单细胞基因数据分析赛题

1、 场景介绍

外周血单个核细胞（Peripheral Blood Mononuclear Cell，PBMC）是外周血中具有单个核的细胞，这些细胞包括淋巴细胞（如，T 细胞、B 细胞、NK 细胞）和单核细胞。PBMC 单细胞 RNA 测序常应用于免疫学领域的科研研究，如自体免疫性疾病, 传染病, 恶性血液病的研究，及疫苗开发等细胞类型鉴定是此类研究的主要任务，快速有效的识别单细胞类型可协助科研人员全面了解研究对象的细胞群微环境。由于受试者的单细胞基因表达数据为敏感数据，需采用适当的数据保护措施。本赛题将采用 TEE 技术实现隐私数据资产的高性能隐私保护计算，依据单细胞基因表达信息建立模型预测细胞类型。

赛题任务

- **计算参与方：**甲方（数据源方）、乙方（TEE 算力提供方）共两方。
- **数据输入：**
    - 为参赛队伍提供含有细胞类型标注的 PBMC 基因表达数据，供解决方案开发使用。数据特征见附录一。
    - 分类目标：10 类，具体见附录一。
- **目标输出：**外周血单个核细胞分类模型和验证结果。
- **技术要求：**使用 Intel SGX V2 TEE（可信执行环境）。
- **安全性要求：**128bits 安全性
    - 可信计算环境（TEE）：参赛队伍请参考 SGX Remote Attestation 协议完成动态密钥交换。本赛题暂不考虑侧信道攻击问题。
- **安全假设：**
    - 可信计算环境（TEE）部分需要支持**恶意模型假设**
- **隐私保护目标：**
    - TEE 算力提供方不能获得数据方的明文原始数据、计算过程中的明文统计信息和明文的计算结果。

- **评测原则：**
  - 解决方案测试时，会使用另一套未公开的数据集进行测试。
  - 结果的正确性（训练好的模型在验证数据集上的准确率）
  - 计算所需的时间（包括预处理、training 时间、testing 时间等）。
  - 计算过程中的内存使用量峰值。
  - 计算过程中的网络通信总流量。
- **评测环境：**
  - TEE 节点：支持 SGX V2 的 TEE 服务器，至少 16GB EPC 内存，可以使用多线程。
  - 暂定 KVM 虚拟机，暂定虚拟机将配备 4 个 CPU 内核 128 GB 内存和 500 GB 评估存储空间。
- **解决方案提交要求：**
  - Training
    - 甲方输入：训练数据集，按格式，带有 label，加密后发给乙方
    - 乙方输入：加密的甲方输入
    - 乙方输出：加密的 model，发给甲方
    - 甲方输出：解密后的 model
  - Testing
    - 甲方输入：model；测试数据集，按格式，不带 label，加密后发给乙方
    - 乙方输入：加密的用户输入
    - 乙方输出：加密的预测 label，发给甲方
    - 甲方输出：解密后的预测 label
  - 详细的 Readme
  - 用于 evaluation 的 script
  - 解决方案提交方式、日志和接口格式详见附录二

附录一：数据格式

数据集文件格式：

训练数据集：train.csv - 格式如下：

- 第一列：label: cell type (ground truth)，包含 10 种分类如下：

  CD14+ Monocyte
  CD19+ B
  CD34+
  CD4+ T Helper2
  CD4+/CD25 T Reg
  CD4+/CD45RA+/CD25- Naive T
  CD4+/CD45RO+ Memory
  CD56+ NK
  CD8+ Cytotoxic T
  CD8+/CD45RA+ Naive Cytotoxic

- 其余列 每一行是个样本的数据；每一列包含 gene 的表达量

label	RP11_34P13_7	RP11_34P13_8	AL627309_1	AP006222_2	RP5_857K21_4
{'CD34+' }	0	0	0	0	0
{'CD19+ B' }	0	0	0	0	0
{'CD4+/CD25 T Reg'}	0	0	0	0	0
{'CD4+ T Helper2' }	0	0	0	0	0
{'CD34+' }	0	0	0	0	0
{'CD34+' }	0	0	0	0	0
{'CD14+ Monocyte' }	0	0	0	0	0
{'CD4+ T Helper2' }	0	0	0	0	0
{'CD4+ T Helper2' }	0	0	0	0	0
{'CD34+' }	0	0	0	0	0

验证数据集：test.csv - 格式如下：

- 每一行是个样本的数据；每一列包含 gene 的表达量，**不包含 label**

RP11_34P13_7	RP11_34P13_8	AL627309_1	AP006222_2	RP5_857K21_4
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

附录二：输出格式

为了更好的评估所构建的模型性能，对模型的输出进行规范，具体要求如下：

- 请将模型的预测结果保存为 **csv** 格式，文件共 **1** 列，包含预测细胞类型(cell type)
- 在写入结果文件时，请保留表头信息，即 cell type
- 保存文件的格式示例如下：

Cell type
CD14+ Monocyte
CD56+ NK
CD4+/CD25 T Reg
…
CD8+ Cytotoxic T

- 具体的提交方式等待后续通知。