《基于隐私保护的机器学习若干技术研究 》
开题报告

数学与信息科学学院

汇报人：刘坤　　　　导师：唐春明教授
汇报时间：2022年03月08日

目录

CONTENTS

01

问题和背景

02

研究动机

03

解决方案

04

目的意义

问题和背景

Problem & Background

✓问题　　✓国内外研究概况

问题和背景

Problem & Background

机器学习

特征提取，模型训练， 查询匹配

现有的计算工作

有效性差，效率低，安全性弱等问题

模型和数据的隐私

场景1 服务器与服务器之间

场景2 用户与服务器之间

场景3 用户与用户之间

隐私性与效率

类同态加密 安全性高 效率低

安全乘法协议

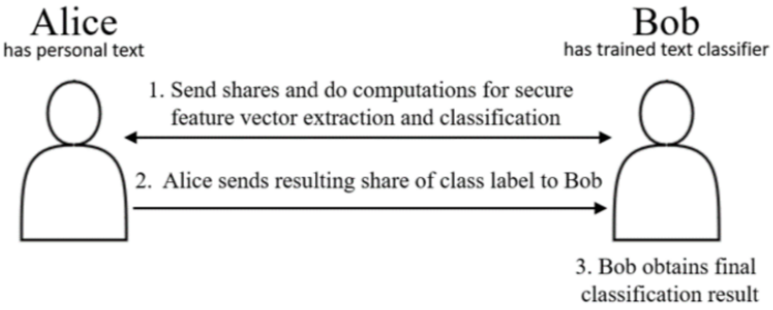
研究的问题

基于隐私保护的朴素贝叶斯分类的安全两方计算

支持分类模型训练的安全外包计算

国内外研究概况

现有的隐私保护朴素贝叶斯协议基于文献[1]，用于文本分类。可信第三方分发乘法三元组。

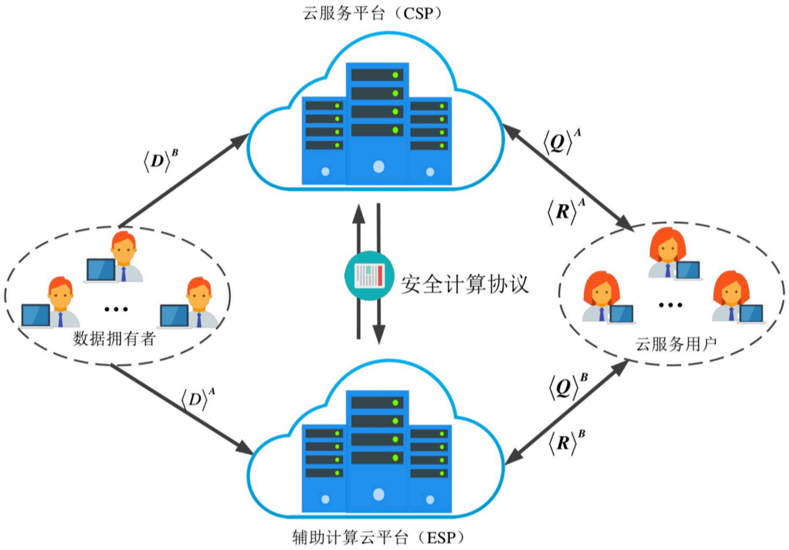

Alice
has personal text
Bob
has trained text classifier
1. Send shares and do computations for secure feature vector extraction and classification
2. Alice sends resulting share of class label to Bob
3. Bob obtains final classification result

$$\hat{c} = \mathrm{argmax}_c \left[\log(\Pr(c)) + \sum_{k=1}^{d} \log(\Pr(x_k|c)) \right]$$

[1] Resende, Amanda, Davis Railsback, Rafael Dowsley, Anderson CA Nascimento, and Diego F. Aranha. "Fast privacy-preserving text classification based on secure multiparty computation." *IEEE Transactions on Information Forensics and Security* (2022).

国内外研究概况

文献[2]利用安全多方计算，在多个数据源参与方垂直或水平分割下，支持隐私保护的线性回归方案。通过使用秘密共享技术，共同训练模型需要参与方时刻保持在线并参与后续的计算。大多数现有方案计算开销大。



[1] Resende, Amanda, Davis Railsback, Rafael Dowsley, Anderson CA Nascimento, and Diego F. Aranha. "Fast privacy-preserving text classification based on secure multiparty computation." *IEEE Transactions on Information Forensics and Security* (2022).

[2] Liu, Lin, Jinshu Su, Rongmao Chen, Ximeng Liu, Xiaofeng Wang, Shuhui Chen, and Hofung Leung. "Privacy-preserving mining of association rule on outsourced cloud data from multiple parties." In *Australasian Conference on Information Security and Privacy*, pp. 431-451. Springer, Cham, 2018.

研究动机

Motivation

✓论文的理论依据 ✓研究方法 ✓研究内容

论文的理论依据

The theoretical basis of the paper

- 通常，在分类学习的安全多方计算协议中，为了降低本地计算量，增加可信第三方，易遭攻击[3]，而且限制了场景；
- 在两方服务器协同训练分类模型中，安全乘法协议和安全两方计算协议可解决上述。

[3] Miller, David J., Zhen Xiang, and George Kesidis. "Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks." *Proceedings of the IEEE* 108, no. 3 (2020): 402-433.

论文的理论依据

The theoretical basis of the paper

- 在外包计算中，采用双云模型，即云服务平台和辅助计算云平台，与传统的云计算模型相比，协议不需要密钥生成中心分发密钥。
- 隐私保护技术采用Paillier同态加密算法和加法秘密共享
- 可进行回归，决策树和k近邻分类

[3] Miller, David J., Zhen Xiang, and George Kesidis. "Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks." *Proceedings of the IEEE* 108, no. 3 (2020): 402-433.

研究方法

Research Method

查阅相关文献

广大电子图书馆

谷歌学术镜像网站

实验重现

用虚拟机在linux系统下重现论文实验

通过阿里云限时免费的在线GPU

笔记

对相同类似的问题进行比较分类
对重要论文研读并笔记
完善个人网页

论文撰写

导师指导

Grammarly等工具

研究内容

Research Contents

隐私计算

安全多方计算协议、秘密共享、安全比较、安全转换（2toQ，Qto2）、安全乘法、混淆电路、Paillier同态加密、安全比特分解、安全定点数截断等

机器学习分类模型

线性回归、逻辑回归、贝叶斯分类模型（朴素、高斯、伯努利）、k近邻和神经网络等

实验编程

利用C++ 、python等语言在linux系统下模拟仿真

解决方案

Solution

✓研究分析 ✓方案及讨论

研究分析

Research Analysis

安全比较协议

1. 求差 $[\![\mathrm{diff}]\!]_q \leftarrow [\![x]\!]_q - [\![y]\!]_q$

2. 最高有效位的秘密共享

安全乘法协议

计算矩阵X $\cdot Y$，寻找均匀随机$([\![U]\!]_q, [\![V]\!]_q, [\![W]\!]_q)$
使得 $W = UV,$

朴素贝叶斯分类

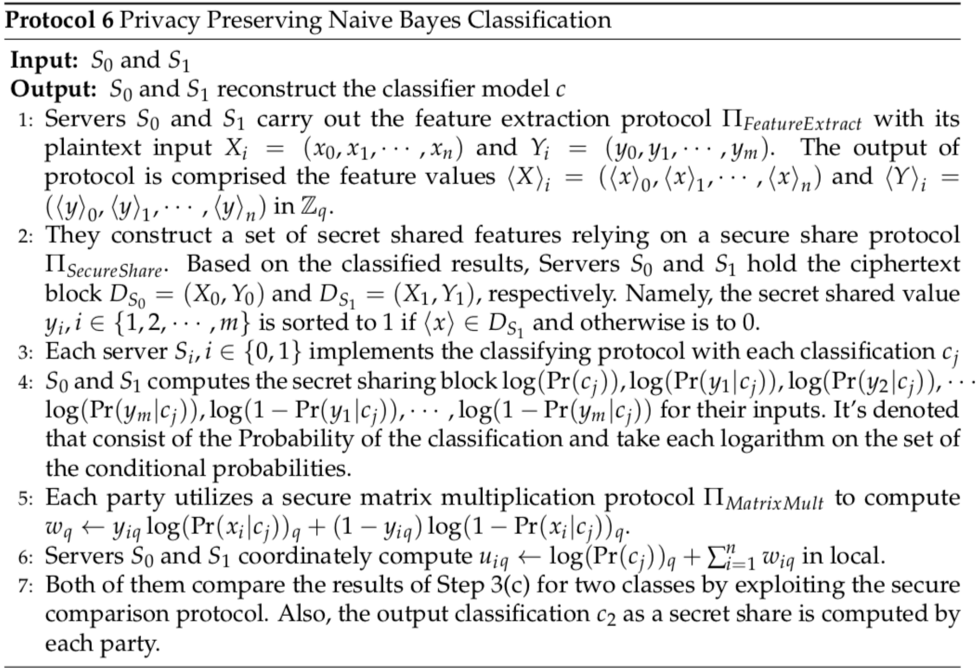
对概率取对数

$$\begin{aligned}\log(\Pr(c|x)) &= \log\left(\Pr(c)\prod_{i=1}^{d}\Pr(x_i|c)\right)\\ &= \log(\Pr(c)) + \sum_{i=1}^{d}\log(\Pr(x_i|c)).\end{aligned}$$

方案及讨论

Solution and Discussion

安全两方计算的朴素贝叶斯分类模型



Protocol 6 Privacy Preserving Naive Bayes Classification

Input: S_0 and S_1

Output: S_0 and S_1 reconstruct the classifier model c

1: Servers S_0 and S_1 carry out the feature extraction protocol $\Pi_{FeatureExtract}$ with its plaintext input $X_i = (x_0, x_1, \cdots, x_n)$ and $Y_i = (y_0, y_1, \cdots, y_m)$. The output of protocol is comprised the feature values $\langle X \rangle_i = (\langle x \rangle_0, \langle x \rangle_1, \cdots, \langle x \rangle_n)$ and $\langle Y \rangle_i = (\langle y \rangle_0, \langle y \rangle_1, \cdots, \langle y \rangle_n)$ in $\mathbb{Z}_q$.

2: They construct a set of secret shared features relying on a secure share protocol $\Pi_{SecureShare}$. Based on the classified results, Servers S_0 and S_1 hold the ciphertext block $D_{S_0} = (X_0, Y_0)$ and $D_{S_1} = (X_1, Y_1)$, respectively. Namely, the secret shared value $y_i, i \in \{1, 2, \cdots, m\}$ is sorted to 1 if $\langle x \rangle \in D_{S_1}$ and otherwise is to 0.

3: Each server $S_i, i \in \{0, 1\}$ implements the classifying protocol with each classification c_j

4: S_0 and S_1 computes the secret sharing block $\log(\Pr(c_j)), \log(\Pr(y_1|c_j)), \log(\Pr(y_2|c_j)), \cdots$ $\log(\Pr(y_m|c_j)), \log(1 - \Pr(y_1|c_j)), \cdots, \log(1 - \Pr(y_m|c_j))$ for their inputs. It's denoted that consist of the Probability of the classification and take each logarithm on the set of the conditional probabilities.

5: Each party utilizes a secure matrix multiplication protocol $\Pi_{MatrixMult}$ to compute $w_q \leftarrow y_{iq} \log(\Pr(x_i|c_j))_q + (1 - y_{iq}) \log(1 - \Pr(x_i|c_j))_q$.

6: Servers S_0 and S_1 coordinately compute $u_{iq} \leftarrow \log(\Pr(c_j))_q + \sum_{i=1}^{n} w_{iq}$ in local.

7: Both of them compare the results of Step 3(c) for two classes by exploiting the secure comparison protocol. Also, the output classification c_2 as a secret share is computed by each party.

- 两方服务器不泄漏各自均能得到模型
- 不涉及双方数据的分割方式
- 加密部分使用到同态，因此协议是可证明完全

感谢各位专家批评指正

THANK YOU FOR WATCHING