文章编号：2095-2163(2022)06-0084-05 中图分类号：TP181 文献标志码：A

基于联邦学习在机场旅客量的预测

林子谦，安艾芝，樊重俊
（上海理工大学 管理学院，上海 200093）

摘 要：基于航空乘客的出行预测对民航机场的建设和运行有重要影响，但是由于数据量不够的原因，致使旅客预测会存在偏差，若将不同的机场数据集结起来训练会涉及商业机密。因此本文提出基于联邦学习的旅客吞吐量的预测方法，首先对参与训练的机场数据进行预处理与归一化，然后使用逻辑回归模型进行训练，同时使用同态加密来保证数据的隐私安全，最终通过服务器的聚合训练出一个泛化能力强的模型。此外，本文还使用了真实的机场吞吐量数据证明该模型的可行性与有效性。

关键词：联邦学习；同态加密；逻辑回归；机场预测

Exploring the application of federated learning on predicting passenger throughput

LIN Ziqian, An Yizhi, FAN Chongjun
(Business School, University of Shanghai for Science and Technology, Shanghai 200093, China)

【Abstract】 The air passenger travel situation is of great importance to the construction and operation of civil airports. This paper proposes a method for predicting passenger throughput based on federal learning. The airport data involved in the training are preprocessed and normalized, and then trained using a logistic regression model, while homomorphic encryption is used to ensure the privacy of the model, finally a model with strong generalization capability is trained through a server party. In addition, this paper also uses real airport throughput data to demonstrate the feasibility and effectiveness of the model.

【Key words】 federated learning; homomorphic encryption; logistic regression; airport prediction

0 引 言

随着经济的快速发展，人民对生活质量的要求越来越高，旅客对空中交通的需求也越来越大，这对机场造成了很大的影响。机场的客流量是民航机场的重要生产指标[1]，这是实现机场资源有效分配的基础，也是进行机场项目投资决策的重要依据。近年来，许多学者提出了基于机器学习和深度学习的方法来进行交通流预测。其中，高伟等学者[2]基于熵值-BP 神经网络的机场旅客吞吐量预测。王超等学者[3]为了提高精度使用了改进的灰色模型并且仿真上海虹桥的吞吐量。李航等学者[4]使用注意力机制加上 EDM 的模型对旅客量进行预测，通过编码器与解码器的功能对数据做正反向的特征提取，提高预测的精度。但是由于机场的价格政策与军机演练等问题，以单个机场的数据不能很好地进行旅客吞吐量预测。由于民用机场相关数据涉及商业机密以及旅客隐私安全等问题，汇总难度大，数据孤岛现象严重。

谷歌提出了一个新的框架，叫做联邦学习[5]，就有效地解决了隐私安全的问题。在联邦学习模式的训练中，每个客户端参与了模型的训练，都可以将其数据存储到当地，而不需要上传。因此，每个机场使用自己的数据从服务器下载模型进行训练，并将训练好的模型上传至网络，将训练的模型或梯度导入服务器以进行汇总，汇总后的模型或梯度信息将由服务器发送到客户端。

在联邦学习过程中，最经常使用的则是同态加密方案。同态加密是一种可以直接对自己的数据进行加密专用算法，其结果与明文下的计算值是相当吻合的，由于传送能力的不同，通常情况下这些加密方案[6]可以分为：部分同态加密方案、有限同态加密方案、全同态加密方案。其中，部分同态加密，也就是只支持相加或相乘，比如 Paillier 方案，只支持密文之间的加法，而不支持密文间的相乘操作[7]。有限同态加密，则是既支持同态相加、又支持同态相乘，但对计算的次数有一定的局限性，例如 Bonh-Goh-Nissim 方案可以支持无限次的同态相加，但最

作者简介：林子谦(1994-)，男，硕士研究生，主要研究方向：人工智能、工业互联网；安艾芝(1997-)，女，硕士研究生，主要研究方向：人工智能、电子商务；樊重俊(1963-)，男，博士，教授，主要研究方向：人工智能、电子商务。
通讯作者：樊重俊 Email：fan.chongjun@163.com

收稿日期：2021-12-28

多只能支持一次同态相乘；而全同态算法，在不需要自举运算的情况下，可以支持任何数量的同态相加和同态乘法。

本文中，设计了基于联邦学习预测吞吐量的模型，首先各个机场对历史数据进行预测处理，在剔除异常值后对吞吐数据进行归一化，消除量纲的影响；其次，通过使用逻辑回归算法与同态加密相结合，使其机场在不泄露相关数据的情况下，能够共享训练一个预测模型。

1　算法原理

1.1　联邦学习

联邦学习是一种新的分布式机器学习技术，技术目的在于确保信息安全与合法合规的前提下，通过对各参与节点进行高效的机器学习，使其能够更好地进行协同训练，从而获得整体的模型。研究中的基本算法并不限于统计机器学习技术，还包括目前快速发展的深层神经网络。具体的联邦学习的目标函数为：

$$\min_w \sum_{k=1}^{m} p_k F_k(w) \tag{1}$$

其中，m 是参与方数量，$p_k \geqslant 0$ 且满足 $\sum_k p_k = 1$，F_k 是第 k 个参与方的本地优化目标函数。本地目标函数一般通过数据上经验风险损失进行定义，如式(2)所示：

$$F_k(w) = \frac{1}{n_k} \sum_{j_k=1}^{n_k} f_{j_k}(w; x_{j_k}, y_{j_k}) \tag{2}$$

其中，第 k 个参与方的梯度为 $g_k = \nabla F_k(w_t)$，学习率为 η，则第 t 轮迭代得到的新参数为：

$$w_{t+1} \leftarrow w_t - \eta \sum_{k=1}^{K} \frac{n_k}{n} g_k \tag{3}$$

每个参与方的本地更新为：

$$w_{t+1}^k \leftarrow w_t^k - \eta \nabla F_k(w^k) \tag{4}$$

1.2　同态加密的逻辑回归

Paillier 半同态加密算法是由 Paillier[7] 在 1999 年提出的，是一种非对称加密算法的实现，可以处理加密后的数据，计算的结果仍然是加密的，拥有密钥的用户对该加密的结果可以进行解密。

本次的训练模型是逻辑回归模型，则用到的激活函数为 $g(\theta x) = \dfrac{1}{1 + e^{-\theta x}}$，$g(z) \geqslant 0.5$ 时标签为 1，$g(z) < 0.5$ 时标签为 0，其目标函数为：

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left(- y_i \boldsymbol{\theta} \boldsymbol{x}_i + \ln(1 + e^{\boldsymbol{\theta} \boldsymbol{x}_i}) \right) \tag{5}$$

假设参与训练的机场的参数分别为 θ^A，θ^B，则二者聚合在一起的目标函数为：

$$L = \sum_{i=1}^{N} \left(- y_i (u_i^A + u_i^B) + \ln(1 + e^{u_i^A + u_i^B}) \right) \tag{6}$$

则机场 A 的模型与机场 B 的模型参数更新为：

$$\boldsymbol{\theta}^A := \boldsymbol{\theta}^A - \eta \frac{\partial L}{\partial \boldsymbol{\theta}^A}$$

$$\boldsymbol{\theta}^B := \boldsymbol{\theta}^B - \eta \frac{\partial L}{\partial \boldsymbol{\theta}^B} \tag{7}$$

由于 Paillier 加密算法只支持加法同态和标量乘法同态，因此文献[8]使用泰勒展开式的方法进行近似原始对数损失的方法。本文首先将公式对数损失函数 $\log(1 + e^{-z})$ 在 $z = 0$ 处的泰勒展开，表达式为：

$$\log(1 + e^{-y\theta^{\mathrm{T}}x}) \approx \log 2 - \frac{1}{2} \boldsymbol{y} \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x} + \frac{1}{8} (\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x})^2 \tag{8}$$

其中的最后一项由于 $y^2 = 1$，因此直接去掉 y，得到：

$$L = \frac{1}{n} \sum_{i=1}^{n} \left\{ \log 2 - \frac{1}{2} \boldsymbol{y}_i \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x}_i + \frac{1}{8} (\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{x}_i)^2 \right\} \tag{9}$$

因此，对应加密后的梯度为：

$$\left\| \frac{\partial L}{\partial \boldsymbol{\theta}} \right\| = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{4} [[\boldsymbol{\theta}^{\mathrm{T}}]] \boldsymbol{x}_i + \frac{1}{2} [[-1]] \boldsymbol{y}_i \right) \boldsymbol{x}_i \tag{10}$$

2　基于联邦学习的机场模型构建

本文的模型当中，首先对 2 个机场数据进行异常值处理，同时进行归一化处理消除量纲的影响。通过逻辑回归的方法进行训练，并使用同态加密的方法进行隐私保护，联邦服务器端通过聚合两者本地模型，最终训练出适用于当地机场评估的模型。机场联邦学习框架如图 1 所示。

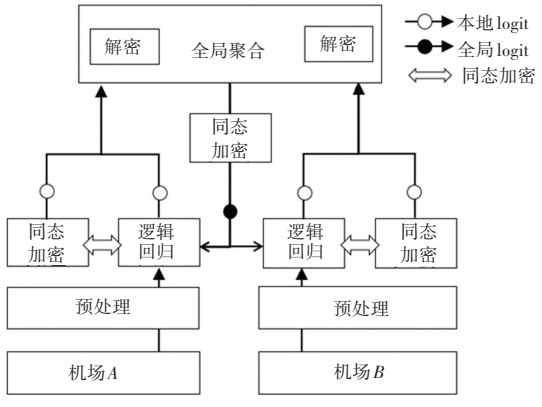

图 1　机场联邦学习框架

Fig. 1　Federated learning framework of the airport

2.1 异常值处理

由于机场会出现极端天气、演练事件的影响,容易出现异常值。因此,当旅客吞吐量数值未能分布在 $(\mu - 3\sigma, \mu + 3\sigma)$ 时,将被判定为异常数据。根据式(11)剔除异常值:

$$\begin{aligned} R_n &= (x_n - \bar{x}) / \mu \\ R_n^* &= (\bar{x} - x_n) / \mu \end{aligned} \tag{11}$$

其中,μ 为已知的总体标准差,$\bar{x}$ 为样本均值,异常值则为 $R_n > R_{0.997}(n)$ 或者 $R_n^* < R_{0.003}^*(n)$。

2.2 数据归一化

假设 X_t 为某机场某一时段的机场旅客吞吐量,则研究推出的归一化公式为:

$$X_t^* = \frac{X_t - X_{\min}}{X_{\max} - X_{\min}} \tag{12}$$

其中,$X_{\min}$ 为数据中旅客量最小值,$X_{\max}$ 为数据中旅客量最大值。

由式(12)可以发现,X_t^* 的取值范围在[0,1]之间,很直观地表现 2 个机场可在同一个区间,保证模型在梯度聚合时不会因为数据不平衡导致偏移的状态。

2.3 同态加密下的训练过程

近年来研究发现通过梯度的传输也会导致数据隐私泄露的风险,因此在传送梯度的过程中进行同态加密是十分重要的。算法步骤具体如下:

Step 1 机场 A 的吞吐量和机场 B 吞吐量分别产生一对公私钥,并将公钥发送到服务器。

Step 2 机场 A 与机场 B 分别计算本地模型的 u_i^A 与 u_i^B,用公钥加密后,将 $[(u_i^A)^2]$ 与 $[(u_i^B)^2]$ 发送给服务器。

Step 3 服务器在接收到 $[(u_i^A)^2]$ 与 $[(u_i^B)^2]$ 后对 2 个机场的模型参数进行解密,同时根据公式(6)计算得到 L_{AB},并且利用公式(8)来求解梯度,再进行聚合。

Step 4 服务器以同样的过程进行加密,传送给机场 A 与机场 B。

Step 5 机场 A 与机场 B 通过解密得到 L_{AB},并根据式(8)得到计算梯度,再使用梯度下降法进行参数更新。此后再次同态加密传送到服务器。

重复 Step1~Step5,直到模型收敛。

2.4 同态加密下的预测过程

当服务器进行询问时,模型部署在机场 A 和机场 B 中,预测过程和上述训练过程类似。对此拟做阐释如下。

Step 1 服务器将预测数据分为 x^A 和 x^B 两部分,利用服务器的公钥加密 x^A 和 x^B 得到 $[x^A]_C$ 和 $[x^B]_C$,分别发送给机场 A 与机场 B 进行计算。

Step 2 在机场 A 和机场 B 上分别计算得到 $[u^A]_C = \boldsymbol{\theta}^A [\boldsymbol{x}^A]_C$ 和 $[u^B]_C = \boldsymbol{\theta}^B [\boldsymbol{x}^B]_C$,并发送给服务器。

Step 3 服务器解密后得到 u^A 和 u^B,$u = u^A + u^B$,计算得到最终输出结果 $\frac{1}{1 + e^{-u}}$。

3 实验及结果分析

3.1 实验环境与数据集准备

根据本文航空旅客吞吐量进行预测,由于选取的实验数据集完整不存在缺失值、但是存在异常值的问题,所以对数据集进行归一化与异常数据剔除的处理。本文所选数据集为某市 2 个机场从 2017 年 1 月 1 日到 2017 年 12 月 31 日的每天旅客吞吐量的数据集。

实验环境配置为 Ubuntu21.0 操作系统,Intel Core i5-8300H,8 GB 内存,Python3.6 编程语言,Pytorch3.6 框架,显卡型号 RTX 3080。其中,模型的学习率为 0.01,动量为 0.9,迭代次数都为 50,且取 80%为训练值,剩下的 20%为预测值。

3.2 评价指数

本文中,分析模型的实验结果,采用平均绝对百分比误差(*MPAE*)作为模型评价函数,以此评价模型的预测效果,具体公式如下:

$$MPAE = \frac{\sum_{i=1}^{n} \frac{|y_i - y_i^*|}{y_i}}{n} \times 100\% \tag{13}$$

3.3 对比实验分析

机场 A 真实数据与逻辑回归预测对比如图 2 所示,机场 A 真实数据与联邦学习预测对比如图 3 所示,机场 B 真实数据与逻辑回归预测对比如图 4 所示,机场 B 真实数据与联邦学习预测对比如图 5 所示。由图 2 可知,很明显在单个机场进行训练的过程中,由于存在一定的噪声的原因,导致模型在拟合的过程中,出现预测过高、或者波动大的问题,图 2 与图 4 都出现了跳动比较大或者预测值过高的情况。而联邦学习后的训练值不论是机场 A、还是机场 B 的拟合度都特别高,特别是机场 A 的拟合不论是时间的波动、还是预测的误差值都十分接近真实。

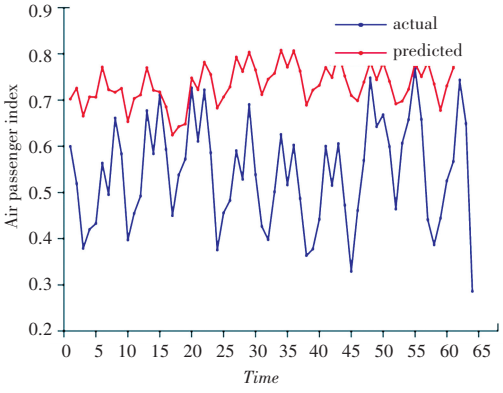

图 2　机场 *A* 真实数据与逻辑回归预测对比图

Fig. 2　Comparison between real data and predictions of logistic regression for airport *A*

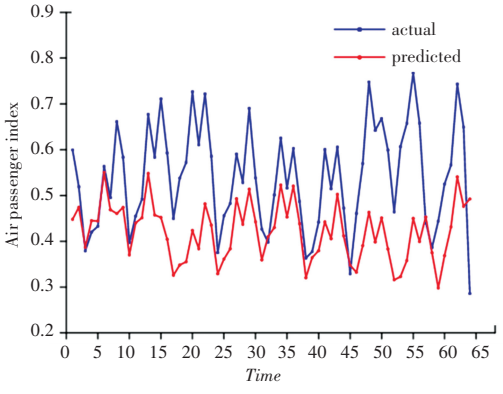

图 3　机场 *A* 真实数据与联邦学习预测对比图

Fig. 3　Comparison between real data and predictions of federated learning for airport *A*

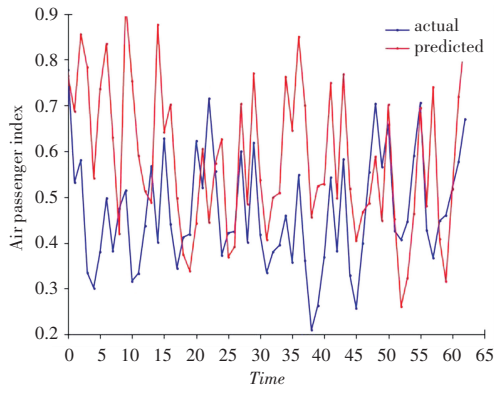

图 4　机场 *B* 真实数据与逻辑回归预测对比图

Fig. 4　Comparison between real data and predictions of logistic regression for airport *B*

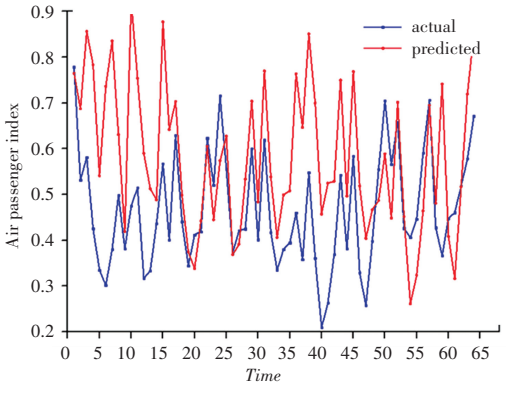

图 5　机场 *B* 真实数据与联邦学习预测对比图

Fig. 5　Comparison of real data and predictions of federated learning for airport *B*

逻辑回归与联邦学习回归 *MPAE* 的对比结果见表 1。由逻辑回归与联邦学习两种方法 *MPAE* 值的对比显示，若直接使用逻辑回归进行预测，分别得到的准确值为 0.220 4 与 0.191 4，而通过将二者聚合之后的模型有着明显的提升的效果。由此也可以证明本方案的可行性与有效性。

表 1　逻辑回归与联邦学习回归 *MPAE* 的对比表

Tab. 1　Comparison table of *MAPE* between logistic regression and federated learning regression

	机场 *A*	机场 *B*
逻辑回归	0.220 4	0.191 4
联邦学习+逻辑回归	0.154 7	0.149 2

4　结束语

通过对旅客吞吐量进行预测，机场可选择在客流量较少的时段内进行维修或其它建设活动，使得对机场运行和管理造成的影响降到最低；同时，在客流高峰期，适当配置与之相适应的内外资源，例如：地面公共交通资源、地勤人员、安保人员等，以保证机场的畅通和生产的安全。本文中，应用在联邦学习的方法使得 2 个机场在发生涉及隐私的情况时会训练一个共同的模型，并且在传递模型的过程中，用到了同态加密，由此保障了数据不会被泄露。本方案是首次将联邦学习应用在 2 个不同的机场旅客量预测当中，未来将会把机场图像分割与其他传感器的数据相结合并扩展到深度学习模型当中。

参考文献

[1] 熊红林，朱人杰，冀和，等. 基于 MI-SVR 模型的航空旅客出行指数预测方法研究[J]. 控制与决策，2021，36(07)：1619-1626.

[2] 高伟，殷小曼. 基于熵值-BP 神经网络的机场旅客吞吐量预测[J]. 计算机仿真，2021，38(10)：64-67.

[3] 王超，李冬，张余. 动态改进灰色模型在机场吞吐量预测中的应用[J]. 计算机仿真，2019，36(12)：74-77，83.

（下转第 96 页）