

保护隐私的集合相似性度量协同计算协议

逯绍锋¹, 胡玉龙², 逯跃锋^{3,4*}

(1. 东北大学 计算机科学与工程学院, 辽宁 沈阳 110189;

2. 中国交通通信信息中心, 北京 100011;

3. 山东理工大学 建筑工程学院, 山东 淄博 255049;

4. 中国科学院 地理科学与资源研究所资源与环境信息系统国家重点实验室, 北京 100101)

摘要:集合相似性度量是机器学习领域的基本问题之一,研究如何在保护数据隐私的前提下计算两个集合间的相似性问题,在保护数据隐私的机器学习、图形识别、生物信息学等方面有着重要的理论意义与应用价值。在机器学习中估算不同样本集合之间的相似性时,通常通过计算集合相似度来对样本之间的相似程度进行估算,这一类集合之间的相似度统称为集合距离。其中,最常用到的集合距离就是杰卡德距离。文中从集合间杰卡德距离入手,首先通过设计一种新的编码方法,对参与计算的数据进行位置数字编码,将相似性度量问题转化为求两集合间相同数字个数问题,进而结合异或思想,借助同态加密体制具体设计了可以保护隐私的集合杰卡德距离协同计算协议,从而解决了集合间相似性度量的隐私保护问题。模拟器证明该协议是安全的,结果分析表明协议可以高效安全地判定出两对象间集合数据的相似性,在保护隐私的集合相似性度量方面,该方法具备一定的普适性。

关键词:隐私保护;安全多方计算;杰卡德距离;集合相似性度量;机器学习

中图分类号:TP309

文献标识码:A

文章编号:1673-629X(2023)01-0137-07

doi:10.3969/j.issn.1673-629X.2023.01.021

Privacy Preserving Set Similarity Measurement Collaborative Computing Protocol

LU Shao-feng¹, HU Yu-long², LU Yue-feng^{3,4*}

(1. School of Computer Science and Engineering, Northeastern University, Shenyang 110189, China;

2. China Transport Telecommunications & Information Center, Beijing 100011, China;

3. School of Civil and Architectural Engineering, Shandong University of Technology, Zibo 255049, China;

4. State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China)

Abstract: Set similarity measurement is one of the basic problems in the field of machine learning. Studying how to calculate the similarity between two sets on the premise of protecting data privacy has important theoretical significance and application value in machine learning, graphics recognition, bioinformatics and so on. When estimating the similarity between different sample sets in machine learning, the similarity degree between samples is usually estimated by calculating the set similarity. This kind of similarity between sets is collectively referred to as set distance. Among them, the most commonly used set distance is Jaccard distance. Starting with the Jaccard distance between sets, we firstly design a new coding method to encode the position numbers of the data involved in the calculation, transform the similarity measurement problem into the problem of finding the number of the same numbers between two sets, and then design a set Jaccard distance collaborative calculation protocol that can protect privacy with the help of homomorphic encryption system. Thus, the privacy protection problem of similarity measurement between sets is solved. The simulator proves that the protocol is secure. The result analysis shows that the protocol can effectively and safely determine the similarity of set data between two objects. This method has certain universality in the measurement of set similarity to protect privacy.

Key words: privacy-preserving; security multi-party computation; Jaccard distance; set similarity measurement; machine learning

收稿日期:2022-03-06

修回日期:2022-07-08

基金项目:国家重点研发计划项目(2018YFC1506506);国家高分辨率对地观测系统重大专项(GFZX0404130304);山东省科技型中小企业创新能力提升工程项目(2021TSGC1056)

作者简介:逯绍锋(1982-),男,博士,CCF会员(J8551G),研究方向为信息安全;通信作者:逯跃锋,副教授,研究方向为数据匹配与隐私保护。

0 引言

机器学习 (Machine Learning, ML) 作为一种实现人工智能的方法,主要是通过算法来解析数据,不断学习,进而对事物做出判断和预测^[1]。机器学习广泛应用于计算机视觉、语音识别和自然语言处理等领域,是近些年学术研究上的重要方向。然而,蓬勃发展的机器学习技术在给人们带来便利的同时,也使数据安全与隐私面临更加严峻的挑战^[2]。目前,研究人员提出了许多解决机器学习中的隐私问题的方法^[3-6]。

解决隐私保护的安全多方计算 (Secure Multi-party Computation, SMC) 由 Yao 在文献[7]中首先提出。在安全多方计算中,参与方将各自的隐私数据输入到一个约定函数进行协同计算,可以实现保证参与方的原始隐私数据不被泄露的同时,输出正确计算结果^[8]。SMC 作为近年来密码学界的研究热点,被认为是解决协同计算问题中保护数据信息隐私的一项核心技术^[9-10]。利用 SMC 来实现保护隐私的机器学习由 Goldwasser 在 CRYPTO2018 上提出^[11],可以在机器学习的明文架构中引入安全多方计算技术实现隐私保护。

在机器学习、数据挖掘等领域,一般都是通过计算样本间的距离来完成对样本数据的相似性度量^[1]。采用什么样的方法来计算两者间距离,需要根据实际进行选择,计算结果直接影响到匹配的准确与否。比如采用欧氏距离计算两者之间的空间距离,使用汉明距离来度量两个文本或者图像之间的差异,使用杰卡德距离 (Jaccard distance) 来度量两个集合之间的相似程度^[12]。文献[13]对 DNA 序列进行 0-1 编码,以 GM 加密方案为主要工具,通过计算两序列的汉明距离的隐私保护方法,解决了保护隐私的 DNA 序列比较问题。保密的集合计算是安全多方计算的一个重要方向,文献[14]针对参与者集合为有限集合子集的问题,根据 Diffie-Hellman 的密钥分配方案基于离散对数困难性,给出一个具体的适用于有限集合子集的交集计算协议。文献[15]针对恶意模型下多方集合交集基数计算问题,提出了一个有效的计算协议。文献[16]通过对任意有理数进行特殊的编码,将有理数域上元素与集合关系问题转化为整数范围内的向量内积问题,以 Paillier 加密方案为主要工具,将两方集合保密计算推广到有理数领域。文献[17]提出了计算集合运算统计量的专用安全计算协议,可以在不泄露交集元素的情况下高效计算集合交集、并集及交集权值和等统计量,解决了隐私保护集合运算解决方案单一性问题。尽管人们针对隐私保护集合及机器学习已经研究并提出了一些安全多方计算解决方案^[18],但关于保护隐私的集合相似性度量问题,在当前的文献研究

中尚未发现。该文尝试通过解决杰卡德距离的安全多方计算问题,来进一步解决关于集合相似度计算的隐私保护问题。

1 相关知识

1.1 半诚实模型下协议的安全性定义

Goldreich 在文献[19]中指出:半诚实模型中的参与者会诚实地执行协议,但他们会根据自己在协议执行过程中获得的信息对其他参与者的秘密数据进行推导,下面给出半诚实模型下的安全性模拟范例。

设分别持有秘密数据 x 和 y 的协同计算参与者 A 和 B ,一起执行协议 π ,协同合作计算函数 $f(x, y) = (f_1(x, y), f_2(x, y))$ 。协议执行中参与计算双方获得信息序列为:

$$\text{view}_1^\pi(x, y) = \{(x, r_1, m_1^1, \dots, m_1^s, f_1(x, y))\} \quad (1)$$

$$\text{view}_2^\pi(x, y) = \{(y, r_2, m_2^1, \dots, m_2^t, f_2(x, y))\} \quad (2)$$

其中, r_1, r_2 分别为两参与者设置的随机数, $m_i^i (i = 1, 2, \dots, s)$ 、 $m_j^j (j = 1, 2, \dots, t)$ 表示两参与者收到的信息, $f_1(x, y)$ 、 $f_2(x, y)$ 表示协议 π 执行后,两参与者得到的输出结果。

定义:在半诚实模型下,对于计算函数 f 的协议 π ,如果存在概率多项式时间算法 S_1 与 S_2 ,使得:

$$\{S_1(x, f_1(x, y))\}_{x, y} \stackrel{c}{=} \{\text{view}_1^\pi(x, y)\}_{x, y} \quad (3)$$

$$\{S_2(y, f_2(x, y))\}_{x, y} \stackrel{c}{=} \{\text{view}_2^\pi(x, y)\}_{x, y} \quad (4)$$

则称 π 保密地计算了函数 f ,其中 $\stackrel{c}{=}$ 表示计算上不可区分。

1.2 杰卡德距离

在机器学习中估算不同样本集合之间的相似性时,通常要计算样本之间的相似度,将这一类集合之间相似度统称为集合距离。其中,最常用到的集合距离就是杰卡德距离。实际上,杰卡德距离和杰卡德相似系数互补,是通过计算集合间的交集和并集的基数来完成对两个集合的相似性衡量的一种集合距离。

杰卡德相似系数^[12]实际就是两个集合交集大小与并集大小的比值(见图1):

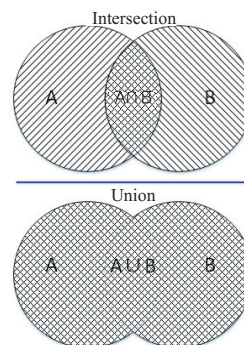


图1 两集合的杰卡德相似系数

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (5)$$

杰卡德距离与杰卡德相似系数互补,其公式可推导如下:

$$J_d(A, B) = 1 - J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \quad (6)$$

由于集合容斥原理, $|A \cup B| = |A| + |B| - |A \cap B|$, 因此杰卡德距离公式可以推演为如下公式:

$$J_d(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} = \frac{|A| + |B| - 2|A \cap B|}{|A| + |B| - |A \cap B|} \quad (7)$$

1.3 异或

异或是一种基于二进制的位运算,该文用符号 \oplus 表示,其运算法则是对运算符两侧数的每一个二进制位,同值取 0,异值取 1。如表 1 所示,可以得到异或运算具有下面的性质。

性质 1: 对于一个数,存在与 0 异或值不变的恒等特性。

表 1 异或运算

A	B	$P = A \oplus B$
0	0	0
0	1	1
1	0	1
1	1	0

性质 2: 对于一个数,存在与自身异或值为零的归零特性。

1.4 GM 同态加密方案

GM 同态加密方案是 Goldwasser 与 Micali 在文献 [20] 中提出的一种概率加密方案。

系统建立。系统选取 $N = pq$, 其中 p, q 为两个大素数, y 为模 N 的非二次剩余。系统公钥为 (N, y) , 私钥为 (p, q) 。

加密过程。明文 $m \in \{0, 1\}$, 任取随机数 $r \in Z_N^*$, 密文: $c = y^m r^2 \bmod N$ 。

解密过程。密文 $c < N$, 输入私钥 (p, q) , 求 Jacobi 符号 $(\frac{c}{N}) = (\frac{c}{p})(\frac{c}{q})$, 明文:

$$m = \begin{cases} 0, & \text{若 } ((\frac{c}{p}) = 1) \wedge ((\frac{c}{q}) = 1) \\ 1, & \text{若 } ((\frac{c}{p}) = -1) \vee ((\frac{c}{q}) = -1) \end{cases} \quad (8)$$

性质 3: GM 加密方案具有异或同态性质。

证明: 设 E 是一个加密算法, $c_1 = E(m_1, r_1)$ 是对 m_1 的加密, $c_2 = E(m_2, r_2)$ 是对 m_2 的加密, 若有 $c_1 c_2 =$

$E(m_1 \oplus m_2, r)$, 其中 r_1, r_2, r 是随机数, \oplus 表示异或, 则称 E 是一个异或同态加密算法。在 Goldwasser-Micali 方案中, 令 $r_1 r_2 = r$, 则有:

$$c_1 c_2 = (g^{m_1} r_1^2 \bmod N) (g^{m_2} r_2^2 \bmod N) = g^{m_1 \oplus m_2} (r_1 r_2)^2 \bmod N = g^{m_1 \oplus m_2} r^2 \bmod N \quad (9)$$

即有, $E(m_1)E(m_2) = E(m_1 \oplus m_2, r)$ 。因此, GM 加密方案具有异或同态性质^[10,20]。

性质 4: GM 加密方案具有自反性质。

证明: 根据前文所述异或运算的性质, 对 GM 加密方案进行异或同态操作时有:

$$c_1 c_2 c_2 = (g^{m_1} r_1^2 \bmod N) (g^{m_2} r_2^2 \bmod N) (g^{m_2} r_2^2 \bmod N) = g^{m_1 \oplus m_2 \oplus m_2} (r_1 r_2)^2 \bmod N = g^{m_1 \oplus 0} r^2 \bmod N = g^{m_1} r^2 \bmod N \quad (10)$$

即经过两次对同一密文的异或, 可使得密文 c_1 变成另一个密文 $c_1 c_2 c_2$, 但不影响解密的明文 m_1 。因此, 由以上推理可得到: 在 GM 加密方案中, 异或具有自反性质。

2 问题描述及转化

集合距离计算是机器学习、人工智能等领域的一个基本问题, 由于经常会用到敏感信息, 因此就不得不考虑参与计算各方数据的隐私问题。进而研究如何在保证各参与方数据隐私的前提下进行集合距离计算也变得十分必要, 该文将保护隐私的集合距离问题描述如下:

问题描述: 假设 Alice 输入集合数据 D_1 , Bob 输入集合数据 D_2 , Alice 和 Bob 想要共同协作计算出 D_1 和 D_2 两个集合的相似度距离, 同时又不泄露各自的元素数据信息。

问题转化: 为了更好地解决上述问题, 该文采用如下称之为位置编码的方法, 对参与计算的数据进行处理。假设存在集合 $U = \{x_1, x_2, \dots, x_i, \dots, x_n\}$, $i \in [1, n]$ 为一个全集, $|U| = n$ 。Alice 拥有集合 $D_1 = \{a_1, a_2, \dots, a_i, \dots, a_{n1}\}$, 其中 $a_i \in U$, Bob 拥有集合 $D_2 = \{b_1, b_2, \dots, b_j, \dots, b_{n2}\}$, 其中 $b_j \in U$, 该文将集合元素按照以下规则先行进行编码:

$$\text{规则: } \begin{cases} c_{aj} = i, & \text{若 } a_j = x_i \\ c_{bj} = i, & \text{若 } b_j = x_i \end{cases} \quad (11)$$

下面以具体的例子来说明问题转化过程。

例 1: 假设存在全集 $U = \{a, b, c, d, e, f, g, h, i, j, k, l\}$ 及两个集合 D_1, D_2 , 分别为: $D_1 = \{f, a, c, e\}$, $D_2 = \{f, a, c, i, l\}$ 。对集合 D_1 按照前文规则 (11) 对其编码, 可得到新的编码集合为 $A = \{6, 1, 3, 5\}$, 同样, 对集合 D_2 编码之后对应的编码集合为 $B = \{6, 1, 3, 9, 12\}$, 如表 2 所示。

表2 对集合进行位置编码

原始数据	位置编码
$D_1 = \{f, a, c, e\}$	$A = \{6, 1, 3, 5\}$
$D_2 = \{f, a, c, i, l\}$	$B = \{6, 1, 3, 9, 12\}$

计算原理说明:经过分析发现, D_1 和 D_2 两个集合经过位置编码后, 计算 D_1 和 D_2 两个集合的相似性度量问题就转化为求 A 和 B 两集合的相同数字个数问题。

该文通过对转化后的问题设计具体协议进行解决, 实现保护隐私的集合相似性度量协同计算。对编码后的两集合中的数字先进行逐位异或运算, 再根据异或运算的归零率特性, 一个数与它本身进行异或运算等于零, 否则为非零, 求出两个集合异或运算后新的集合中的零的个数, 即可求出 A 和 B 两集合相似性度量值, 对基数较小集合, 不足部分补零。例如, 在上述例1中两个集合 D_1 和 D_2 经过位置编码后, 再进行逐位异或运算的结果为: $C = \{0, 0, 0, 1, 1\}$, 结果集中0的个数为3, 从而可知 D_1 和 D_2 两个集合的相似性距离值为1/2。

3 具体协议

协议1: 保护隐私的集合杰卡德距离协同计算协议。

输入: Alice 和 Bob 分别输入私密的集合:

$D_1 = \{a_1, a_2, \dots, a_i, \dots, a_{n1}\}$, $D_2 = \{b_1, b_2, \dots, b_j, \dots, b_{n2}\}$, 其中 $a_i \in U$, $b_j \in U$, U 为包含集合的全集。

输出: 两集合的杰卡德距离 J_d 。

步骤1: (1) Bob 利用前述的编码规则将集合 D_2 进行位置编码, 生成新的集合 B , $B = \{b'_1, b'_2, \dots, b'_m\}$ 。

(2) Bob 生成自己的 GM 加密系统的公钥为 $PK = (\delta, n)$, 私钥为 $SK = (p, q)$ 。

(3) Bob 保存私钥 SK , 用公钥 PK 将集合 B 逐位加密, 获得密文 $E(B) = \{E(b'_1), E(b'_2), \dots, E(b'_m)\}$, 并将加密后的密文信息发给 Alice。

步骤2: (1) Alice 依前述的编码规则将集合 D_1 进行位置编码, 生成新的集合 A , $A = \{a'_1, a'_2, \dots, a'_m\}$ 。

(2) Alice 使用公钥 PK 对自有的集合 A 逐位加密, 得到密文 $E(A) = \{E(a'_1), E(a'_2), \dots, E(a'_m)\}$ 。

(3) Alice 将收到的 $E(A)$ 和 $E(B)$ 两两相乘, 并进行同态操作:

$$\begin{aligned}
 C &= E(A)E(B) = \\
 &\{E(a'_1)E(b'_1), E(a'_2)E(b'_2), \dots, \\
 &E(a'_m)E(b'_m)\} = \\
 &\{E(a'_1 \oplus b'_1), E(a'_2 \oplus b'_2), \dots, E(a'_m \oplus b'_m)\} = \\
 &E(A \oplus B) =
 \end{aligned}$$

$$\{E(c_1), E(c_2), \dots, E(c_m)\}$$

(4) Alice 选取随机置换 T 对上述结果进行置换, 得到随机置换后的 m 个密文:

$$\begin{aligned}
 (E(A \oplus B))^T &= \{E(a'_1 \oplus b'_1), E(a'_2 \oplus b'_2), \dots, \\
 &E(a'_m \oplus b'_m)\}^T
 \end{aligned}$$

并将该密文发回给 Bob。

(5) Alice 计算自己集合的基数 $CA = |D_1|$, 随同之前 $E(A)$ 一并发给 Bob。

步骤3: (1) Bob 收到置换后的 m 个密文后, 进行逐次解密, 得到:

$$(A \oplus B)^T = \{(a'_1 \oplus b'_1), (a'_2 \oplus b'_2), \dots, (a'_m \oplus b'_m)\}^T$$

$$\text{令 } C' = (A \oplus B)^T = \{c'_1, c'_2, \dots, c'_m\}。$$

(2) Bob 计算 $\sum_{i=1}^m \{C' | c'_i = 0\}$, 由于 $E(A \oplus B)$ 置换后得到的 $(A \oplus B)^T$ 和 $(A \oplus B)$ 两者序列中0的总数是不变的, 因此, 可由 Bob 统计出0的总数。

$$|D_1 \cap D_2| = |A \cap B| = \sum_{i=1}^m \{C' | c'_i = 0\}$$

(3) Bob 计算集合的大小值 $CB = |D_2|$, 并完成两集合的杰卡德相似性值计算。

$$J_d(A, B) = \frac{CA + CB - 2 |A \cap B|}{CA + CB - |A \cap B|}$$

(4) Bob 将 J_d 值传送给 Alice。

分析: 在协议1中, 由于 Bob 发送给 Alice 的是自己对集合 D_2 进行位置编码后的密文 $E(B)$, 其他合作计算方不能解密, 因此也不能从中得到集合 D_2 的隐私信息。另一方面, Alice 将自己的集合 D_1 进行位置编码并加密后获得 $E(A)$, 两者相乘后得到 $E(A)E(B) = E(A \oplus B)$, 若直接将计算结果传给 Bob, 由于 Bob 可以解密得到 $A \oplus B$, 在这个基础上, Bob 就可以通过加密机制的自反特性 $A \oplus B \oplus B$ 得到 A , 从而得到 Alice 的隐私集合 D_1 的信息。因此, 为了防止这种隐私泄露, Alice 将 $E(A \oplus B)$ 随机置换得到 $(E(A \oplus B))^T$, 这样使得 Bob 解密后只能得到 $(A \oplus B)^T$, 由于置换函数为 Alice 独自掌握, 因此 Bob 无法从中获取到 A 的信息, 从而 Alice 保护了自己的隐私。

4 协议分析

4.1 正确性分析

根据 GM 的同态加密方案的异或同态性, 任何数字对自身进行异或结果为零。并且在协议第3步虽然进行了置换, 但置换前后集合中零的总数是不变的, 即:

$$|D_1 \cap D_2| = |A \cap B| = \sum_{i=1}^n \{C | c_i = 0\} =$$

$$\sum_{i=1}^n \{C' | c'_i = 0\} \quad (12)$$

由于 $CA = |D_1|$ 、 $CB = |D_2|$ 的值不变,因此所得 J_d 的值即为两集合的杰卡德距离值。协议 1 是正确的。

4.2 安全性分析

定理 1: 协议 1 能够保密地协同计算出两集合的杰卡德距离值。

证明: 下面用模拟范例的方法来严格证明协议的安全性。

首先, 构造模拟器 S_1 , 令 $f_1(A, B) = f_2(A, B) = J_d(A, B)$, 并将 $(A, f_1(A, B))$ 输入 S_1 。

具体步骤如下:

第 1 步: S_1 接受输入 $(A, f_1(A, B)) = (A, J_d(A, B))$, 由于 S_1 拥有 $J_d(A, B)$, 它选取一个随机的集合, $\bar{B} = \{s'_1, s'_2, \dots, s'_n\}$, 满足 $J_d(A, B) = J_d(A, \bar{B})$ 。按照协议 S_1 将 $\{s'_1, s'_2, \dots, s'_n\}$ 进行位置编码后得到 $B' = \{b'_1, b'_2, \dots, b'_n\}$ 。对 B' 逐位加密后得到 n 长密文:

$$E(B') = \{E(b'_1), E(b'_2), \dots, E(b'_n)\}$$

第 2 步: S_1 将集合 D_1 进行位置编码得到:

$$A' = \{a'_1, a'_2, \dots, a'_n\}$$

然后, 再逐位加密, 得到 n 长密文:

$$E(A') = \{E(a'_1), E(a'_2), \dots, E(a'_n)\}$$

然后将收到的 $E(B')$ 和 $E(A')$ 两两相乘, 并进行同态操作:

$$\begin{aligned} C &= E(A')E(B') = \\ &\{E(a'_1)E(b'_1), E(a'_2)E(b'_2), \dots, \\ &E(a'_n)E(b'_n)\} = \\ &\{E(a'_1 \oplus b'_1), E(a'_2 \oplus b'_2), \dots, E(a'_n \oplus b'_n)\} = \\ &E(A' \oplus B') = \{E(c_1), E(c_2), \dots, E(c_n)\} \end{aligned}$$

S_1 选取随机置换 T 对上述结果进行置换, 得到随机置换后的 n 个密文:

$$(E(A' \oplus B'))^T = \{E(a'_1 \oplus b'_1), E(a'_2 \oplus b'_2), \dots, E(a'_n \oplus b'_n)\}^T$$

第 3 步: 将此 n 个密文逐次解密得到:

$$(A' \oplus B')^T = \{(a'_1 \oplus b'_1), (a'_2 \oplus b'_2), \dots, (a'_n \oplus b'_n)\}^T$$

令 $C' = (A' \oplus B')^T = \{c'_1, c'_2, \dots, c'_n\}$, 计算出:

$$|A \cap \bar{B}| = \sum_{i=1}^n \{C' | c'_i = 0\}$$

即两集合的杰卡德距离值:

$$J_d(A, \bar{B}) = \frac{CA + \bar{C}\bar{B} - 2|A \cap \bar{B}|}{CA + \bar{C}\bar{B} - |A \cap \bar{B}|}$$

在本协议中,

$$\text{view}_1^\pi(A, B) = \{A, E(A), E(B), E(C), J_d(A, B)\}$$

$$S_1(A, \bar{B}) = \{A, E(A), E(B'), E(C'), J_d(A, \bar{B})\}$$

由于 $J_d(A, B) = \sum_{i=1}^n \{C | c_i = 0\}$, $J_d(A, \bar{B}) =$

$\sum_{i=1}^n \{C' | c'_i = 0\}$, 且 $J_d(A, B) = J_d(A, \bar{B})$, 因此,

$$\sum_{i=1}^n \{C | c_i = 0\} = \sum_{i=1}^n \{C' | c'_i = 0\}.$$

同时, 由于 $E(B) = \{E(b_1), E(b_2), \dots, E(b_n)\}$, $E(B') = \{E(b'_1), E(b'_2), \dots, E(b'_n)\}$, 即 $E(B)$ 和 $E(B')$ 都是同一概率性公钥算法加密结果, 因此 $E(B)$ 和 $E(B')$ 计算不可分, 即 $E(B) \stackrel{c}{=} E(B')$ 。

又因为 $E(C) = (E(A)E(B))^T$, $E(C') = (E(A)E(B'))^T$, 且随机置换为同一置换 T , 从而可知:

$$(E(A)E(B))^T \stackrel{c}{=} (E(A)E(B'))^T, \text{ 即 } E(C) \stackrel{c}{=} E(C')$$

因此可得:

$$\{S_1(A, f_1(A, B))\}_{a, b \in U} \stackrel{c}{=} \{\text{view}_1^\pi(A, B)\}_{a, b \in U}$$

同理, 用类似的方法可以构造模拟器 S_2 , 使得:

$$\{S_2(A, f_2(A, B))\}_{a, b \in U} \stackrel{c}{=} \{\text{view}_2^\pi(A, B)\}_{a, b \in U}$$

因此, 协议 1 可以保密地协同计算出两集合的杰卡德距离值。

4.3 计算与通信复杂性分析

计算复杂性。文中协议借助 GM 同态加密方案来解决问题, 计算开销主要来自于模乘运算, 因此以模乘运算次数作为衡量指标。在协议 1 中, Alice 执行加密 m 次, Bob 执行加密、解密各 m 次, 由 GM 同态加密方案可知, 加密一次需要 2 次模乘运算, 解密一次需要 $\log p$ (p 为 GM 同态加密方案中的大素数私钥) 次模乘运算, 而 $m \leq n$, 因此协议 1 最多需进行 $4n + n \log p$ 次模乘运算。

通信复杂性。一般用协议交换信息的比特数或者通信轮数作为衡量通信复杂度指标, 在安全多方计算中通常用轮数^[10]。在协议 1 中, 计算双方共需进行两轮通信。

文献[13]基于 GM 同态方案解决汉明距离, 经过改造后也可用于计算集合相似度, 但文中协议要对数据长度进行扩展, 因此协议双方计算共需要做 $4n(4 + \log p)$ 模乘运算, 但协议同样需要 2 轮通信。

文献[14]提出协议 4 经过变形后可保密计算两个集合交集的基数, 用于集合相似度评价。该协议是借助 Diffie-Hellman 的密钥分配方案设计的解决方案, 协议中计算双方共需要做 $4n$ 次模乘运算, 进行 3 轮通信。

协议效率分析比较如表 3 所示。

表3 协议效率分析比较

协议	计算复杂性	通信复杂性
文献[13]协议	$4n(4 + \log p)$	2
文献[14]协议4	$4n$	3
文中协议	$4n + n \log p$	2

5 文中协议应用推广

从协议1可以看出,协同计算的核心是求解参与计算的两个集合的交集的基数,通过对协议1的必要改造后,可以构建其他类似的保护隐私的相似性度量协同计算协议,完成对Dice系数、Tversky系数、集合余弦相似性的计算。

Dice系数最初是被用来定量表达两个不同物种在自然界中的关联程度^[21],作为一种集合相似性度量函数,如今被广泛应用于深度学习图像分割和目标检测中。其公式含义可解读为真实数据集和预测数据集的正确部分之和占两者之和的比例,具体公式如下所示:

$$D(A, B) = \frac{2|A \cap B|}{|A| + |B|} \quad (13)$$

Tversky系数是由Tversky等在文献[22]中提出的用于特征相似性测度的函数,杰卡德系数和Dice系数均是Tversky系数的参数取特定值时的一种表达。

$$T(A, B) = \frac{|A \cap B|}{\alpha|A - B| + \beta|B - A| + |A \cap B|} \quad (14)$$

余弦距离和余弦相似度互补,集合的余弦距离是通过计算两集合夹角的余弦值作为衡量两个集合之间的相似性的一种方法^[12],具体公式如下:

$$1 - \cos(A, B) = 1 - \frac{|A \cap B|}{\sqrt{|A| \times |B|}} \quad (15)$$

事实上对于文本相似性也可以借助shingling算法将文档转化为集合,再借助文中协议进行保密协同计算出其相似性^[23]。对于海量数据,可以借助最小哈希和局部敏感哈希^[24]寻找相似集合的算法实现数据的相似性度量保密协同计算。

6 结束语

集合相似性度量问题是机器学习领域的基本问题之一,在能够保护隐私的前提下计算两集合对象间的相似性距离值是密码学中一个新的问题,在机器学习、图形识别、人工智能、生物信息学等方面有着重要的应用。该文设计了一种新的编码方法,将相似性度量问题转化为计算两集合相同数字个数的问题,使集合相似性度量距离计算问题普适化,并结合异或的思想,借助同态加密算法来解决杰卡德距离计算的隐私保护问题。然后,用模拟器证明了该协议是安全的,并分析了

协议可以高效安全地计算两集合对象间的杰卡德距离。最后,将问题扩展至其他方面,证明该协议具备一定的普适性,经过必要的改造后可以解决保密计算Dice系数、Tversky系数、集合余弦相似性等问题。由于该协议仍是基于半诚实模型,恶意模型下的保护隐私的集合相似性度量协同计算问题将是后续研究的方向。

参考文献:

- [1] MITCHELL T M. Machine learning[M]. New York: McGraw Hill, 2003: 2-3.
- [2] 魏立斐,陈聪聪,张蕾,等. 机器学习的安全问题及隐私保护[J]. 计算机研究与发展, 2020, 57(10): 2066-2085.
- [3] MOHASSEL P, RINDAL P. ABY3: a mixed protocol framework for machine learning[C]//Proc of the 2018 ACM SIGSAC conf on computer and communications security. New York: ACM, 2018: 3552.
- [4] JUVEKAR C, VAIKUNTANATHAN V, CHANDRAKASAN A. GAZELLE: a low latency framework for secure neural network inference[C]//Proc of the 27th USENIX conf on security symp. Berkeley: USENIX Association, 2018: 1651-1668.
- [5] CHAUDHARI H, CHOUDHURY A, PATRA A, et al. ASTRA: high throughput 3PC over rings with application to secure prediction[C]//Proc of the 2019 ACM SIGSAC conf on cloud computing security workshop. New York: ACM, 2019: 81-92.
- [6] BYALI M, CHAUDHARI H, PATRA A, et al. FLASH: fast and robust framework for privacy-preserving machine learning[J]. Proceeding on Privacy Enhancing Technologies, 2020, 2020(2): 459-480.
- [7] YAO A C. Protocols for secure computations[C]//Proceedings of the 23rd annual IEEE symposium on foundations of computer science. Chicago: IEEE, 1982: 160-164.
- [8] 罗永龙,黄刘生,荆巍巍,等. 保护私有信息的叉积协议及其应用[J]. 计算机学报, 2007, 30(2): 248-254.
- [9] SHEN E, VARIA M, CUNNINGHAM R K, et al. Cryptographically secure computation[J]. Computer, 2015, 48(4): 78-81.
- [10] 陈振华,李顺东,陈立朝,等. 点和区间关系的全隐私保密判定[J]. 中国科学: 信息科学, 2018, 48(2): 187-204.
- [11] GOLDWASSER S. From dea to impact, the cryptostory: what's next? [C/OL]//IACR distinguished lecture of CRYPTO2018. California, USA: [s. n.], 2018. <https://www.iacr.org/cry-ptodb/data/pap-er.php?pubkey=29941>.
- [12] 庞俊,谷峪,许嘉,等. 相似性连接查询技术研究进展[J]. 计算机科学与探索, 2013, 7(1): 1-13.
- [13] 马敏耀,徐艺,刘卓. 隐私保护DNA序列汉明距离计算问题[J]. 计算机应用, 2019, 39(9): 2636-2640.
- [14] 周素芳,李顺东,郭奕旻,等. 保密集合相交问题的高效计

- 算[J]. 计算机学报, 2018, 41(2): 464–480.
- [15] ZHANG E, LIU F H, LAI Q, et al. Efficient multi-party private set intersection against malicious adversaries[C]//Proceedings of the 2019 ACM SIGSAC conference on cloud computing security workshop. New York: ACM, 2019: 93–104.
- [16] 窦家维, 刘旭红, 王文丽. 有理数域上两方集合的高效保密计算[J]. 计算机学报, 2020, 43(8): 1397–1413.
- [17] 宋祥福, 盖敏, 赵圣楠, 等. 面向集合计算的隐私保护统计协议[J]. 计算机研究与发展, 2020, 57(10): 2221–2231.
- [18] 郭娟娟, 王琼霄, 许新, 等. 安全多方计算及其在机器学习中的应用[J]. 计算机研究与发展, 2021, 58(10): 2163–2186.
- [19] GOLDREICH O. Foundations of cryptography: basic applications[M]. London: Cambridge University Press, 2004: 599–729.
- [20] GOLDWASSER S, MICALI S. Probabilistic encryption[J]. Journal of Computer and System Sciences, 1984, 28(4): 270–299.
- [21] DICE L R. Measures of the amount of ecologic association between species[J]. Ecology, 1945, 26(3): 297–302.
- [22] TVERSKY A. Features of similarity[J]. Psychological Review, 1977, 84: 327–352.
- [23] 李志华, 陈超群, 李村, 等. 基于关键词重提取的密文文本相似性度量方法研究[J]. 计算机科学, 2016, 43(8): 95–99.
- [24] LESKOVEC J, RAJARAMAN A, ULLMAN J D. Mining of massive datasets[M]. Cambridge: Cambridge University Press, 2018: 1520–1521.
- +++++
- (上接第136页)
- 数据过滤方案[J]. 计算机工程与应用, 2015, 51(24): 78–85.
- [11] 王斌. 工业物联网信息安全防护技术研究[D]. 成都: 电子科技大学, 2018.
- [12] 袁勇, 王飞跃. 区块链技术发展现状与展望[J]. 自动化学报, 2016, 42(4): 481–494.
- [13] 沈苏彬, 毛燕琴, 李莉. 一种面向非数字货币的区块链通用应用方案[J]. 南京邮电大学学报: 自然科学版, 2019, 39(1): 1–11.
- [14] 赵阔, 邢永恒. 区块链技术驱动下的物联网安全研究综述[J]. 信息安全学报, 2017(5): 1–6.
- [15] FROMKNECHT C, VELICANU D, YAKOUBOV S. A decentralized public key infrastructure with identity retention[J]. IACR Cryptol. ePrint Arch., 2014, 26(7): 1–16.
- [16] 周启惠, 邓祖强, 邹萍, 等. 基于区块链的防护物联网设备DDoS攻击方法[J]. 应用科学学报, 2019, 37(2): 213–223.
- [17] ZHANG Z K, CHO M C Y, WANG C W, et al. IoT security: ongoing challenges and research opportunities[C]//2014 IEEE 7th international conference on service-oriented computing and applications. Matsue: IEEE, 2014: 230–234.
- [18] 陈诗鹏, 陈彬, 代明军, 等. 一种基于区块链的物联网架构[J]. 物联网学报, 2020, 4(2): 78–83.
- [19] LI D, PENG W, DENG W, et al. A blockchain-based authentication and security mechanism for IoT[C]//2018 27th international conference on computer communication and networks (ICCCN). Hangzhou: IEEE, 2018: 1–6.