, 2018, 6:59329-59335.

[20] LIU J, LI T, XIE P, et al. Urban big data fusion based on deep learning: an overview[J]. Information Fusion, 2020, 53: 123-133.

[21] ZHU Z, DONG S, YU C, et al. A text hybrid clustering algorithm based on HowNet semantics[J]. Key Engineering Materials, 2011, 474/475/476: 2071-2078.

This work is partially supported by the National Natural Science Foundation of China (61572170, 61170254), the Tangshan Science and Technology Project (18120203A).

LUO Changyin, born in 1994, M. S. candidate. His research interests include data security.

CHEN Xuebin, born in 1970, Ph. D., professor. His research interests include data security, IoT security, network security.

MA Chundi, born in 1999. His research interests include network security.

WANG Junyu, born in 1996, M. S. candidate. Her research interests include network security.