同态加密在联邦学习中的运用研究性学习研究报告

一、相关知识

1. 同态加密算法

同态加密是一种特殊的加密方法，它可以直接对数据执行加法或乘法的计算操作，并且在计算过程中不会泄漏原文信息，计算的结果仍然是加密的。同态加密可以分为全同态加密、部分同态加密和半同态加密。

2. 联邦学习

通过联邦学习训练得到的模型和数据集集中在中心服务器上进行训练得到的模型相比效果相同，但是联邦学习存在隐私泄露的风险。在联邦学习的过程中有大量中间参数需要交换，有学者指出可以根据中间参数来推断已获取的信息是否来自某个特定的参与者。在联邦学习系统中，如果参与者是恶意的，那么在训练过程中，参与训练的模型数据可能会被污染。即使在训练之初可以避免恶意参与者的存在，参与者也可能会导致数据信息的泄漏。[1]

所以联邦学习中需要使用同态加密算法来加强安全性：为防止进行多次梯度交换导致数据泄露，可以将同态加密算法与横向联邦学习模型相结合；为防止交换多次计算结果导致数据泄露，可以将同态加密算法与纵向联邦学习模型相结合。实现数据的可用不可见，保障本地用户数据的安全。

二、联邦学习中 RSA 算法的应用

1. RSA 算法

RSA 加密算法（Rivest-Shamir-Adleman）是 Ron Rivest、Adi Shamir 和 Leonard Adleman 三人于 1977 年在文献中首次提出。[3]由于 RSA 是一种非对称加密算法，因此在公开密钥加

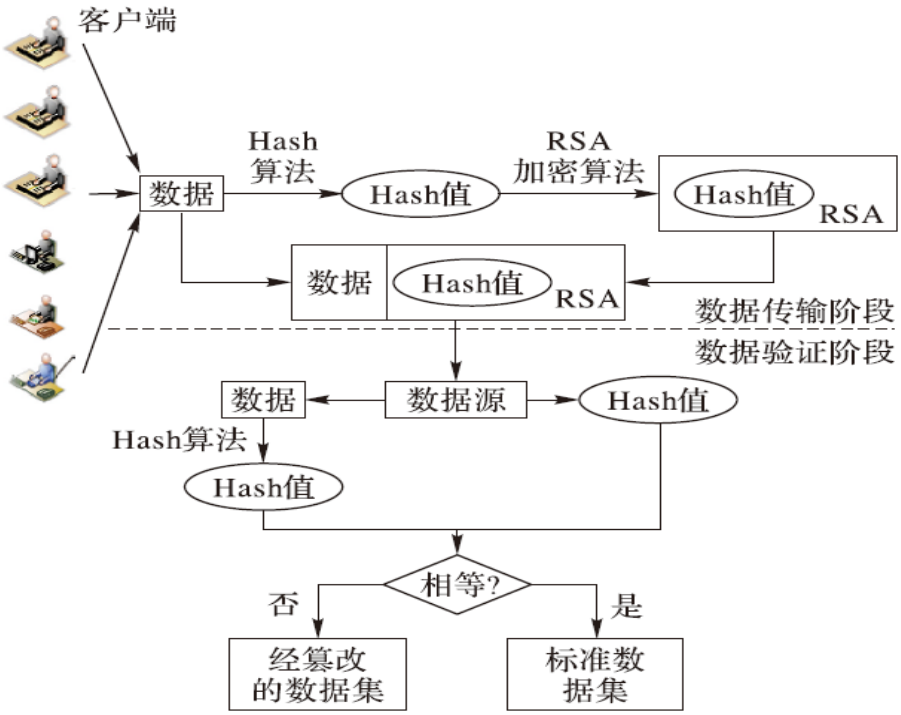
密和电子商业中被广泛应用。例如，提出的基于强认证技术的会话初始协议安全认证模型使用 RSA 数字签名来保证消息传输的机密性、真实性、完整性和不可否认性。[4]

2. 安全性与完整性

联邦增量学习算法利用联邦学习框架和区块链的特性。首先，使用 RSA 加密算法对每个客户端的数据进行哈希计算，并将哈希值与数据一起传输到各数据源。数据源重新计算哈希值，确保数据在收集阶段的安全和完整性。

模型传输的安全性保证：

使用 RSA 加密算法生成的公钥 Pij 对初始全局模型 Hi 进行加密，并将加密后的模型传输到各数据源。数据源利用相应的私钥 pij 进行解密，并用解密后的模型进行训练，以确保在模型传输过程中的安全性。



模型储存的安全性保证：

可信第三方利用 ECDSA 生成密钥对 Pi+2 和 pi+2j。使用私钥 pi+2j 对 ti 时间段的初始全局模型 Hi、本地模型 hi1、hi2 等所有相关模型以及更新的全局模型 hi 进行签名。签名后的模型数据传输到区块 i，区块 i 利用公钥 Pi+2 进行验证，并逐个存储在区块 i 的数据块中。[5]

3. 复杂度与计算开销

在联邦学习中采用 RSA 加密会带来一些复杂度和计算开销的问题。[6]

具体而言，联邦增量学习算法的时间复杂度可表示为：

O(((n * log(n) * d * k) + N^3 + W * k + 2 + G * k + 2) * I)

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，I 表示轮数。

因此，虽然 RSA 加密在保护隐私和安全方面具有优势，但在联邦学习中可能会增加显著的计算和通信开销。为了解决这些问题，研究人员正在不断探索更高效的加密技术和隐私保护方案，以在保障数据隐私的同时降低计算和通信成本。例如，基于同态加密（Homomorphic Encryption）或差分隐私（Differential Privacy）等技术，可以在一定程度上缓解这些问题。

三、联邦学习中 Paillier 算法的应用

1. Paillier 算法

paillier 加密算法是一种最著名的半同态加密方案，具有效率高、安全性证明完善的特点。包括密钥生成、加密过程、解密过程三个阶段。

2. 密钥生成阶段

1）随机选择两个大素数 (p) 和 (q)，确保它们的长度相等，并满足条件 $\gcd(pq,(p-1)(q-1))=1$，这确保了模数 $N=pq$和欧拉函数 $lambda=\mathrm{lcm}(p-1,q-1)$。

计算 $N=pq$ 和$lambda=\mathrm{lcm}(p-1,q-1)$。

3）随机选择 $g \in \mathbb{Z}_N^*$，确保满足条件：$\gcd\left(L\left(g^{\lambda} mod N^2\right), N\right) = 1$，其中 $L(x) = \frac{x-1}{N}$

4）计算私钥参数 μ，使其满足$\mu \equiv \text{modinv}\left(L\left(g^{\lambda} mod N^2\right), N\right)$。

最终，公钥为 (N, g)，私钥为 (λ, μ)。

3. 加密过程

1） 选择一个整数 m，其中 $m \in \mathbb{Z}_N$.

2） 选择一个随机整数 r，其中 $r \in \mathbb{Z}_N^*$.

3）计算密文 c，满足：$c \equiv g^m \cdot r^N \pmod{N^2}$

4. 解密过程

1） 计算$L(x) \equiv \frac{x-1}{N} \pmod{N}$

2）计算 $u \equiv L(c^{\lambda} \pmod{N^2}) \pmod{N}$.

3）计算明文 $m \equiv u \cdot \mu \pmod{N}$.[7]

5. 算法改进

在算法加密的过程中，输入为 n, p, q, g, m，输出为 c。为了提高算法的加密速度，可以利用中国剩余定理的空间特性对数据进行预处理，将在 Z_n 下的模指数运算转化到 Z_p和 Z_q 上，从而提升 Paillier 算法的效率。

具体步骤如下：

1）计算模逆： 计算 p^2和q^2的模逆 μ_1 和 μ_2，其中$\omega_1 = (p^2)^{-1} \text{mod} p^2$ $\omega_2 = (q^2)^{-1} \text{mod} q$ 然后令$\omega \equiv \omega_1 \cdot \omega_2$，$l_1 = \omega \cdot p^2$，$l_2 = \omega \cdot q^2$

2）随机选择 R： 随机选取一个整数 R，使得 $R \in \mathbb{Z}_n^*$.

3）计算模指数运算： 计算$k_1 = g^m \mathrm{mod} p^2$,$k_2 = g^m \mathrm{mod} q^2$,$k_3 = R^n \mathrm{mod} p^2$,$k_4 = R^n \mathrm{mod} q^2$ ，

4）中国剩余定理： 利用中国剩余定理公式计算 $g^m \mathrm{mod} n^2$ 和 $R^n \mathrm{mod} n^2$分别等于 c_1和 c_2，其中$c_2 = R^n \mathrm{mod} n^2$ ，$c_1 = g^m \mathrm{mod} n^2$

5）合并结果： 根据模运算的规则 $(a \times b) \mathrm{mod} n \equiv (a \mathrm{mod} n) \times (b \mathrm{mod} n)$ 得到$c = g^m \cdot R^n \mathrm{mod} n^2$ 。[8]

四、联邦学习中 CKKS 算法的应用

1. CKKS 算法简介

CKKS 是一种高效的全同态加密方案，其支持对浮点数近似计算，通过 SIMD (Single Instruction Multiple Data)技术可以将多个明文打包为单个密文，同时对多个比特处理，提高加密与解密的效率。[9]

2. CKKS 加密结构

使用时需要的数学符号[10]

R=Z[X]/ΦM(X)表示分圆多项式环，ΦM(X)表示 M 阶的分圆多项式。本文中

ΦM(X)=XM/2+1

e←DG(σ2)示从离散高斯分布中采样 N 个系数从而获得多项式 e∈Rq

s←HWT(h)示从{1，-1， 0}均均采样 N 个系数并保证非零系数的个数恰为 h，这些系数组成了多项式 s∈Rq

v←ZO(ρ)示从{1，-1， 0}均匀采样 N 个系数并保证 1 出现的概率为ρ/2， -1 出现的概为ρ/2，0 出现的概率为 1-ρ

密钥生成 KeyGen(1λ)算法如下[11]

1）令 qι=pι forι= 1，…，L

2）采样多项式 s←HWT(h)，a←U(RqL)，e← DG(σ2)

3）输出私钥 sk←(1，s)； 输出公钥 pk←(b，a)∈RqL2，这里 b=-as+e mod qL

加密 Encpk (m)算法如下[12]

采样多项式 v←ZO(0.5)，e0，e1←DG(σ2)

输出密文 ct←v·pk + (m+ e0，e1)(mod qL)

解密 Decsk(ct)算法如下:

令密文 ct= (c0，c1)∈RqL2，输出明文多项式 m←c0+c1·s(mod qL)

CKKS 是目前唯一一个支持浮点数运算的全同态加密方案，允许对实数或复数的浮点数进行加法和乘法的同态运算。此外，相比于其他同态加密方案，CKKS 具有高效的加解密、较小的密文大小等优势，非常适合用于隐私保护的机器学习。

3. CKKS 算法的应用

若双方的向量内积计算有两个，参与方 A，B 各有 一个向量 VA ={ a0，a1，…，an }，VB ={ b0，b1，…，bn }，要求 计算两个向量的内积，且不泄露各自的隐私 A 使 用 自 己 的 公 钥 （CKKS 算 法） 加 密 数 据 EncA (VA)，将密文发送给 B；B 用 A 的公钥 （CKKS 算 法） 加密数据 EncA (VB)； B 计算 C0=EncA (VA) EncA (VB)= EncA (a0·b0)，…，EncA (an - 1·bn - 1)； B 对 C0 移位 n/2 得到 C0'， 求和得 C1 =C0 +C0'； B 对 C1 移位 n/4 得到 C1'，求和得 C2； B 得到最终结果，将其发送给 A；A 解密得到结果。CKKS 可以使用公式直接对加密 参数 θ 和数据(x，y)进行运算。[13]

五、联邦学习中的 Gentry 方案

1. 背景

在前文中，我们已经提到了 RSA 算法与 Paillier 算法，它们对同态加密的发展都起到了至关重要的作用。但 RSA 算法仅对乘法同态，Paillier 算法仅对加法同态，它们都仅是部分同态。Boneh-Goh-Nissim 方案确实既满足加法同态又满足乘法同态，但也仅能进行一次乘法运算。因此它也称为近似全同态。[14]

2. 密文生成

定义：b 为要加密的一个位。其取值为 0 或 1。p 为加密 b 用的秘钥，必须是奇数。x 为掩盖 b 的噪音，是随机数。k 也是随机数。将 b 加密为 c=b+2x+kp。

3. 解密

要解 b 的值时，只需让 c mod p，即（b+2x）mod p。由于 2x 均为偶数，不影响其奇偶性，因此可以得到 b 的值。

4. 局限性

随着计算的进行，噪音 x 会不断增长，尤其是在进行乘法时急剧增长。待噪音增长到一定程度，便会超过计算机限制，使 c 变成负数并出现奇偶翻转。这也就意味着计算结果的正确性无法保证。

5. 改进——Bootstrapping 办法

1） 定义：p_1 与 p_2 分别为两个秘钥，b 为要加密的位
2） 用 p_1 将 b 加密成 c_1
3） 刚刚被转化过来的 c_1 噪音含量很低，可以进行数次计算
4） 多次计算后得到 c_2，如果此时 c_2 的噪音含量已经达到一定程度，则跳转到 5，否则跳转到 3
5） 用秘钥 p_2 将秘钥 p_1 加密成 p_{11}、将 c_2 加密成 d_0
6） 用 p_{11} 将 d_0 解密成 d_1，此时 d_1 可视为 b 只经过 p_2 加密形成的密文，在 p_1 加密中的噪音已经由于解密而消除
7） 继续计算，噪音达到一定程度后重复上述步骤

6. 缺陷

由于解密需要占用不小的空间，所以每轮计算的最后需要保留解密的空间，导致计算次数减少，解密次数增多。而解密的步骤往往也较为复杂，这就导致此方法计算量很大，效率低下。

六、联邦学习中的 Boneh-Goh-Nissim 方案

1. Boneh-Goh-Nissim 方案简介

Boneh-Goh-Nissim 方案是利用双线性映射实现的支持任意数量加法和一次乘法的有限次全同态加密算法，它通过保持密文大小不变来支持任意数量的加法和一次乘法。该方案的难度基于子群决策问题——子群决策问题简单地确定一个元素是否是复合阶 n = pq 的群 G 的子群 G 的成员，其中 p 和 g 是不同的素数。BGN 通常基于椭圆曲线密码学，使用椭圆曲线上的点作为公钥，而私钥则是与之相关的离散对数

2. Boneh-Goh-Nissim 方案原理[15]

1) Enc（加密）算法

要加密消息 m，使用预先计算的 g 和 h 从集合[0，1，... ，n - 1]中选取并加密随机数 r，如下所示:

$$c=E(m)=g^{m}h^{r}\bmod n \quad (19)$$

2) Dec（解密）算法

要解密密文 c，首先计算 c`= cq1 =(gmhr)q1=(gq1)m (注意 hq1 =1modn)和 g’=gq1 使用密钥 q1 完成解密：

$$m=D(c)=\log_{g'}c' \quad (20)$$

为了有效地解密，消息空间应保持较小，因为离散对数无法快速计算。

3) 加法、乘法同态的实现

使用密文 E(m1)= c1 和 E(m2)= c2 对明文 m1 和 m2 进行同态加法，如所示:

$$c=C_1C_2h^{r}=(g^{m1}h^{r1})(g^{m2}h^{r2})h^{r}=g^{m1+m2}h^{r'}, \quad (21)$$

其中 r = r1 +r2 +r，可以看出 m1 + m2 可以很容易地从生成的密文 c 中恢复

要执行同态乘法，请使用 g1 与阶 n 和 h1 与阶 g1 并设置 g1 =e(g，g)，h1=e(g，h) 和 h = gαq2。然后，使用密文 c1 =E(m1)和 c2=E(m2)计算消息 m1 和 m2 的同态乘法如下:

$$c = e(c_1, c_2)h_1^{r} = e(g^{m1}h^{r1}, g^{m2}h^{r2})h_1^{r'} = g_1^{m1m2}h^{m1r2+r2m1+\alpha q2r1r2+r} = g_1^{m1m2}\, h_1^{r'} \quad (22)$$

可以看出，r’像 r 一样均匀分布，因此 m1m2 可以从得到的密文 c 中正确恢复。但是，c 现在在组 G1 中，而不是 G 中。因此，在 G 中不允许进行另一个同乘法运算，因为集合 G1 中没有配对。但是，在 G 中生成的密文仍然允许无限数量的同态加法。

3. BGN 在联邦学习中可能存在的缺点

1) 不安全性

在某些情况下 BGN 算法可能会导致加密失效。

2) 缺乏普遍性

适合较复杂模型的加密，但过于复杂（乘法超过一次）则会失去准确性，在加密模型近似椭圆曲线同态加密时其实用性才能最大化。

七、联邦学习中的全同态加密算法 BGV、BFV 方案

1. 背景

在上文提到的 Bootstrapping 技术被提出后， 2011 年两位大佬 Brakerski 和 Vaikuntanathan 提出了一个新的全同态加密体系，这一体系基于格加密的另一种假设 Learning With Errors（LWE），他们发明的全同态系统简称为 BGV 系统，这是一个有限级数的同态加密系统，但是可以通过 Bootstrapping 的方式来变成全同态系统

2. 全同态加密算法 BGV、BFV 方案

作为第二代全同态加密算法的主要构成，BGV、BFV 方案的主要特征包括： 1. 能在整数向量上进行高效的 SIMD 计算，即允许在单个指令周期内对多个数据元素执行相同的操作，这种能力可以用于同时处理加密数据中的多个位或多个加密数据元素，从而加快加密计算的速度 2. 能够快速完成高精度整数算术与快速向量的标量乘法 3. 引入 Leveled design 操作，可以层次化管理噪声，即将同态操作划分为不同的级别，每个级别包含一组操作。这样的设计有助于，使系统能够在不丢失准确性的情况下进行更多的计算。

3. 参数设置

选择大素数 (q)，定义多项式环 R_q = \mathbb{Z}_q[X]/(X^n + 1)，其中(n)是多项式的次数选择两个小整数 (t) 和 (m)，其中 (t) 是噪声大小，(m) 是模数。

4. 密钥生成

生成两个 (m)-bit 随机数 (s) 和 (e)，并计算 (a = -s + 2e)。生成一个 (m)-bit 随机数 (x)，并计算 (b = s + 2x)。公开 (a) 和 (b)，保留 (s) 作为私钥。

5. 加密过程

将明文 (m) 转换为多项式 (m(x))。生成一个 (t)-bit 随机数 (r)，并计算 (u = a + r).计算 (v = b + m \cdot q + r \cdot X^n).密文为 ((u, v))。

6. 解密过程

使用私钥 (s) 计算 (u \cdot s)。从结果中减去 (v)，然后将结果模 (q)，得到估计的 (m \cdot q)。除以 (q) 得到估计的 (m)。

7. 在联邦学习中的应用

全同态加密算法的一个运用是在隐私集合求交技术，即如何在允许两方在保持各自数据私密性的同时计算出两个集合的交集的技术。[16]

输入:数据集 DA 和 DB

输出:DA∩DB.

1）B 公司根据全同态加密算法生成公私钥对.

2）B 公司对集合 DB 中的元素进行加密,得到密文集合{c1,c2,……，$c_{|S_B|}$}并将其发给 A 公司.

3）A 公司对每个密文 ci，进行如下操作:①随机产生非零随机数 ri⊆ $\mathbb{Z}_n$; ②计算$d_i = r_i\prod_{x\in S_A}(c_i - x)$. A 公司将集合{d1,d2,……，$d_{|D_A|}$}发送给 B 公司

4）B 公司对{d1,d2,……，$d_{|D_A|}$}进行解密，解密为 0 的元素就是交集中的元素.

[1] LI T, SAHU A K, TALWALKAR A, et al. Federated learning:challenges, methods, and future directions[J]. IEEE signal processing magazine, 2020, 37(3):50-60.

[2] Chen Zhang，Yu Xie，Hang Bai，Bin Yu，WeihongLi，Yuan Gao.A survey on federated learning[N].Knowledge-Based Systems,2021-3-15

[3] C. Sun, A. Shrivastava, S. Singh and A. Gupta, "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 843-852, doi: 10.1109/ICCV.2017.97.

[4] H. Cho and Y. Baek, "Design and Implementation of a Smart Air Quality Monitoring and Purifying System for the School Environment," 2022 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 2022, pp. 1-4, doi: 10.1109/ICCE53296.2022.9730505.]H. Cho and Y. Baek, "Design and Implementation of a Smart Air Quality Monitoring and Purifying System for the School Environment," 2022 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 2022, pp. 1-4, doi: 10.1109/ICCE53296.2022.9730505.

[5] LIU J, LI T, XIE P, et al. Urban big data fusion based on deep learning: An overview[J/OL]. Information Fusion, 2020: 123-133. http://dx.doi.org/10.1016/j.inffus.2019.06.016. DOI:10.1016/j.inffus.2019.06.016.

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

[7] Pallier P.Public-key cryptosystems based on composite degree residue classes[C].Proceedings of EuroCrypt 99, Springer Verlag LNCS Series, 1999:223-238.

[8] 尚家秀,吴宗航,史腾飞.基于中国剩余定理的 Paillier 加密改进方法[J].信息技术与信息化,2023(03):133-136.

[9] LI Z R，HUANG Z C，CHEN C C，et al. Quantification of the leakage in federated learning［J］. ArXivpreprint ArXiv：1910.054672019.https://doi.org/10.48550/arXiv.1910.05467

[10] ZILLER A，TRASK A，LOPARDO A，et al. Pysyft：A library for easy federated learning［J］. Federated Learning Systems： Towards Next-Generation AI，2021：111-139.

[11] LIU Y，FAN T，CHEN T J，et al. Fate：An industrial grade platform for collaborative learning with data protection［J］. The Journal of Machine Learning Research，2021，22（1）：10320-10325

[12] Cheon J H , Kim A , Kim M ,et al.Homomorphic Encryption for Arithmetic of Approximate Numbers[J].Springer, Cham, 2017.DOI:10.1007/978-3-319-70694-8_15.

[13] Aono Y, Hayashi T, Wang L, et al. Privacy-preserving deep learning via additively homomorphic encryption[J]. IEEE Transactions on Information Forensics and Security, 2017, 13(5): 1333-1345

[14] Craig Gentry.A fully homomorphic encryption scheme[C].2009.Available from ProQuest Dissertations & Theses Global. (305003863). Retrieved from

https://www.proquest.com/dissertations-theses/fully-homomorphic-encryption-scheme/docview/305003863/se-2

[15] Lee, Hyang-Sook et al. 'On Insecure Uses of BGN for Privacy Preserving Data Aggregation Protocols'. 1 Jan. 2022 : 91 – 101.

[16] CHEN H,LAINE K,RINDAL P. Fast private set intersection from homomorphic encryption[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security,2017:1243-1255.