

# 基于同态加密的分布式隐私保护线性回归分析模型

李 娟, 马 飞

(北方民族大学 计算机科学与工程学院, 宁夏 银川 750021)

**摘 要:** 线性回归是对数据进行统计分析的重要方法之一, 但对具有隐私保护的分布式多数据源线性回归分析的研究还较少. 针对该问题, 利用加性同态加密, 提出了一种在非信任的分布式环境下具有隐私保护的协作线性回归分析模型. 该模型充分利用分布式环境的计算能力, 由各数据源端与服务器端协作进行线性回归参数的计算, 各数据源的敏感数据在整个分析过程中都处于保密状态. 最后分析了模型在半诚实模式下的安全性, 并对所实现模型进行了性能测试.

**关键词:** 隐私保护; 同态加密; 线性回归; 半诚实模式

**中图分类号:** TP309.7

**文献标识码:** A

**文章编号:** 1000-7180(2016)01-0110-04

## A Model on Distributed Privacy Preserving Linear Regression Analysis Based on Homomorphic Encryption

LI Juan, MA Fei

(School of Computer Science and Engineering, Beifang University of Nationalities, Yinchuan 750021, China)

**Abstract:** Linear regression is one of the important methods of statistical analysis, but the research on linear regression analysis with privacy preserving for multiple data source in distributed environment is relatively less. To solve this problem, by using additive homomorphic encryption, a model on collaborative linear regression analysis with privacy preserving in non-trusted distributed environment is presented. The model makes full use of the calculating ability of distributed environment, the clients and statistical server collaboratively calculate linear regression parameters, and the sensitive data of each data source are protected in analysis process. Finally, the security of model is discussed in Semi-Honest Mode, and the performance of model is tested.

**Key words:** Privacy Preserving; homomorphic encryption; linear regression; Semi-Honest Mode

### 1 引言

统计分析已被广泛应用于研究与生活领域, 而目前的统计分析数据常常来自多个数据源, 数据的拥有者出于研究与商业目的, 希望进行合作统计分析, 但由于数据的隐私性, 数据的持有者希望在对数据进行合作统计分析的同时能保护数据隐私. 目前多数据源的隐私保护计算方案主要是安全多方计算<sup>[1-3]</sup>, 但该方案由于涉及繁琐的加解密步骤、复杂的协议交互过程, 从而较为低效. 如何利用分布式环境的计算与存储能力, 由各数据的持有者协作进行统计分析且不破坏数据隐私, 这是本文重点研究及

解决的问题.

本文利用 Paillier 加密算法<sup>[4-5]</sup>的加法同态性设计了一种在分布式环境下对多数据源进行具有隐私保护的协作统计计算线性回归参数方案. 该方案由提供统计数据的用户端与服务器端协作完成整个统计计算过程, 方案不但充分利用了分布式环境计算与存储能力, 而且在整个统计分析过程中用户的隐私数据都处于保密状态.

### 2 线性回归与加性同态

#### 2.1 加性同态

公钥加密方案:  $\epsilon = (\text{KeyGen}, \text{Enc}, \text{Dec})$ , 其中  $k$



### 3.2.1 阶段 1

假设 CClientX、CClientY 为叶结点, CClientZ 为二者的父结点. CClientX、CClientY 与 CClientZ 分别具有隐私数据对:  $(x_a, y_a)$ 、 $(x_b, y_b)$ 、 $(x_c, y_c)$ . 各 CClient 用  $K_{pk}$  加密隐私数据:  $x_i, y_i, x_i \cdot y_i, x_i^2$  及 1, 得到:  $E_{pk}(x_i), E_{pk}(y_i), E_{pk}(x_i^2), E_{pk}(x_i y_i) E_{pk}(1)$ .

CClientX 与 CClientY 把  $E_{pk}(x_a), E_{pk}(y_a), E_{pk}(x_a y_a), E_{pk}(x_a^2), E_{pk}(x_b), E_{pk}(y_b), E_{pk}(x_b y_b), E_{pk}(x_b^2)$  及两个  $E_{pk}(1)$  交于其父节点 CClientZ 做统计计算, 根据加同态性质有下式成立:

$$E_{pk}(x_a + x_b + x_c) = E_{pk}(x_a) E_{pk}(x_b) E_{pk}(x_c) \quad (6)$$

$$E_{pk}(y_a + y_b + y_c) = E_{pk}(y_a) E_{pk}(y_b) E_{pk}(y_c) \quad (7)$$

$$E_{pk}(x_a y_a + x_b y_b + x_c y_c) = E_{pk}(x_a y_a) E_{pk}(x_b y_b) E_{pk}(x_c y_c) \quad (8)$$

$$E_{pk}(1 + 1 + 1) = E_{pk}(1) E_{pk}(1) E_{pk}(1) \quad (9)$$

$$E_{pk}(x_a^2 + x_b^2 + x_c^2) = E_{pk}(x_a^2) E_{pk}(x_b^2) E_{pk}(x_c^2) \quad (10)$$

CClientZ 把计算结果向其父结点传递, 其父节点重复计算过程.

### 3.2.2 阶段 2

数据经过计算与传递, KClient 得到以下加密中间统计结果并交于统计服务器:

$$E_{pk}(u_y) = \prod_{i=1}^n E_{pk}(y_i); E_{pk}(u_x) = \prod_{i=1}^n E_{pk}(x_i);$$

$$E_{pk}(z_{xy}) = \prod_{i=1}^n E_{pk}(x_i y_i); E_{pk}(n) = \prod_{i=1}^n E_{pk}(1);$$

$$E_{pk}(w_x) = \prod_{i=1}^n E_{pk}(x_i^2).$$

### 3.2.3 阶段 3

统计服务器解密后得到明文  $z_{xy} = \sum_{i=1}^n x_i y_i, u_x = \sum_{i=1}^n x_i, u_y = \sum_{i=1}^n y_i, w_x = \sum_{i=1}^n x_i^2$  及  $n$ , 代入式(2), 所得计算结果即为回归参数的值.

## 4 模型安全性分析与性能测试

### 4.1 模型安全性分析

半诚实模型<sup>[8]</sup>: 参与计算的模型元素都遵循协议要求的步骤, 但希望能额外获得除最终结果以外的其他信息.

攻击方案 1: 全部 CClient 可信, SServer 不可信.

攻击目的: 所有 CClient 遵循方案所要求的计算步骤, 但非可信 SServer 想获得各 CClient 的隐私数据.

模型安全性分析: 方案中, 加密数据经过计算向根节点传递后, SServer 最终获得的是加密的中间结果, 所以即使 SServer 对其解密, 也无法推出 CClient 各自原始隐私数据.

攻击方案 2: 部分 CClient 共谋, SServer 可信.

攻击目的: SServer 可信, 部分 CClient 共谋希望推出其他非共谋的 CClient 的隐私数据.

模型安全性分析: 由于所有 CClient 的隐私数据及中间计算结果都被可信 SServer 的公钥所加密, 而在 SServer 私钥安全的情况下, 共谋 CClient 是无法对其解密, 故非共谋 CClient 的隐私信息在计算的任何阶段都不会被泄露.

攻击方案 3: 部分 CClient 与 SServer 共谋.

攻击目的: 部分 CClient 与 SServer 共谋, 希望推出非共谋的 CClient 的隐私数据.

模型安全性分析: 设  $n$  为 CClient 的总数,  $m$  为其中共谋 CClient 的数量. (1) 当  $m=n-1$  时, 即只有一个 CClient 是非共谋, 其他  $n-1$  个 CClient 都与 SServer 共谋, 当非共谋 CClient 把自己的隐私数据用 SServer 的公钥加密后交于其父结点, 其父结点与其他共谋 CClient 仍然按照协议要求的步骤执行, 但都只用加密后的“0”与非共谋的 CClient 的加密隐私数据做同态计算并把结果向根传递, SServer 收到的数据其实就是非共谋 CClient 的加密隐私数据, SServer 对其解密即可获得其隐私数据. (2) 当  $m < n-1$  时, 即非共谋的 CClient 个数  $\geq 2$  时, 按照协议的计算步骤, SServer 收到的一定是非共谋 CClient 隐私数据做加法之后的中间结果的加密形式, 所以 SServer 即使对其解密, 也无法从隐私数据“和”中反推出非共谋 CClient 各自的隐私数据. 故对本方案而言, 只要满足  $m < n-1$  的条件, 则共谋的 CClient 与 SServer 不能推出非共谋 CClient 的隐私数据.

### 4.2 模型性能测试

测试模型在计算回归参数时的时间代价.

由于方案采用分布式计算回归参数值, 计算的时间代价主要集中于: (1) 客户端对数据进行 Paillier 算法加密, 及进行加性同态计算时所用时间; (2) 统计服务器对中间统计结果进行解密及计算最终结果所用时间.

模型实现环境参数:模型中的各组件都采用性能一致的 Acer 笔记本电脑,共设置了 8 台电脑,其中具有 CClient 身份的电脑共 6 台, KClient 和 SServer 各一台. 设备通过 IEEE 802. 11g 网络进行互联. 具体参数如表 1 所示.

表 1 模型实现环境参数

|          |                               |
|----------|-------------------------------|
| OS       | Windows 7 32-bit professional |
| CPU      | 1. 70 GHz,2. 86 GB RAM        |
| Software | VC++Crypto++库                 |

为了保证 Paillier 加密算法具有足够的语义安全,设置算法参数如表 2 所示.

表 2 Paillier 加密算法参数

|                           |      |
|---------------------------|------|
| $N$                       | 1024 |
| $g$                       | 160  |
| 公钥 $\langle N, g \rangle$ | 1184 |

基于以上参数, Paillier 加密过程实质是做两次 1024 bit 的幂运算,一次 2048 bit 的乘法运算,而解密过程是做一次 2048 bit 的幂运算. 私钥与公钥利用生成算法在模型运行前按要求位数已生成. 模型的客户端模块(CClient)和服务端模块(SServer)都处于工作状态. 共进行了 15 次运行测试.

符号说明如下:

$t_E(\cdot)$  表示单次 Paillier 算法加密操作所用时间;

$t_D(\cdot)$  表示单次 Paillier 算法解密操作所用时间;

$t_H(\cdot)$  表示单次同态操作所用时间;

$t_{out}$  指 SServer 对收到的中间结果进行解密所用——时间与计算最终统计结果的时间之合.

$t_E(\cdot), t_H(\cdot)$  是在 6 台 CClient 上分别进行测试得到相应的数值,  $t_D(\cdot)$  与  $t_{out}$  是在 SServer 上测试得到. 单个 CClient 做加密及同态运算所花费时间与方案规模(参与统计计算的客户端数)无关. SServer 最后对中间结果解密及计算最终统计结果所用时间也与规模无关. 计算“回归参数”时间代价见表 3.

表 3 计算“回归参数”时间代价

|              | Max   | Min   | Mean   |
|--------------|-------|-------|--------|
| $t_E(\cdot)$ | 40. 7 | 34. 2 | 38. 01 |
| $t_H(\cdot)$ | 76. 8 | 70. 6 | 73. 76 |
| $t_D(\cdot)$ | 42. 1 | 37. 1 | 39. 92 |
| $t_{out}$    | 43. 1 | 36. 8 | 40. 13 |

4. 3 方案特点

用户数据由客户端独立存储,方便用户对其访问控制. 统计计算由客户端与服务器共同完成,降低了传统服务器因集中统计计算与存储带来的安全隐患及计算复杂性高的问题,适用于大数据量计算. 数据在传递的同时进行统计计算,不必预先合并数据. 除中间与最终统计结果以外,用户隐私数据在整个统计计算过程中都被保护. 加密操作仅在客户端提供原始统计数据时出现,而解密操作也仅在统计服务器对中间结果解密时发生,所以方案只涉及很少的加解密过程.

除所用公式不同以外,本文的模型完全可应用到计算样本均值、频率分布、相关系数等统计量的隐私保护计算中.

5 结束语

本文利用 Paillier 加密算法的加法同态性,提出了一种在互不信任的分布式环境下具有隐私保护的线性回归参数计算模型. 该模型充分利用分布式环境的计算与存储能力,由客户端与统计服务器协作进行回归参数计算. 方案不但能保护用户数据的隐私性,而且降低了计算复杂度.

由于医疗系统中监测患者体征的可穿戴设备、智能电网中的智能电表等应用都需要采集设备环境或用户的体征数据,而这些数据都具有很强的隐私性,对这类数据进行统计分析时,就可采用本文提出的模型.

对于部分客户端与统计服务器进行共谋对模型进行攻击时,必须满足非共谋客户端数大于 1 时模型才能正确执行,这将是以后需要对模型完善的地方.

参考文献:

[1] 叶云. 保护私有数据的合作计算问题及其应用研究[D]. 安徽:中国科学技术大学,2012.

[2] 耿涛. 安全多方计算若干问题以及应用研究[D]. 北京:北京邮电大学,2012.

[3] 张斌. 高效安全的多方计算基础协议及应用研究[D]. 济南:山东大学,2012:77-81.

[4] 白健,杨亚涛,李子臣. Paillier 公钥密码体制同态特性及效率分析[J]. 北京电子科技学院学报, 2012, 29 (4):77-81.

[5] 周青婷. 基于 Paillier 密码体制的点积协议研究[D]. 云南:云南大学,2012.

AFDX 交换机警管模块的警管算法,并基于该算法对 AFDX 警管模块进行硬件设计.通过仿真实验证明了该算法可以满足 ARINC 664 规范中对 AFDX 网络流量控制的要求<sup>[5]</sup>,对进入 AFDX 交换机内的数据帧进行警管,具有良好的流量控制功能.本算法已经在相关的 AFDX 交换机设计中得以应用,接下来我们将继续研究该警管算法在不同网络交换系统中实现的可能性.

参考文献:

[1] (美)Andrew S Tanenbaum. 计算机网络(第四版)[M]. 潘爱民,译. 北京:清华大学出版社,2004.

[2] Roy Courtney. Circulation prior to adoption consideration Draft3 of project paper 644: Aircraft data network, part 7-avionics full duplex switched ethernet (AFDX) network, september 13, 2004 [S]. Stockholm: AEEC, 2004.

[3] Teresa Schuster, Dinesh Verma. Networking concepts

comparison for avionics architecture[C]//IEEE/AIAA 27th Digital Avionics Systems Conference, st. paul,  $\mu$ N:IEEE,2008:1-11.

[4] 刘增基,鲍民权,邱智亮. 交换原理与技术[M]. 北京:人民邮电出版社,2007.

[5] 王红春,张涛,牛文生. 面向 AFDX 网络的交换机流量管制算法[J]. 西北工业大学学报, 2012, 30(02): 301-304.

作者简介:

张栩培 男,(1991-),硕士研究生. 研究方向为计算机控制技术. E-mail:panfeng1102@163. com.

贺占庄 男,(1962-),博士,研究员,博士生导师. 研究方向为计算机系统结构、SoC 和 IP 核设计.

王瑞晓 女,(1985-),工程师. 研究方向为计算机系统结构.

李龙飞 男,(1988-),博士研究生. 研究方向为计算机系统结构.

(上接第 113 页)

[6] 陈智罡,王箭,宋新霞. 全同态加密研究[J]. 计算机应用研究,2014,30(4):35-37.

[7] Gentry C. Fully homomorphic encryption using ideal lattices [C] // Proceedings of the forty-first annual ACM Sysposium on Theory of Computing. NY, USA: IEEE,2009:169-178.

[8] Kantarcioglu, M Kardes. Privacy-preserving data mining in the malicious model[C]// Proceeding of the Seventh IEEE International Data Uining Wo9rkshops, O-

maha, NE:IEEE, 2007:717-722.

作者简介:

李娟 女,(1975-),硕士,副教授. 研究方向为云计算、网络安全、社交网络与隐私保护.

马飞 男,(1976-),博士,副教授. 研究方向为数据挖掘、云计算,社交网络分析与隐私保护.

E-mail:feixiangflying33@nxu. edu. cn.