Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

联邦学习

硕士研究生　王殿元

2020年06月07日

内容提要

- 背景简介
- 基本概念
- 算法原理
  - 纵向联邦学习 + SecureBoost
  - 横向联邦学习 + Fedavg
- 应用总结
- 研究展望
- 参考文献

预期收获

- 预期收获
  - 1. 了解横向、纵向联邦学习的基本思想
  - 2. 理解纵向联邦学习框架下SecureBoost的算法原理
  - 3. 理解横向联邦学习框架下Fedavg的算法原理
  - 4.了解联邦学习的应用前景

背景简介

- AI成功秘诀

computing power　　big data

1 ZB = $10^{21} Byte$

Intel i386
Intel i486
Intel Pentium
Intel core
NVIDIA GPU

Google TPU

2008 0.5ZB

2010 1.2ZB

2012 2.0ZB

2015 8.5ZB

2020 40ZB

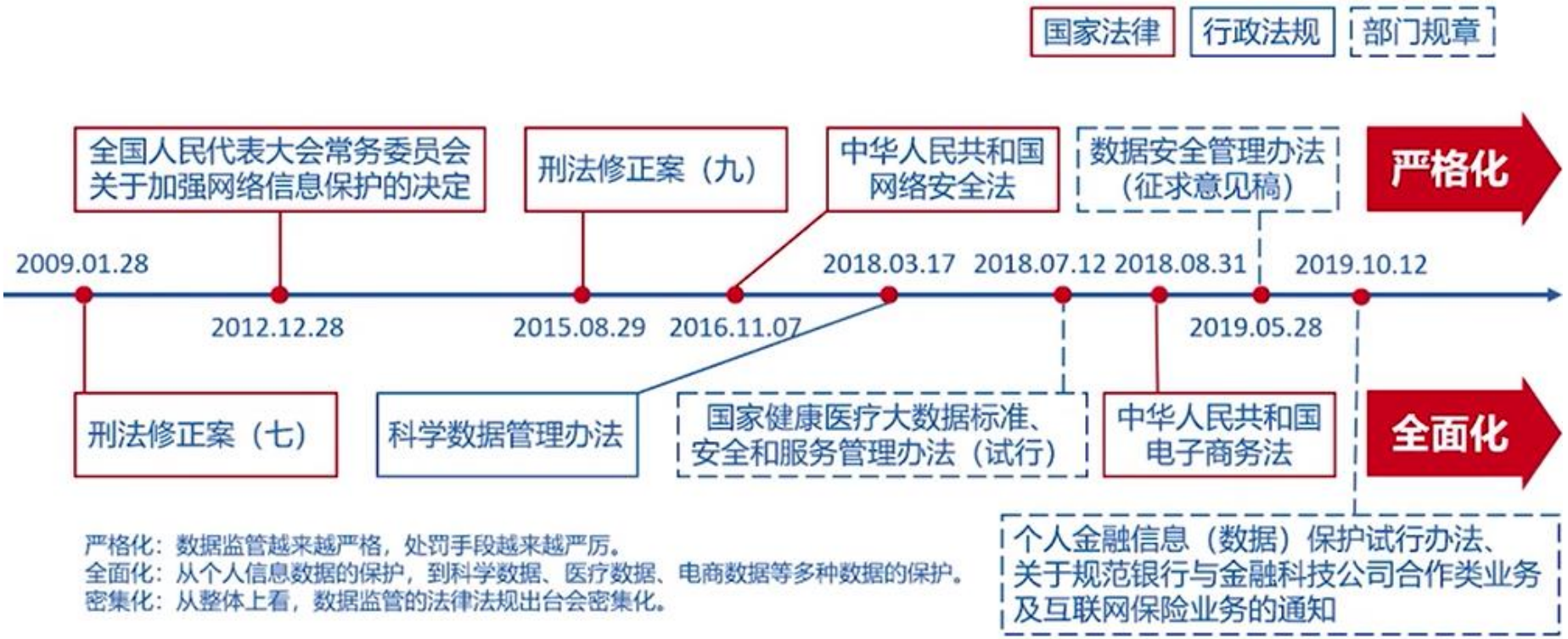
The world’s most valuable resource is no longer oil , but data.

《The Economist》– May 2017

—— 来自互联网数据中心(IDC)

背景简介

- 数据监管
  - 严格化、全面化、密集化
  - 全球化
    - 欧盟 GDPR、美国 CCPA



背景简介

- 数据孤岛
  - 独立存储，独立维护，彼此间相互孤立，形成了物理上的孤岛

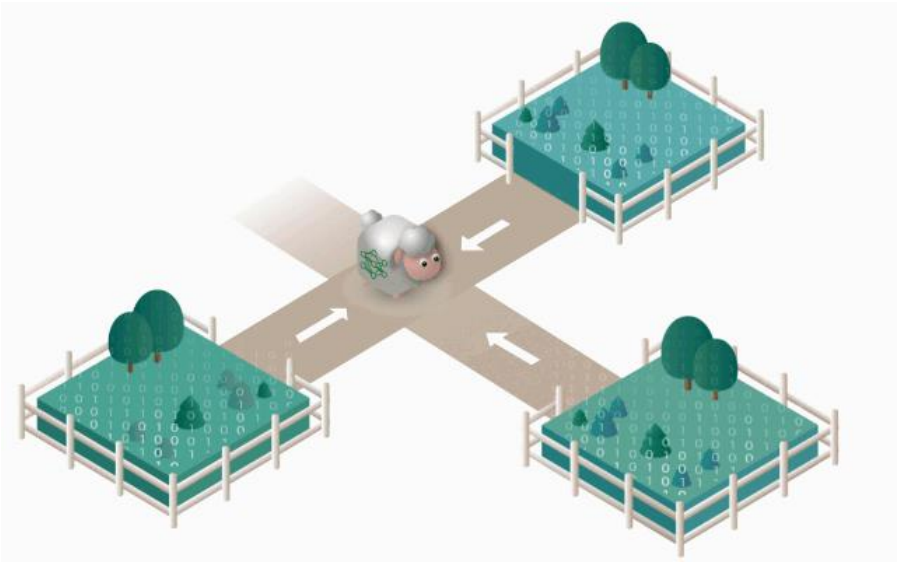
基本概念

- 联邦学习的思想
  - 联邦学习（Federated Learning）在 2016 年由谷歌最先提出，其设计目标是在保障大数据交换时的信息安全、保护终端数据和个人数据隐私、保证合法合规的前提下，在多参与方或多计算结点之间开展高效率的机器学习，解决数据孤岛的问题。

- 联邦学习的特点
  - 数据隔离
  - 无损
  - 对等
  - 共同效益

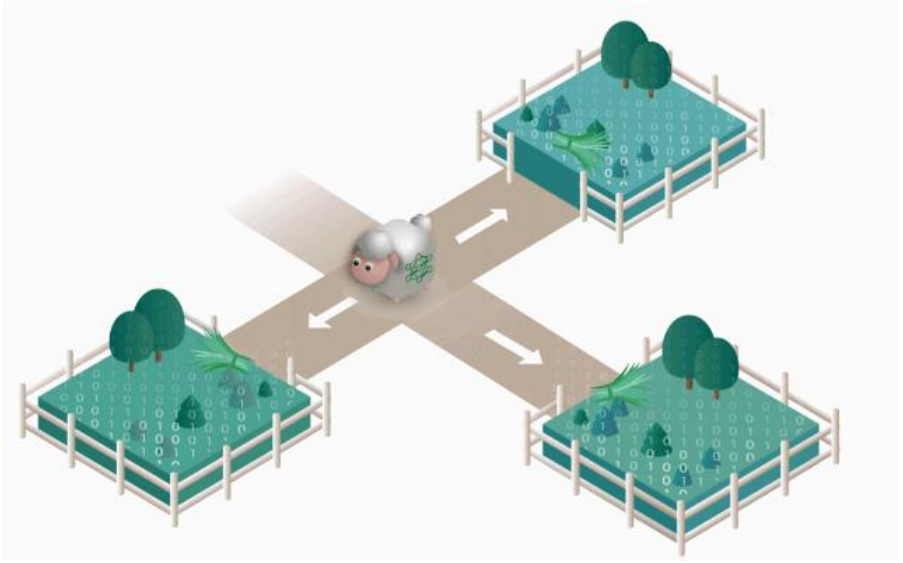
基本概念

- 传统学习训练模型　　VS　　联邦学习训练模型



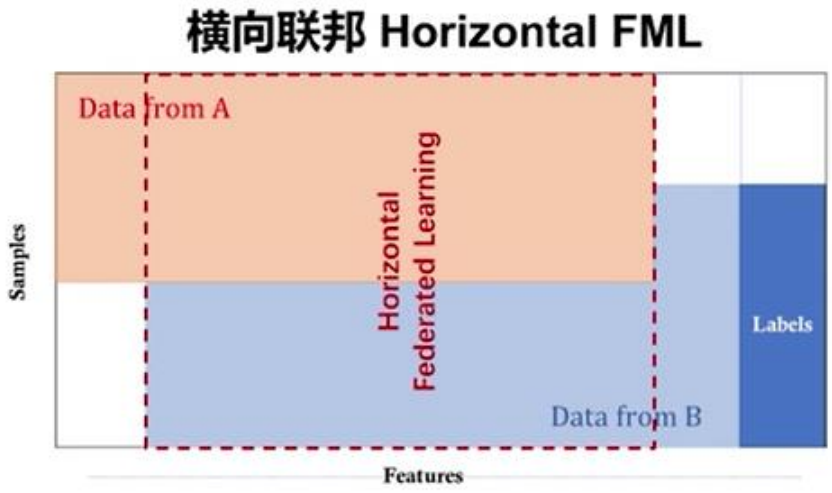
模型不动数据动

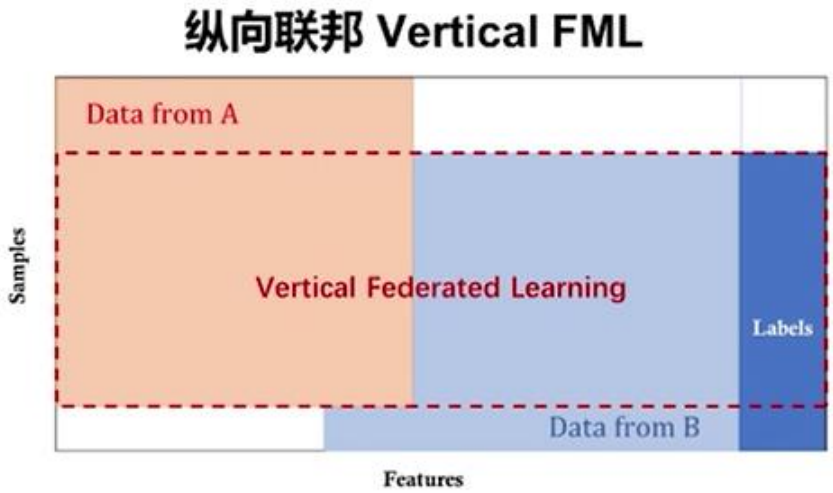
数据不动模型动



基本概念

- 联邦学习的分类
  - 横向联邦
    - 适用于参与者的数据特征维度相同，而样本ID不同。例如两家不同地区的银行的客户数据。
  - 纵向联邦
    - 适用于参与者的数据ID重叠较多，而数据特征维度不同。例如同一地区的银行和电商的共同客户的数据。





基本概念

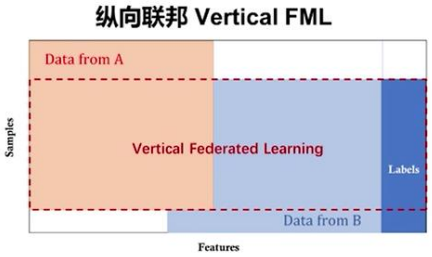
- **联邦学习的关键——加密/解密**
  - **同态加密** Homomorphic Encryption（HE）
    - **概念：同态加密是基于数学难题的计算复杂性理论的密码学技术。对经过同态加密的数据进行处理得到一个输出，将这一输出进行解密，其结果与用同一方法处理未加密的原始数据得到的输出结果是一样的。**
    - **特性：加法同态：** $[\![u]\!] + [\![v]\!] = [\![u + v]\!]$；**数乘同态：** $n \cdot [\![u]\!] = [\![n \cdot u]\!]$
  - RSA**加密**
    - **非对称加密算法，加密过程中生成公钥** n，**加密算法**e，**解密算法**d。**在加密传输的过程中传输公钥** {n , e} 。
  - **哈希机制**
    - **通过散列算法将任意长度输入变换成固定长度的输出，是一种压缩映射。**
  - **差分隐私** Differential Privacy（DP）
    - **保护数据的一种密码学机制**
    - **《差分隐私原理及应用》——邹森** 2020.05.17

纵向联邦学习

纵向联邦学习-实例

适用：参与者的数据ID重叠较多，而数据特征维度不同。

举例：某银行与其合作企业联合训练一个信用逾期评估模型，银行有Y（逾期表现），期望优化本方的Y预测模型



数据方：样本ID相同

◆ 设定：

✓ 只有银行拥有 Y=“逾期表现”

✓ 合作企业无法暴露含有隐私的 X

◆ 传统建模方法问题：

✓ 合作企业缺乏 Y 无法独立建模

✓ X 数据全量传输到银行不可行

◆ 期望结果：

✓ 保护隐私条件下，建立联合模型

✓ 联合模型效果超过单边数据模型

合作企业				银行			
ID 电话号	X1 账龄	X2 月薪	X3 等级	ID 电话号	X4 征信	X5 评分	Y 表现
U1	7	7500	B	U1	600	90	无
U2	6	9800	A	U2	700	95	无
U3	1	3500	C	U3	400	75	有
U4	8	7500	B	U4	580	85	有
U5	9	9999	A	U8	739	98	无
U6	5	6900	B	U9	520	80	有
U7	5	8800	A	U10	699	94	无

纵向联邦学习-算法原理

算法1：基于隐私保护的样本id匹配算法

T	寻找多个参与方之间的数据交集
I	$A \ni \{u_1, u_2, u_3, u_4\}$； $B \ni \{u_1, u_2, u_3, u_5\}$
P	RSA + 哈希机制的安全求交方案
O	交集$C \ni \{u_1, u_2, u_3\}$

算法2：联合训练算法 —— SecureBoost

T	协同学习一种共享的梯度树递增模式，既无损又安全
I	多个参与方各自保存管理的数据
P	纵向联邦学习 + SecureBoost
O	全局最优模型（效果优于单边数据模型）

Cheng, K. , Fan, T. , Jin, Y. , Liu, Y. , Chen, T. , & Yang, Q. (2019). Secureboost: a lossless federated learning framework. *arXiv preprint arXiv:1901.08755*.

基于隐私保护的样本id匹配算法流程

Part A

$X_A: \{u_1, u_2, u_3, u_4\}$

Part B

$X_B: \{u_1, u_2, u_3, u_5\}$

Step1
生成一对密钥

发送公钥

Step2
使用公钥加密一组随机数
使用哈希压缩原始数据

发送加密后的数据

Step3
使用私钥加密
使用哈希压缩原始
数据再用私钥加密

发送加密后的数据

Step4
消去随机数
得到两个集合的交集信息
得到id

发送交集信息

Step5
得到id

基于隐私保护的样本id匹配算法原理

Part A

X_A: $\{u_1, u_2, u_3, u_4\}$

Part B

X_B: $\{u_1, u_2, u_3, u_5\}$

$RSA: n, e, d$

$public\ key: (n, e)$

Y_A=$\{r_i{}^e H(u_i) | u_i \in X_A\ r_i$:rand$\}$

Y_A=$\{r_1{}^e H(u_1), r_2{}^e H(u_2), r_3{}^e H(u_3), r_4{}^e H(u_4)\}$

$Z_A = (r_i{}^e H(u_i))^{\,d} = r_i * (H(u_i))^d \ \%\ \mathrm{n}$
$|r_i{}^e H(u_i) \in Y_A$

$Z_B = \{\mathrm{H}((H(u_j)^d))\%\ \mathrm{n}\ |u_j \in X_B\}$

Z_A, Z_B

$D_A = \{H(r_i * (H(u_i))^{\,d}/r_i) = H((H(u_i))^d)$
$|r_i * (H(u_i))^{\,d} \in Z_A$

$I = D_A \cap Z_B$
$= \{H((H(u_1))^d), H((H(u_2))^d), H((H(u_3))^d)\}$

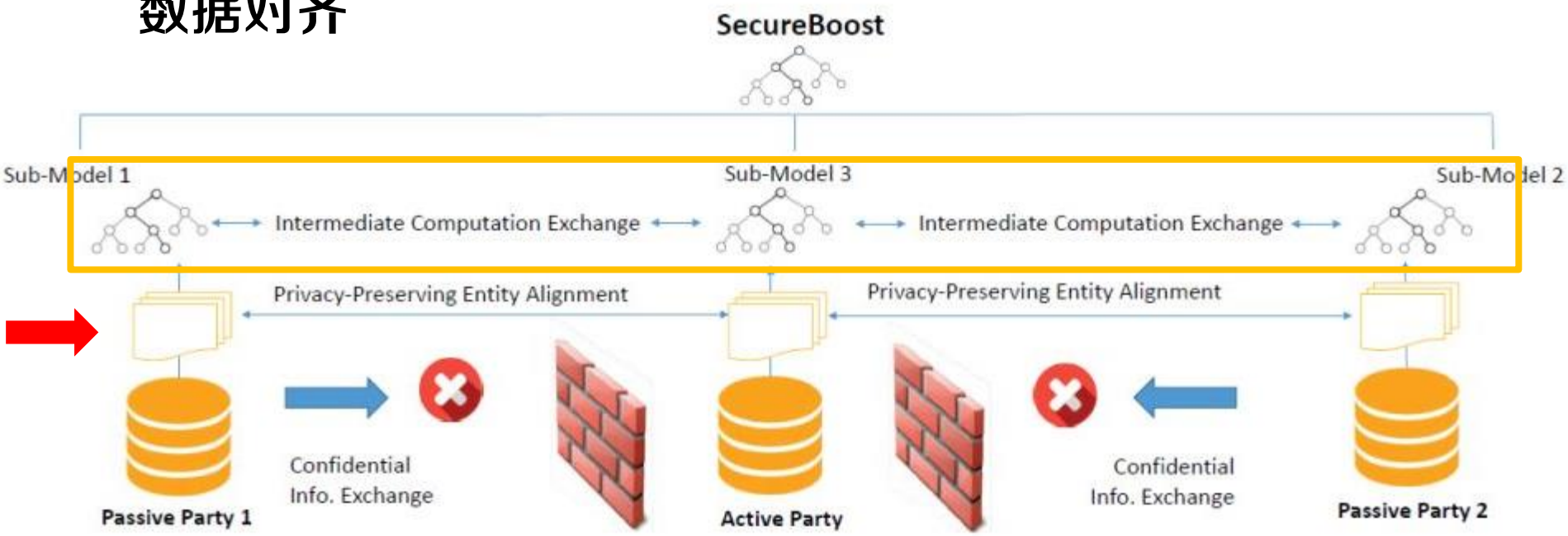
I

$I, D_A => \{u_1, u_2, u_3\}$

$I, Z_B => \{u_1, u_2, u_3\}$

SecureBoost算法原理

Step1
数据对齐

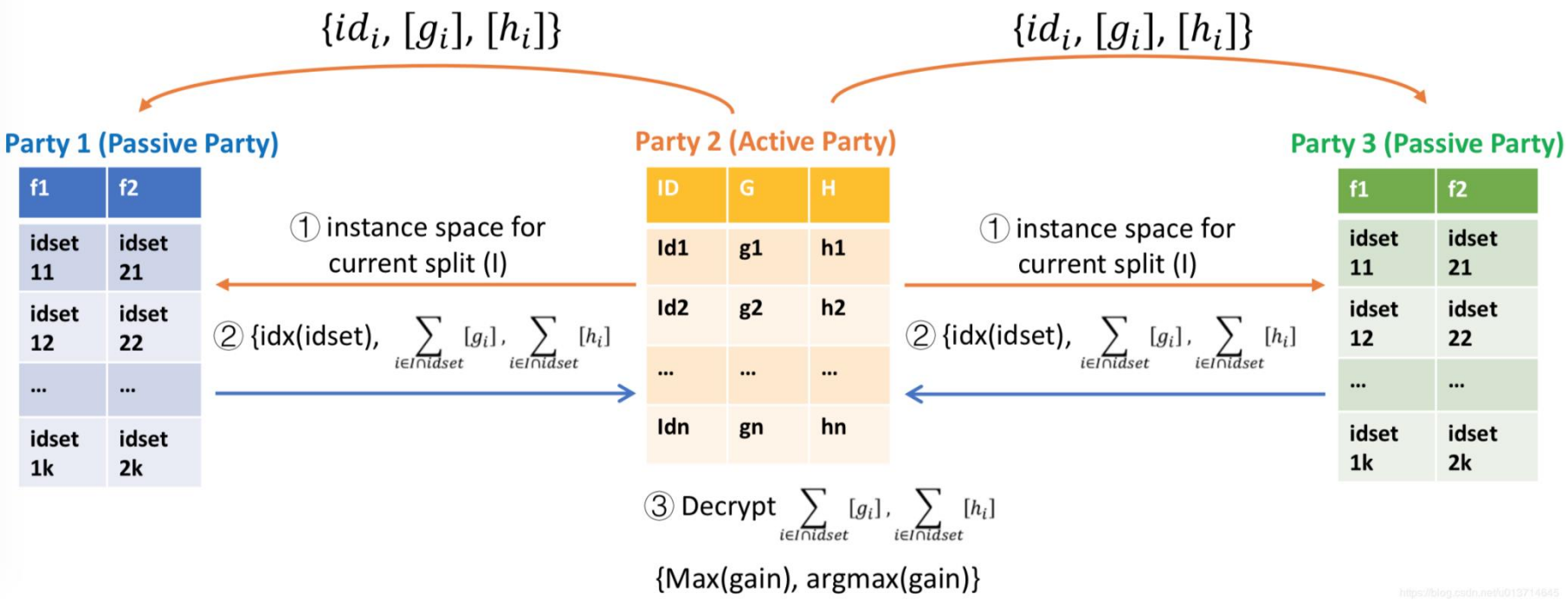


Step2
模型联合训练

SecureBoost算法原理

Step2.1
联邦计算信息增益

$[g_i]$: hormomorphic encrypted g_i
$[h_i]$: hormomorphic encrypted h_i



SecureBoost算法原理

Step2.2
联邦构建树结构

Party 1 (Passive Party)

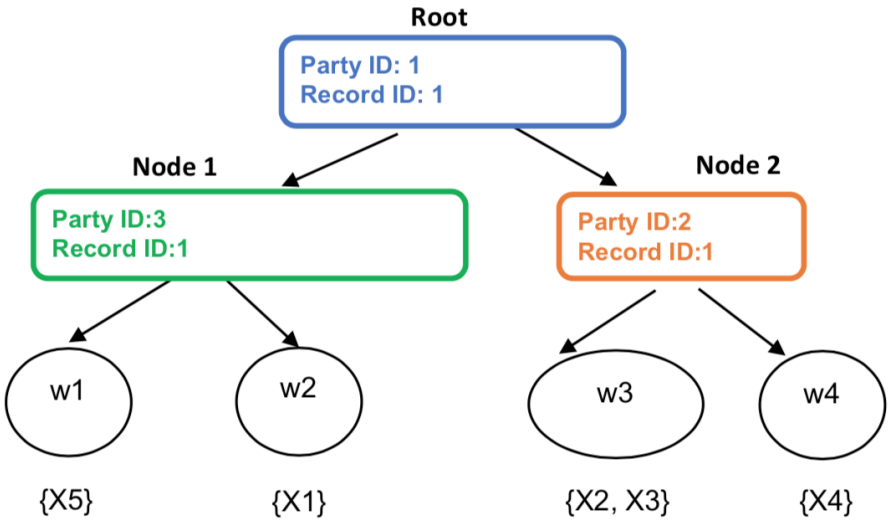
Example	Bill Payment	Education
X1	3102	2
X2	17250	3
X3	14027	2
X4	6787	1
X5	280	1

Party 2 (Active Party)

Example	Age	Gender	Marriage	Label
X1	20	1	0	0
X2	30	1	1	1
X3	35	0	1	1
X4	48	0	1	2
X5	10	1	0	3

Party 3 (Passive Party)

Example	Amount of given credit
X1	5000
X2	300000
X3	250000
X4	300000
X5	200



Lookup table

Party 1:

Record ID	Feature	threshold value
1	Bill Payment	5000

Party 2:

Record ID	Feature	threshold value
1	Age	40

Party 3:

Record ID	Feature	threshold value
1	Amount of given credit	800

SecureBoost算法原理

Step2.3
联邦预测

Training Set

Party 1 (Passive Party)

Example	Bill Payment	Education
X1	3102	2
X2	17250	3
X3	14027	2
X4	6787	1
X5	280	1

Party 2 (Active Party)

Example	Age	Gender	Marriage	Label
X1	20	1	0	0
X2	30	1	1	1
X3	35	0	1	1
X4	48	0	1	2
X5	10	1	0	3

Party 3 (Passive Party)

Example	Amount of given credit
X1	5000
X2	300000
X3	250000
X4	300000
X5	200

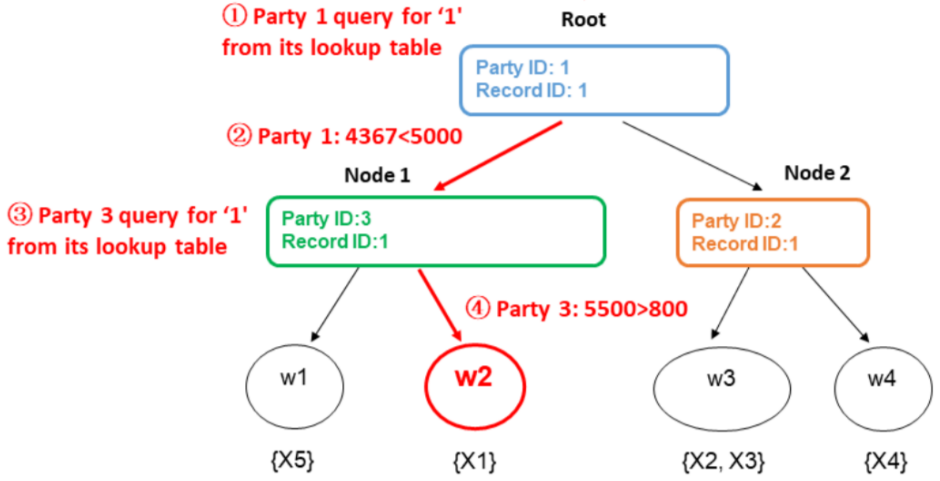
Predict

Example	Bill Payment	Education
X6	4367	2

Example	Age	Gender	Marriage	Label
X6	28	1	0	0

Example	Amount of given credit
X6	5500

input



Lookup table

Party 1:

Record ID	Feature	threshold value
1	Bill Payment	5000

Party 2:

Record ID	Feature	threshold value
1	Age	40

Party 3:

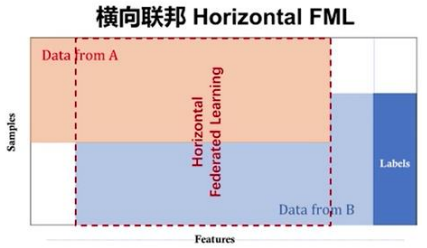
Record ID	Feature	threshold value
1	Amount of given credit	800

横向联邦学习

横向联邦学习-实例

适用：参与者的数据特征维度相同，用户id不同。

举例：某银行与其合作企业联合训练一个反洗钱模型。



数据方：特征维度相同

◆ 设定：

✓ 双方均拥有 Y=“是否存在洗钱行为”

✓ 双方均有相同的 X

✓ 双方不能暴露自己的 {X,Y}

◆ 传统建模方法问题：

✓ 双方的各自样本数量过少

◆ 期望结果

✓ 保护隐私条件下，建立联合模型

✓ 联合模型效果超过单边数据模型

合作企业

ID 电话号	X1 不明资金笔数	X2 大额交易笔数	Y 表现数据
U1	7	15	有
U2	6	20	有
U3	1	5	无
U4	1	0	无
U5	2	1	无
U6	50	50	有
U7	60	6	有

银行

ID 电话号	X1 不明资金笔数	X2 大额交易笔数	Y 表现数据
U8	100	100	有
U9	0	1	无
U10	0	0	无
U11	3	50	有
U12	5	13	无
U13	60	5	有
U14	20	25	有

横向联邦学习-算法原理

算法3：联合训练方法 —— Fedavg

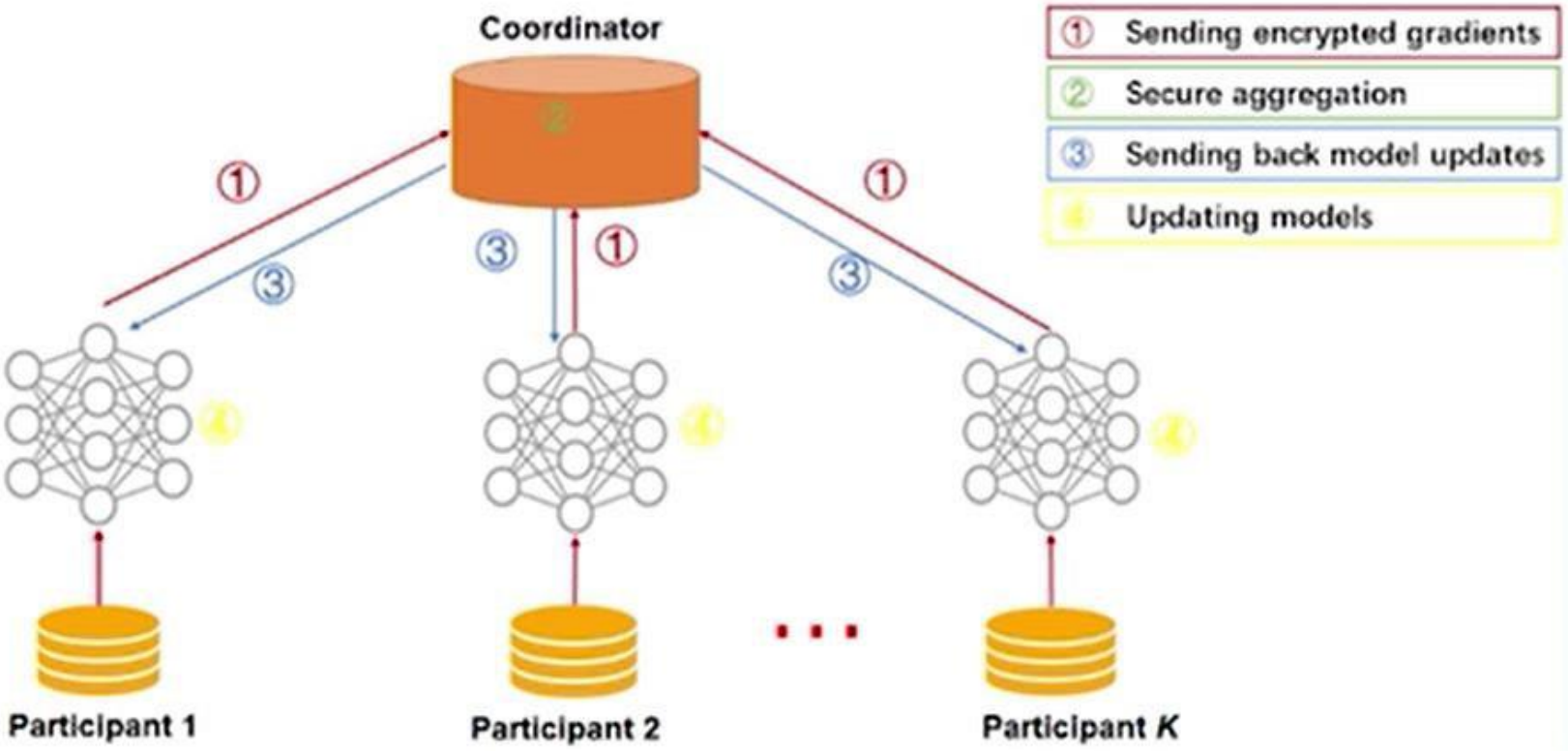
T	在保证数据安全的前提下联合训练多个参与方的数据
I	多个参与方各自保存管理的数据
P	横向联邦学习 + Fedavg
O	全局最优模型（效果优于单边数据模型）

P	多个参与方所包含的样本过少
C	多个参与方所包含特征相同且均有标签
D	联合训练时参数更新方法
L	ACM Transaction 2019

Yang Q , Liu Y , Chen T , et al. Federated Machine Learning: Concept and Applications[J]. Acm Transactions on Intelligent Systems, 2019, 10(2):12.1-12.19.

横向联邦学习-算法原理

- 参与者之间不需要交换信息，通过Fedavg联合更新参数



应用总结

- 联邦推荐（电影 + 图书）
- 信贷管理（行政、税务、舆情、司法…）
- 联邦成像（医疗图像数据）
- 智慧城市

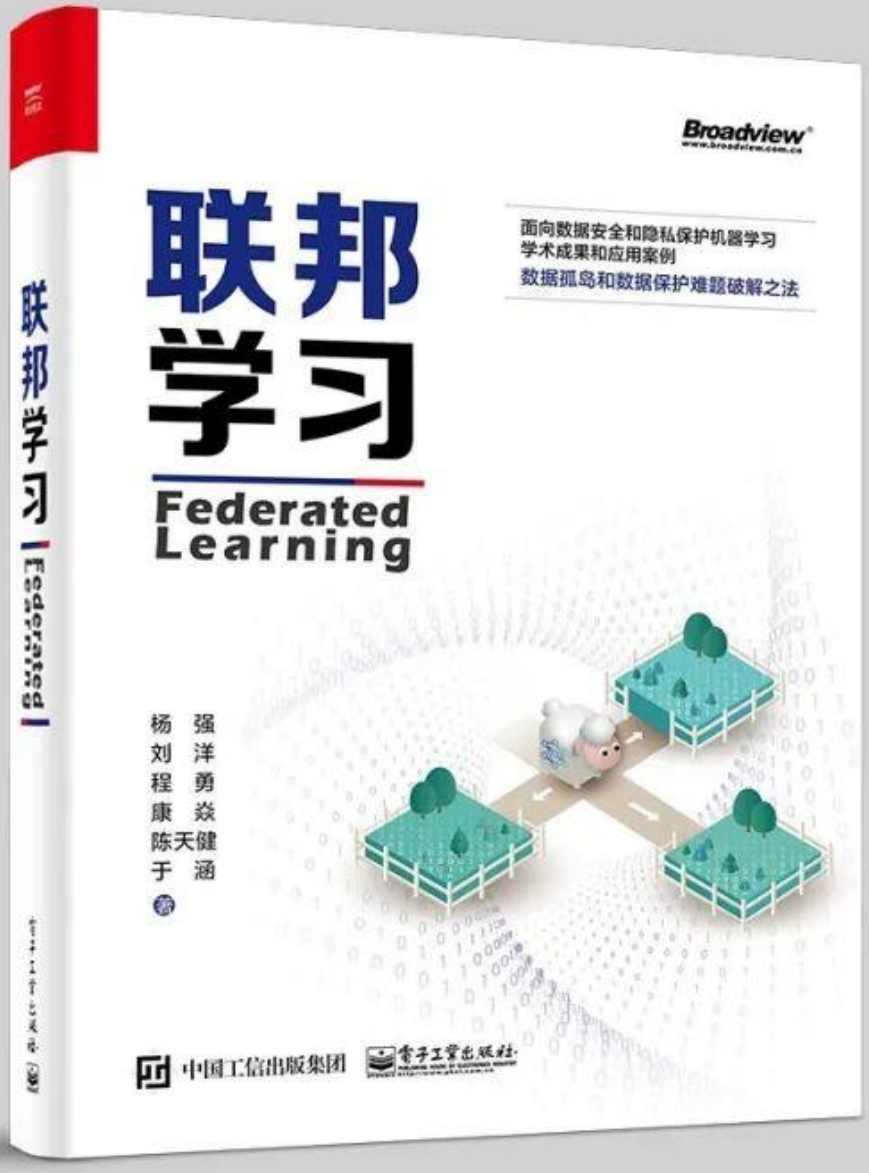
研究展望

参考文献

- Cheng, K. , Fan, T. , Jin, Y. , Liu, Y. , Chen, T. , & Yang, Q. (2019). Secureboost: a lossless federated learning framework. *arXiv preprint* arXiv:1901.08755.https://www.fedai.org/research/publications/secureboost-a-lossless-federated-learning-framework/
- Yang Q , Liu Y , Chen T , et al. Federated Machine Learning: Concept and Applications[J]. Acm Transactions on Intelligent Systems, 2019, 10(2):12.1-12.19. https://www.fedai.org/research/publications/federated-machine-learning-concept-and-applications/
- FATE**开源代码** https://github.com/FederatedAI/FATE

数据不动模型动，风险不增效益增

推荐


Broadview
联邦学习
Federated Learning
面向数据安全和隐私保护机器学习
学术成果和应用案例
数据孤岛和数据保护难题破解之法
杨 强
刘 洋
程 勇
康 焱
陈天健
于 涵
著
中国工信出版集团
电子工业出版社


第一章 联邦学习总览
1.联邦学习提出的背景
2.联邦学习定义
3.联邦学习分类
4.联邦学习平台
5.联邦学习标准
6.FedAI生态
分类
第四章 横向联邦学习
1.定义与应用场景
2.系统架构
3.联邦平均算法
4.最新研究进展
5.挑战与展望
第五章 纵向联邦学习
1.定义与应用场景
2.系统架构
3.联邦线性回归
4.联邦XGBoost
5.挑战与展望
第六章 联邦迁移学习
1.定义与应用场景
2.系统架构
3.模型训练过程
4.模型推理过程
5.挑战与展望
第九章 联邦强化学习
1.定义与应用场景
2.系统架构
3.横向联邦强化学习
4.纵向联邦强化学习
5.挑战与展望
联邦学习—面向隐私保护的机器学习
第二章 联邦学习中的隐私保护技术
1.面向隐私保护的机器学习
2.安全多方计算
3.同态加密
4.差分隐私
5.在机器学习中的应用
第三章 分布式机器学习(DML)
1.面向扩展性的DML
1.1 常见DML平台
1.2 常见DML模式
2.面向隐私保护的DML
2.1 隐私保护机器学习
2.2 隐私保护决策树
2.3 隐私保护SGD
3.挑战与展望
第七章 联邦学习的激励机制
1.激励机制定义与分类
2.联邦激励机制
3.贡献评估
4.基于贡献的收益分配
5.挑战与展望
支撑技术
第八章 联邦学习与其他深度学习领域
1.联邦学习与CV
2.联邦学习与NLP
3.联邦学习与RS
4.挑战与展望
第十章 联邦学习潜在应用场景
1.智慧金融
2.智慧医疗
3.智慧城市
4.区块链与分布式商业
5.5G移动通信
6.挑战与展望
应用

谢谢！

大成若缺，其用不弊。大盈若冲，其用不穷。大直若屈。大巧若拙。大辩若讷。静胜躁，寒胜热。清静为天下正。