(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115438370 A
(43)申请公布日 2022.12.06

(21)申请号 202210940253.1

(22)申请日 2022.08.05

(71)申请人 北京富算科技有限公司
地址 100020 北京市朝阳区东三环中路9号19层2201

(72)发明人 陈立峰 卞阳 尤志强 王兆凯

(74)专利代理机构 上海弼兴律师事务所 31283
专利代理师 罗朗 李静

(51)Int.Cl.
G06F 21/62(2013.01)
G06K 9/62(2022.01)
G06N 20/00(2019.01)

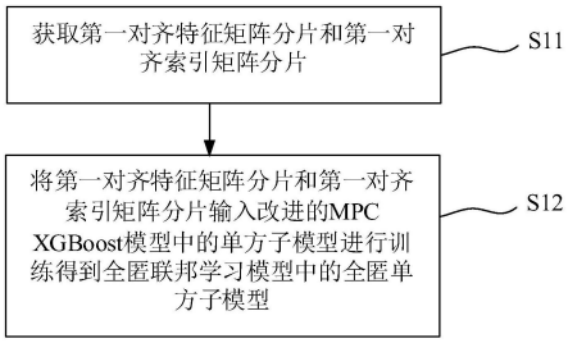
权利要求书3页 说明书12页 附图10页

(54)发明名称

全匿联邦学习模型的训练方法、设备和存储介质

(57)摘要

本发明提供一种全匿联邦学习模型的训练方法、设备和存储介质，方法包括：获取第一对齐特征矩阵分片和第一对齐索引矩阵分片；将第一对齐特征矩阵分片和第一对齐索引矩阵分片输入改进的MPC XGBoost模型中的单方子模型训练得到全匿联邦学习模型中的全匿单方子模型。本发明中求交信息以分片形式存在，在全匿框架中中间数据也是碎片形式，各参与方的对齐特征矩阵分片包括了碎片密态化的共同用户信息并且矩阵的行高尽量小；对齐特征矩阵分片对齐各参与方的共同用户的数据，并在密态计算中使得共同用户的数据相加为不变，非共同用户的数据相加为零，密态分片也避免了参与方识别出非交集用户的置零数据，从而使得整个流程的安全性大大提升，不会暴露任何数据。



CN 115438370 A

1.一种全匿联邦学习模型的训练方法,其特征在于,两个参与方包括第一参与方和第二参与方,所述训练方法应用于所述第一参与方,所述训练方法包括:

获取第一对齐特征矩阵分片和第一对齐索引矩阵分片;

将所述第一对齐特征矩阵分片和所述第一对齐索引矩阵分片输入改进的MPC XGBoost模型中的单方子模型进行训练得到全匿联邦学习模型中的全匿单方子模型;

其中,多个所述单方子模型联合进行训练,所述第一对齐特征矩阵分片包括了碎片密态化的共同用户的信息,并且矩阵的行高与所述两个参与方中数据最少的样本量相同;所述第一对齐特征矩阵分片使得所述第一参与方中的共同用户对应的特征分片数据与所述第二参与方中的共同用户对应的特征分片数据对齐,并在密态计算中使得所述共同用户对应的特征分片数据相加为不变,使得非共同用户对应的特征分片数据相加为零。

2.如权利要求1所述的全匿联邦学习模型的训练方法,其特征在于,所述获取第一对齐特征矩阵分片和第一对齐索引矩阵分片,包括:

获取所述第一参与方的第一样本数据;

对所述第一样本数据进行分片得到第一特征矩阵分片和第二特征矩阵分片,并将所述第二特征矩阵分片发送至所述两个参与方中的第二参与方;

接收所述第二参与方发送的第三特征矩阵分片;

获得第一求交结果分片;

比较所述第一特征矩阵分片和所述第三特征矩阵分片的行高,以最小的行高作为对齐特征矩阵的行高,基于所述第一求交结果分片和所述第三特征矩阵分片通过MPC协议的乘法得到第三中间特征矩阵分片;

对所述第一求交结果分片按行求和得到所述第一对齐索引矩阵分片;

基于所述第一特征矩阵分片和所述第一对齐索引矩阵分片通过点乘得到第一中间特征矩阵分片;

将所述第一中间特征矩阵分片和所述第三中间特征矩阵分片进行拼接得到所述第一对齐特征矩阵分片。

3.如权利要求2所述的全匿联邦学习模型的训练方法,其特征在于,所述获得第一求交结果分片,包括:

获取第一求交数据集合,其中,所述第一求交数据集合包括所述第一参与方的用户数据;

将所述第一求交数据集合进行分片得到第一分片和第二分片,其中,所述第一分片和所述第二分片均保留所述第一求交数据集合中每一条数据的一部分信息;

将所述第二分片发送至第二参与方,并接收所述第二参与方发送的第三分片,其中,所述第三分片是第二求交数据集合的一个分片,所述第二求交数据集合包括所述第二参与方的用户数据;

基于所述第一分片和所述第三分片通过MPC协议的比较得到第一求交结果分片,其中,所述第一求交结果分片以碎片信息的形式指示所述第一参与方和所述第二参与方的交集用户。

4.如权利要求3所述的全匿联邦学习模型的训练方法,其特征在于,所述基于所述第一分片和所述第三分片通过MPC协议的比较得到第一求交结果分片,包括:

将所述第一分片和所述第三分片中每一位置上的数值进行两两比较是否相等得到所述第一求交结果分片;其中,所述第一分片为矩阵;

其中,若相等则将所述第一求交结果分片的对应位置设置为一;若不相等则将所述第一求交结果分片的对应位置设置为零。

5.如权利要求2所述的全匿联邦学习模型的训练方法,其特征在于,在训练过程中,所述单方子模型执行以下步骤:

获取随机种子、第一预测值分片和第一标签分片;

执行构建树的迭代直至满足迭代停止条件。

6.如权利要求5所述的全匿联邦学习模型的训练方法,其特征在于,在所述构建树的迭代中,所述单方子模型执行以下步骤:

对所述第一对齐特征矩阵分片进行按行样本采样得到第一采样特征矩阵分片;

基于所述第一预测值分片和所述第一标签分页通过MPC协议的密态计算得到第一初始一阶导数分片和第一初始二阶导数分片;

基于所述第一初始一阶导数分片、所述第一初始二阶导数分片和所述第一对齐特征矩阵分片通过MPC协议的点乘得到第一最终一阶导数分片和第一最终二阶导数分片;

基于所述第一采样特征矩阵分片通过MPC协议的密态计算得到第一最大值转置分片和第一最小值转置分片;

获取第一辅助计算矩阵分片;

基于所述第一最大值转置分片、所述第一最小值转置分片和所述第一辅助计算矩阵分片划分出B个桶,并确定每个桶的边界通过MPC协议的密态计算得到第一分桶边界分片;

遍历所述第一采样特征矩阵分片中每列特征数据,提取每列特征数据的列特征分片,基于所述列特征分片和所述第一分桶边界分片进行MPC协议的范围比较,得到矩阵内容进行B2A转换生成特征分桶矩阵分片;

拼接所有特征分桶矩阵分片得到最终特征分桶稀疏矩阵分片;

基于所述第一最终一阶导数分片、所述第一最终二阶导数分片和所述最终特征分桶稀疏矩阵分片通过MPC协议的矩阵乘法得到第一直方图分片;

基于所述第一最终一阶导数分片和所述第一最终二阶导数分片通过MPC协议的密态计算为达到停止分裂条件的节点赋值,得到当前停止分裂的节点的第一节点值分片;

更新树结构;

使用更新后的树来预测原始数据,更新所述第一预测值分片。

7.如权利要求6所述的全匿联邦学习模型的训练方法,其特征在于,在所述第一参与方为标签方时,所述第一样本数据包括y标签;

所述获取随机种子、第一预测值分片和第一标签分片,包括:

生成所述随机种子,初始化预测值,将所述预测值和所述y标签分别进行分片得到第一预测值分片、第二预测值分片、第一标签分片和第二标签分片;

将所述随机种子、所述第二预测值分片和所述第二标签分片发送至所述第二参与方;

所述获取第一辅助计算矩阵分片,包括:

构建辅助计算矩阵,并将所述辅助计算矩阵进行分片得到第一辅助计算矩阵分片和第二辅助计算矩阵分片,将所述第二辅助计算矩阵分片发送至所述第二参与方;

在所述基于所述第一最终一阶导数分片和所述第一最终二阶导数分片通过MPC协议的密态计算为达到停止分裂条件的节点赋值的步骤之前，所述单方子模型还执行以下步骤：

将所述第一直方图分片发送至所述第二参与方，并接收所述第二参与方发送的第二直方图分片；其中，所述第二直方图分片为所述第二参与方计算得到的直方图分片；

根据所述第一直方图分片和所述第二直方图分片得到直方图明文；

对于所述直方图明文根据最优分割公式进行最优分割点计算，得到待分裂节点的最优分割点；

在所述更新树结构的步骤之前，所述单方子模型还执行以下步骤：

将所述最优分割点的信息发送至所述第二参与方；

在所述更新树结构的步骤之后，所述单方子模型还执行以下步骤：

将下一级的节点信息发送至所述第二参与方。

8.如权利要求6所述的全匿联邦学习模型的训练方法，其特征在于，在所述第一参与方为非标签方时，所述获取随机种子、第一预测值分片和第一标签分片，包括：

接收所述第二参与方发送的所述随机种子、所述第一预测值分片和所述第一标签分片；其中，所述第一预测值分片为所述第二参与方将初始化后的预测值进行分片得到的多个预测值分片中任一个，所述第一标签分片为所述第二参与方将y标签进行分片得到的多个标签分片中任一个；

所述获取第一辅助计算矩阵分片，包括：

接收所述第二参与方发送的所述第一辅助计算矩阵分片；其中，所述第一辅助计算矩阵分片为所述第二参与方将构建的辅助计算矩阵进行分片得到的多个辅助计算矩阵分片中任一个；

在所述基于所述第一最终一阶导数分片和所述第一最终二阶导数分片通过MPC协议的密态计算为达到停止分裂条件的节点赋值的步骤之前，所述单方子模型还执行以下步骤：

将所述第一直方图分片发送至所述第二参与方；

在所述更新树结构的步骤之前，所述单方子模型还执行以下步骤：

接收所述第二参与方发送的最优分割点的信息；

在所述更新树结构的步骤之后，所述单方子模型还执行以下步骤：

接收所述第二参与方发送的下一级的节点信息。

9.一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，其特征在于，所述处理器执行计算机程序时实现权利要求1至8中任一项所述的全匿联邦学习模型的训练方法。

10.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现权利要求1至8中任一项所述的全匿联邦学习模型的训练方法。

全匿联邦学习模型的训练方法、设备和存储介质

技术领域

[0001] 本发明涉及计算机技术领域，尤其涉及一种全匿联邦学习模型的训练方法、设备和存储介质。

背景技术

[0002] 随着信息化在人们生活方方面面普及，越来越多的数据在各种场景下生成，如商场的消费记录，打车的行程记录，看病的就诊记录等等。这些数据是个人隐私的一部分，极其敏感也极其有价值，为了使这些数据不外泄许多单位会选择不对外开放仅内部使用，这就产生了一个个“数据孤岛”。然而信息技术的发展依赖数据合理地开放使用，将多种单位数据的结合可以辅助许多产业的发展，如：医院患者病例数据和药品厂商的生产数据结合可以促进新药的研发，多个银行的流水数据结合使用可以辅助联合风控，保险公司保单数据和医院患者病例数据结合使用可以进行高效理赔等等。

[0003] 联邦学习本质上是一种分布式机器学习框架，其做到了在保障数据隐私安全及合法合规的基础上，实现数据共享，共同建模。有了联邦学习技术的广泛使用，可以打破“数据孤岛”尴尬处境，让技术使用方在“数据不出门”的情况下和其他使用联合建模(即多方建模)，共同创造数据价值。多方建模首先要进行安全求交(PSI)。安全求交可以简单理解为有N方，每方都有各自业务的用户(医院的患者，银行的储户，手机应用的注册用户等等)，使用安全求交技术获得N方的共同用户，且求交完后不会给对方暴露非交集用户的任何信息，求交的匹配信息可以是手机号码，身份证等可表达唯一身份的信息。

[0004] 但是安全求交得到的交集用户在一些敏感场景下会不适用，主要原因在于最终结果为明文，会相互暴露交集用户，如银行a有用户集合A，医院b有用户集合B，A和B数据集和进行安全求交后会得到A∩B＝C，这里数据集合C是双方明文知晓的，即银行a知道C数据集合里面的储户都在医院b里有就诊记录，医院b也知道C数据集合里面的患者在银行a里有账户。这种情况对一些单位是不可接受的，安全性差，无法保护用户隐私，限制了纵向联邦学习在高安全性要求的机构或者场景下的使用。另外，在纵向联邦学习中，如果不进行交集共享，传统的算法设计无法有效进行联合建模。

发明内容

[0005] 本发明要解决的技术问题是为了克服现有技术中安全求交会相互暴露交集用户，安全性差，无法保护用户隐私的缺陷，提供一种全匿联邦学习模型的训练方法、设备和存储介质。

[0006] 本发明是通过下述技术方案来解决上述技术问题：

[0007] 本发明提供一种全匿联邦学习模型的训练方法，两个参与方包括第一参与方和第二参与方，所述训练方法应用于所述第一参与方，所述训练方法包括：

[0008] 获取第一对齐特征矩阵分片和第一对齐索引矩阵分片；

[0009] 将所述第一对齐特征矩阵分片和所述第一对齐索引矩阵分片输入改进的MPC(多

方安全计算)XGBoost(一种联邦学习模型)模型中的单方子模型进行训练得到全匿联邦学习模型中的全匿单方子模型；

[0010] 其中，多个所述单方子模型联合进行训练，所述第一对齐特征矩阵分片包括了碎片密态化的共同用户的信息，并且矩阵的行高与所述两个参与方中数据最少的样本量相同；所述第一对齐特征矩阵分片使得所述第一参与方中的共同用户对应的特征分片数据与所述第二参与方中的共同用户对应的特征分片数据对齐，并在密态计算中使得所述共同用户对应的特征分片数据相加为不变，使得非共同用户对应的特征分片数据相加为零。

[0011] 较佳地，所述获取第一对齐特征矩阵分片和第一对齐索引矩阵分片，包括：

[0012] 获取所述第一参与方的第一样本数据；

[0013] 对所述第一样本数据进行分片得到第一特征矩阵分片和第二特征矩阵分片，并将所述第二特征矩阵分片发送至所述两个参与方中的第二参与方；

[0014] 接收所述第二参与方发送的第三特征矩阵分片；

[0015] 获得第一求交结果分片；

[0016] 比较所述第一特征矩阵分片和所述第三特征矩阵分片的行高，以最小的行高作为对齐特征矩阵的行高，基于所述第一求交结果分片和所述第三特征矩阵分片通过MPC协议的乘法得到第三中间特征矩阵分片；

[0017] 对所述第一求交结果分片按行求和得到所述第一对齐索引矩阵分片；

[0018] 基于所述第一特征矩阵分片和所述第一对齐索引矩阵分片通过点乘得到第一中间特征矩阵分片；

[0019] 将所述第一中间特征矩阵分片和所述第三中间特征矩阵分片进行拼接得到所述第一对齐特征矩阵分片。

[0020] 较佳地，所述获得第一求交结果分片，包括：

[0021] 获取第一求交数据集合，其中，所述第一求交数据集合包括所述第一参与方的用户数据；

[0022] 将所述第一求交数据集合进行分片得到第一分片和第二分片，其中，所述第一分片和所述第二分片均保留所述第一求交数据集合中每一条数据的一部分信息；

[0023] 将所述第二分片发送至第二参与方，并接收所述第二参与方发送的第三分片，其中，所述第三分片是第二求交数据集合的一个分片，所述第二求交数据集合包括所述第二参与方的用户数据；

[0024] 基于所述第一分片和所述第三分片通过MPC协议的比较得到第一求交结果分片，其中，所述第一求交结果分片以碎片信息的形式指示所述第一参与方和所述第二参与方的交集用户。

[0025] 较佳地，所述基于所述第一分片和所述第三分片通过MPC协议的比较得到第一求交结果分片，包括：

[0026] 将所述第一分片和所述第三分片中每一位置上的数值进行两两比较是否相等得到所述第一求交结果分片；其中，所述第一分片为矩阵；

[0027] 其中，若相等则将所述第一求交结果分片的对应位置设置为一；若不相等则将所述第一求交结果分片的对应位置设置为零。

[0028] 较佳地，在训练过程中，所述单方子模型执行以下步骤：

[0029] 获取随机种子、第一预测值分片和第一标签分片；

[0030] 执行构建树的迭代直至满足迭代停止条件。

[0031] 较佳地，在所述构建树的迭代中，所述单方子模型执行以下步骤：

[0032] 对所述第一对齐特征矩阵分片进行按行样本采样得到第一采样特征矩阵分片；

[0033] 基于所述第一预测值分片和所述第一标签分页通过MPC协议的密态计算得到第一初始一阶导数分片和第一初始二阶导数分片；

[0034] 基于所述第一初始一阶导数分片、所述第一初始二阶导数分片和所述第一对齐特征矩阵分片通过MPC协议的点乘得到第一最终一阶导数分片和第一最终二阶导数分片；

[0035] 基于所述第一采样特征矩阵分片通过MPC协议的密态计算得到第一最大值转置分片和第一最小值转置分片；

[0036] 获取第一辅助计算矩阵分片；

[0037] 基于所述第一最大值转置分片、所述第一最小值转置分片和所述第一辅助计算矩阵分片划分出B个桶，并确定每个桶的边界通过MPC协议的密态计算得到第一分桶边界分片；

[0038] 遍历所述第一采样特征矩阵分片中每列特征数据，提取每列特征数据的列特征分片，基于所述列特征分片和所述第一分桶边界分片进行MPC协议的范围比较，得到矩阵内容进行B2A(一种矩阵转换方法，即布尔转数值，True转1，False转0)转换生成特征分桶矩阵分片；

[0039] 拼接所有特征分桶矩阵分片得到最终特征分桶稀疏矩阵分片；

[0040] 基于所述第一最终一阶导数分片、所述第一最终二阶导数分片和所述最终特征分桶稀疏矩阵分片通过MPC协议的矩阵乘法得到第一直方图分片；

[0041] 基于所述第一最终一阶导数分片和所述第一最终二阶导数分片通过MPC协议的密态计算为达到停止分裂条件的节点赋值，得到当前停止分裂的节点的第一节点值分片；

[0042] 更新树结构；

[0043] 使用更新后的树来预测原始数据，更新所述第一预测值分片。

[0044] 较佳地，在所述第一参与方为标签方时，所述第一样本数据包括y标签；

[0045] 所述获取随机种子、第一预测值分片和第一标签分片，包括：

[0046] 生成所述随机种子，初始化预测值，将所述预测值和所述y标签分别进行分片得到第一预测值分片、第二预测值分片、第一标签分片和第二标签分片；

[0047] 将所述随机种子、所述第二预测值分片和所述第二标签分片发送至所述第二参与方；

[0048] 所述获取第一辅助计算矩阵分片，包括：

[0049] 构建辅助计算矩阵，并将所述辅助计算矩阵进行分片得到第一辅助计算矩阵分片和第二辅助计算矩阵分片，将所述第二辅助计算矩阵分片发送至所述第二参与方；

[0050] 在所述基于所述第一最终一阶导数分片和所述第一最终二阶导数分片通过MPC协议的密态计算为达到停止分裂条件的节点赋值的步骤之前，所述单方子模型还执行以下步骤：

[0051] 将所述第一直方图分片发送至所述第二参与方，并接收所述第二参与方发送的第二直方图分片；其中，所述第二直方图分片为所述第二参与方计算得到的直方图分片；

[0052] 根据所述第一直方图分片和所述第二直方图分片得到直方图明文；

[0053] 对于所述直方图明文根据最优分割公式进行最优分割点计算，得到待分裂节点的最优分割点；

[0054] 在所述更新树结构的步骤之前，所述单方子模型还执行以下步骤：

[0055] 将所述最优分割点的信息发送至所述第二参与方；

[0056] 在所述更新树结构的步骤之后，所述单方子模型还执行以下步骤：

[0057] 将下一级的节点信息发送至所述第二参与方。

[0058] 较佳地，在所述第一参与方为非标签方时，所述获取随机种子、第一预测值分片和第一标签分片，包括：

[0059] 接收所述第二参与方发送的所述随机种子、所述第一预测值分片和所述第一标签分片；其中，所述第一预测值分片为所述第二参与方将初始化后的预测值进行分片得到的多个预测值分片中任一个，所述第一标签分片为所述第二参与方将y标签进行分片得到的多个标签分片中任一个；

[0060] 所述获取第一辅助计算矩阵分片，包括：

[0061] 接收所述第二参与方发送的所述第一辅助计算矩阵分片；其中，所述第一辅助计算矩阵分片为所述第二参与方将构建的辅助计算矩阵进行分片得到的多个辅助计算矩阵分片中任一个；

[0062] 在所述基于所述第一最终一阶导数分片和所述第一最终二阶导数分片通过MPC协议的密态计算为达到停止分裂条件的节点赋值的步骤之前，所述单方子模型还执行以下步骤：

[0063] 将所述第一直方图分片发送至所述第二参与方；

[0064] 在所述更新树结构的步骤之前，所述单方子模型还执行以下步骤：

[0065] 接收所述第二参与方发送的最优分割点的信息；

[0066] 在所述更新树结构的步骤之后，所述单方子模型还执行以下步骤：

[0067] 接收所述第二参与方发送的下一级的节点信息。

[0068] 本发明还提供一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行计算机程序时实现前述的全匿联邦学习模型的训练方法。

[0069] 本发明还提供一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现前述的全匿联邦学习模型的训练方法。

[0070] 本发明的积极进步效果在于：以MPC秘密共享为基础，各参与方的求交信息都是以分片的形式存在，在全匿框架中进行数值计算时，中间数据也是以碎片形式计算，每一参与方的对齐特征矩阵分片包括了碎片密态化的共同用户的信息，并且矩阵的行高与参与方中数据最少的样本量相同；每一参与方的对齐特征矩阵分片使得本参与方(即第一参与方)中的共同用户对应的特征分片数据与其他参与方(即第二参与方)中的共同用户对应的特征分片数据对齐，并在密态计算中使得共同用户对应的特征分片数据相加为不变，使得非共同用户对应的特征分片数据相加为零，密态分片也避免了参与方识别出非交集用户的置零数据，从而使得整个流程的安全性大大提升，不会暴露任何数据。

附图说明

[0071] 图1为本发明的实施例1的全匿联邦学习模型的训练方法的流程图。

[0072] 图2A为本发明的实施例1的全匿联邦学习模型的训练方法中的步骤S11的一具体实施方式的流程图。

[0073] 图2B为本发明的实施例1的全匿联邦学习模型的训练方法中的得到对齐特征矩阵分片以及对齐索引矩阵分片示例的流程图。

[0074] 图3A为本发明的实施例1的全匿联邦学习模型的训练方法中的步骤S114的一具体实施方式的流程图。

[0075] 图3B为本发明的实施例1的全匿联邦学习模型的训练方法中的全匿安全求交示例的流程图。

[0076] 图4为本发明的实施例1的全匿联邦学习模型的训练方法的一具体实施方式的流程图。

[0077] 图5A为本发明的实施例1的全匿联邦学习模型的训练方法中的步骤S22的一具体实施方式的流程图。

[0078] 图5B1为本发明的实施例1的全匿联邦学习模型的训练方法中的训练示例的流程图的上半部分。

[0079] 图5B2为本发明的实施例1的全匿联邦学习模型的训练方法中的训练示例的流程图的下半部分。

[0080] 图6为本发明的实施例1的全匿联邦学习模型的训练方法中的步骤S21的一具体实施方式的流程图。

[0081] 图7为本发明的实施例1的全匿联邦学习模型的训练方法中的步骤S2205的一具体实施方式的流程图。

[0082] 图8为本发明的实施例1的全匿联邦学习模型的训练方法中的计算最优分割点的一具体实施方式的流程图。

[0083] 图9为本发明的实施例2的电子设备的结构示意图。

具体实施方式

[0084] 下面通过实施例的方式进一步说明本发明，但并不因此将本发明限制在所述的实施例范围之中。

[0085] 需要说明的是，在本发明的描述中，术语“第一”、“第二”仅用于描述目的，对同类型数据进行区分，而不能理解为指示或暗示相对重要性。

[0086] 实施例1

[0087] 本实施例提供一种全匿联邦学习模型的训练方法，两个参与方包括第一参与方和第二参与方。训练方法应用于第一参与方，参照图1，训练方法包括：

[0088] S11、获取第一对齐特征矩阵分片和第一对齐索引矩阵分片。

[0089] S12、将第一对齐特征矩阵分片和第一对齐索引矩阵分片输入改进的MPC XGBoost模型中的单方子模型进行训练得到全匿联邦学习模型中的全匿单方子模型。

[0090] 其中，多个单方子模型联合进行训练，第一对齐特征矩阵分片包括了碎片密态化的共同用户的信息，并且矩阵的行高与两个参与方中数据最少的样本量相同。第一对齐特

征矩阵分片使得第一参与方中的共同用户对应的特征分片数据与第二参与方中的共同用户对应的特征分片数据对齐,并在密态计算中使得共同用户对应的特征分片数据相加为不变,使得非共同用户对应的特征分片数据相加为零。

[0091] 例如,两个参与方分别为Guest和Host,Guest有一个对齐特征矩阵分片<cf1>以及一个对齐索引矩阵分片<s_c1>,Host有一个对齐特征矩阵分片<cf2>以及一个对齐索引矩阵分片<s_c2>。

[0092] 对于Guest,第一参与方是Guest,第二参与方是Host,第一对齐特征矩阵分片是<cf1>,第一对齐索引矩阵分片是<s_c1>。对于Host,第一参与方是Host,第二参与方是Guest,第一对齐特征矩阵分片是<cf2>,第一对齐索引矩阵分片是<s_c2>。

[0093] 对齐特征矩阵分片<cf1>和<cf2>包括了碎片密态化的共同用户的信息,并且矩阵的行高与Guest和Host中数据最少的样本量相同。对齐索引矩阵分片<s_c1>和<s_c2>使得Guest和Host的对齐特征矩阵分片对齐并在密态计算中使得共同用户对应的特征分片数据相加为不变,非共同用户对应的特征分片数据相加为零。密态分片也避免了参与方识别出非交集用户的置零数据。这样设置的矩阵的行高确保了用户数据在训练过程中不暴露。训练全过程都是分片状态的计算,双方都不知道实际哪些数据起了训练作用,没有起作用的非交集用户的用户数据在<cf1>和<cf2>里都会置零。非交集用户的数据置零了就不会对模型训练产生影响。非交集用户的置零数据不会暴露的原因是,交集用户的碎片态数据在Guest方可能是-6,在Host方是6,两者相加是零,就无法和非交集用户的数据区分开了。

[0094] 本实施例以MPC秘密共享为基础,各参与方的求交信息都是以分片的形式存在,在全匿框架中进行数值计算时,中间数据也是以碎片形式计算,每一参与方的对齐特征矩阵分片包括了碎片密态化的共同用户的信息,并且矩阵的行高与参与方中数据最少的样本量相同;每一参与方的对齐特征矩阵分片使得本参与方(即第一参与方)中的共同用户对应的特征分片数据与其他参与方(即第二参与方)中的共同用户对应的特征分片数据对齐,并在密态计算中使得共同用户对应的特征分片数据相加为不变,使得非共同用户对应的特征分片数据相加为零,密态分片也避免了参与方识别出非交集用户的置零数据,从而使得整个流程的安全性大大提升,不会暴露任何数据。

[0095] 具体实施时,参照图2A,步骤S11包括:

[0096] S111、获取第一参与方的第一样本数据。

[0097] S112、对第一样本数据进行分片得到第一特征矩阵分片和第二特征矩阵分片,并将第二特征矩阵分片发送至两个参与方中的第二参与方。

[0098] S113、接收第二参与方发送的第三特征矩阵分片。

[0099] S114、获得第一求交结果分片。

[0100] S115、比较第一特征矩阵分片和第三特征矩阵分片的行高,以最小的行高作为对齐特征矩阵的行高,基于第一求交结果分片和第三特征矩阵分片通过MPC协议的乘法得到第三中间特征矩阵分片。

[0101] S116、对第一求交结果分片按行求和得到第一对齐索引矩阵分片。

[0102] S117、基于第一特征矩阵分片和第一对齐索引矩阵分片通过点乘得到第一中间特征矩阵分片。其中,通过点乘抹除第一参与方的未对齐数据。

[0103] S118、将第一中间特征矩阵分片和第三中间特征矩阵分片进行拼接得到第一对齐

特征矩阵分片。

[0104] 其中,图2B示出了Guest和Host交互得到各自的对齐特征矩阵分片<cf1>和<cf2>以及对齐索引矩阵分片<s_c1>和<s_c2>的流程图。

[0105] 在上述示例中,Guest的样本数据为FA,FA为特征矩阵,shape(矩阵形状)为(m,r),m为Guest的样本数量,r为Guest的特征数量,Guest的特征矩阵分片为<fa1>和<fa2>,Guest的求交结果分片为<c1>,Guest的中间特征矩阵分片为<cfa1>和<cfb1>,Guest的对齐索引矩阵分片为<s_c1>,Guest的对齐特征矩阵分片为<cf1>。Host的样本数据为FB,FB为特征矩阵,shape(矩阵形状)为(n,t),n为Host的样本数量,t为Host的特征数量,Host的特征矩阵分片为<fb1>和<fb2>,Host的求交结果分片为<c2>,Host的中间特征矩阵分片为<cfa2>和<cfb2>,Host的对齐索引矩阵分片为<s_c2>,Host的对齐特征矩阵分片为<cf2>。

[0106] 对于Guest,第一参与方是Guest,第二参与方是Host,第一样本数据是FA,第一特征矩阵分片是<fa1>,第二特征矩阵分片是<fa2>,第三特征矩阵分片是<fb1>,第一求交结果分片是<c1>,第一对齐索引矩阵分片是<s_c1>,第一中间特征矩阵分片是<cfa1>,第二中间特征矩阵分片是<cfa2>,第三中间特征矩阵分片是<cfb1>,第四中间特征矩阵分片是<cfb2>,第一对齐特征矩阵分片是<cf1>。

[0107] 对于Host,第一参与方是Host,第二参与方是Guest,第一样本数据是FB,第一特征矩阵分片是<fb1>,第二特征矩阵分片是<fb2>,第三特征矩阵分片是<fa1>,第一求交结果分片是<c2>,第一对齐索引矩阵分片是<s_c2>,第一中间特征矩阵分片是<cfb1>,第二中间特征矩阵分片是<cfb2>,第三中间特征矩阵分片是<cfa1>,第四中间特征矩阵分片是<cfa2>,第一对齐特征矩阵分片是<cf2>。

[0108] 图中,单箭头虚线表示传递分片数据,双箭头虚线表示在MPC协议的算子计算(例如乘法和点乘)中传递中间计算数据,MPC协议的算子计算为现有技术,此处不再赘述。

[0109] 该示例中以Guest的样本数量小于Host的样本数量举例。如果Host的对齐样本数量远小于Guest,可以按列求和重新计算对齐索引矩阵分片,这样可以缩小聚合后的特征矩阵大小。

[0110] 具体实施时,参照图3A,步骤S114包括:

[0111] S1141、获取第一求交数据集合,其中,第一求交数据集合包括第一参与方的用户数据。

[0112] S1142、将第一求交数据集合进行分片得到第一分片和第二分片,其中,第一分片和第二分片均保留第一求交数据集合中每一条数据的一部分信息。

[0113] S1143、将第二分片发送至第二参与方,并接收第二参与方发送的第三分片,其中,第三分片是第二求交数据集合的一个分片,第二求交数据集合包括第二参与方的用户数据。

[0114] S1144、基于第一分片和第三分片通过MPC协议的比较得到第一求交结果分片,其中,第一求交结果分片以碎片信息的形式指示第一参与方和第二参与方的交集用户。

[0115] 其中,图3B示出了Guest和Host全匿安全求交的流程图。

[0116] 在上述示例中,Guest的求交数据集合为A,A分片后得到两个分片<a1>和<a2>,求交结果分片为<c1>。

[0117] Host的求交数据集合为B,B分片后得到两个分片<b1>和<b2>,求交结果分片为<c2

>。

[0118] 对于Guest和Host,其各自需要的数据(例如第一求交数据集合等)的确定方法与前述内容类似,此处不再赘述。

[0119] 图中,单箭头虚线表示传递分片数据,双箭头虚线表示在MPC协议的算子计算(例如乘法和点乘)中传递中间计算数据,MPC协议的算子计算为现有技术,此处不再赘述。

[0120] 具体实施时,步骤S1144包括:

[0121] 将第一分片和第三分片中每一位置上的数值进行两两比较是否相等得到第一求交结果分片。其中,第一分片为矩阵。

[0122] 其中,若相等则将第一求交结果分片的对应位置设置为一。若不相等则将第一求交结果分片的对应位置设置为零。

[0123] 具体实施时,参照图4,在训练过程中,单方子模型执行以下步骤:

[0124] S21、获取随机种子、第一预测值分片和第一标签分片。

[0125] S22、执行构建树的迭代直至满足迭代停止条件。

[0126] 其中,随机种子是为了在接下来创建随机采样时能使得两个参与方的数据进行对齐。

[0127] 具体实施时,参照图5A,在构建树的迭代中,单方子模型执行以下步骤:

[0128] S2201、对第一对齐特征矩阵分片进行按行样本采样得到第一采样特征矩阵分片。

[0129] S2202、基于第一预测值分片和第一标签分页通过MPC协议的密态计算得到第一初始一阶导数分片和第一初始二阶导数分片。

[0130] S2203、基于第一初始一阶导数分片、第一初始二阶导数分片和第一对齐特征矩阵分片通过MPC协议的点乘得到第一最终一阶导数分片和第一最终二阶导数分片。

[0131] S2204、基于第一采样特征矩阵分片通过MPC协议的密态计算得到第一最大值转置分片和第一最小值转置分片。

[0132] S2205、获取第一辅助计算矩阵分片。

[0133] S2206、基于第一最大值转置分片、第一最小值转置分片和第一辅助计算矩阵分片划分出B个桶,并确定每个桶的边界通过MPC协议的密态计算得到第一分桶边界分片。

[0134] S2207、遍历第一采样特征矩阵分片中每列特征数据,提取每列特征数据的列特征分片,基于列特征分片和第一分桶边界分片进行MPC协议的范围比较,得到矩阵内容进行B2A转换生成特征分桶矩阵分片。

[0135] S2208、拼接所有特征分桶矩阵分片得到最终特征分桶稀疏矩阵分片。

[0136] S2209、基于第一最终一阶导数分片、第一最终二阶导数分片和最终特征分桶稀疏矩阵分片通过MPC协议的矩阵乘法得到第一直方图分片。

[0137] S2210、基于第一最终一阶导数分片和第一最终二阶导数分片通过MPC协议的密态计算为达到停止分裂条件的节点赋值,得到当前停止分裂的节点的第一节点值分片。

[0138] S2211、更新树结构。

[0139] S2212、使用更新后的树来预测原始数据,更新第一预测值分片。

[0140] 其中,对于对齐特征矩阵分片按行采样是样本采样,按列采样是特征采样,通过采样改变矩阵的形状。

[0141] 预测值的明文更新会有可能被推测出来交集内容,导致交集样本暴露,因此全匿

情况下使用预测值分片以避免暴露交集样本。使用标签分片是为了通过MPC协议的数值计算得到一阶梯度分片和二阶梯度分片。

[0142] 图5B1和5B2示出了Guest和Host联合进行全匿联邦学习模型训练的示例的流程图，图5B1为流程图的上半部分，图5B2为流程图的下半部分，图在粗虚线为上半部分和下半部分的分界线。图中，单箭头虚线表示传递分片数据，双箭头虚线表示在MPC协议的算子计算(例如乘法和点乘)中传递中间计算数据，MPC协议的算子计算为现有技术，此处不再赘述。

[0143] Guest的预测值分片为<p1>，标签分片为<y1>，采样特征矩阵分片为<cf_sam1>，初始一阶导数分片为<g1′>和初始二阶导数分片为<h1′>，最终一阶导数分片为<g1>和最终二阶导数分片为<h1>，最大值转置分片为<f_max1>，最小值转置分片为<f_min1>，辅助计算矩阵分片为<B_range1>，分桶边界分片为<bin_split1>，列特征分片为<cf_i1>，特征分桶矩阵分片为<bin_i1>，最终特征分桶稀疏矩阵分片为<b1>，直方图分片为<h_histo1>，节点值分片为<w1i>。

[0144] Host的预测值分片为<p2>，标签分片为<y2>，采样特征矩阵分片为<cf_sam2>，初始一阶导数分片为<g2′>和初始二阶导数分片为<h2′>，最终一阶导数分片为<g2>和最终二阶导数分片为<h2>，最大值转置分片为<f_max2>，最小值转置分片为<f_min2>，辅助计算矩阵分片为<B_range2>，分桶边界分片为<bin_split2>，列特征分片为<cf_i2>，特征分桶矩阵分片为<bin_i2>，最终特征分桶稀疏矩阵分片为<b2>，直方图分片为<h_histo2>，节点值分片为<w2i>。

[0145] 对于Guest和Host，其各自需要的数据(例如第一标签分片等)的确定方法与前述内容类似，此处不再赘述。

[0146] 示例中，使用的计算公式具体如下：

[0147] $$<g1'>=1\oslash(1\oplus e^{-<p1>})\ominus<y1>$$

[0148] $$<g2'>=1\oslash(1\oplus e^{-<p2>})\ominus<y2>$$

[0149] $$<h1'>=e^{-<p1>}\oslash((1\oplus e^{-<p1>})\otimes(1\oplus e^{-<p1>}))$$

[0150] $$<h2'>=e^{-<p2>}\oslash((1\oplus e^{-<p2>})\otimes(1\oplus e^{-<p2>}))$$

[0151] $$<bin\_split1>=(<f\_min1>\oplus(<f\_max1>\ominus<f\_min1>)\oslash B)\otimes<B\_range1>$$

[0152] $$<bin\_split2>=(<f\_min2>\oplus(<f\_max2>\ominus<f\_min2>)\oslash B)\otimes<B\_range2>$$

[0153] $$Gain=\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right]-\gamma$$

[0154] $$<w1_i>=\sum_{j\in I}<g1_j>\oslash(\sum_{j\in I}<h1_j>\oplus\lambda)$$

[0155] $$<w2_i>=\sum_{j\in I}<g2_j>\oslash(\sum_{j\in I}<h2_j>\oplus\lambda)$$

[0156] 上面公式均为MPC密态计算过程，⊗为MPC乘法(MPC协议的乘法)，∅为MPC除法

(MPC协议的除法)，⊕为MPC加法(MPC协议的加法)，⊖为MPC减法(MPC协议的减法)。I表示停止分裂节点下的样本集合，i表示停止分裂节点的序号，j表示I集合中的每个样本，e为常数。

[0157] 通过最大值转置分片和最小值转置分片来构建分片状态的直方图分桶边界。同时，为了构建分片状态的直方图分桶边界，还需要构建一个辅助计算矩阵分片，内容为[0,1,2,3...,B]，B为分桶数。

[0158] 依次求每个特征的值在不同边界范围之间的布尔结果，这个过程也是mpc比较，最后将布尔结果矩阵转为算数结果<bin_i>，shape为(m,B)，m为Guest方所有样本数量，B为分桶数量，<bin_i>记录的原始值明文是一个稀疏的0/1矩阵。

[0159] 将所有特征的<bin_i>矩阵进行拼接得到最终特征分桶稀疏矩阵分片<b>，shape为(m,r*B+t*B)，m为Guest方所有样本数量，B为分桶数量，r为Guest的特征数量，t为Host的特征数量。

[0160] Gain表示增益计算，G_R表示分裂右侧样本一阶梯度和，G_L表示分裂左侧样本一阶梯度和，H_R表示分裂右侧二阶梯度和，H_L表示分裂左侧阶梯度和，λ和γ均为计算系数以防止除数为零。Gain的计算为现有技术，此处不再赘述。

[0161] 具体实施时，在第一参与方为标签方时，第一样本数据包括y标签。

[0162] 参照图6，步骤S21包括：

[0163] S211、生成随机种子，初始化预测值，将预测值和y标签分别进行分片得到第一预测值分片、第二预测值分片、第一标签分片和第二标签分片。

[0164] S212、将随机种子、第二预测值分片和第二标签分片发送至第二参与方。

[0165] 参照图7，步骤S2205包括：

[0166] S22051、构建辅助计算矩阵，并将辅助计算矩阵进行分片得到第一辅助计算矩阵分片和第二辅助计算矩阵分片，将第二辅助计算矩阵分片发送至第二参与方。

[0167] 参照图8，在步骤S2210之前，单方子模型还执行以下步骤：

[0168] S31、将第一直方图分片发送至第二参与方，并接收第二参与方发送的第二直方图分片。其中，第二直方图分片为第二参与方计算得到的直方图分片。

[0169] S32、根据第一直方图分片和第二直方图分片得到直方图明文。

[0170] S33、对于直方图明文根据最优分割公式进行最优分割点计算，得到待分裂节点的最优分割点。

[0171] 其中，通过双方的直方图分片得到直方图明文，进而能够计算得到待分裂节点的最优分割点，进一步构建树。

[0172] 在步骤S2211之前，单方子模型还执行以下步骤：

[0173] 将最优分割点的信息发送至第二参与方。其中，最优分割点的信息即为节点分裂信息。

[0174] 在步骤S2211之后，单方子模型还执行以下步骤：

[0175] 将下一级的节点信息发送至第二参与方。

[0176] 其中，标签方需要提供y标签数据。

[0177] 本实施例中，以MPC秘密共享为基础，标签方需要提供y标签数据，各参与方的求交信息都是以分片的形式存在，在全匿框架中进行数值计算时，中间数据也是以碎片形式计

算,每一参与方的对齐特征矩阵分片包括了碎片密态化的共同用户的信息,并且矩阵的行高与参与方中数据最少的样本量相同;每一参与方的对齐特征矩阵分片使得本参与方(即第一参与方)中的共同用户对应的特征分片数据与其他参与方(即第二参与方)中的共同用户对应的特征分片数据对齐,并在密态计算中使得共同用户对应的特征分片数据相加为不变,使得非共同用户对应的特征分片数据相加为零,密态分片也避免了参与方识别出非交集用户的置零数据,从而使得整个流程的安全性大大提升,不会暴露任何数据。

[0178] 具体实施时,在第一参与方为非标签方时,参照图6,步骤S21包括:

[0179] S213、接收第二参与方发送的随机种子、第一预测值分片和第一标签分片。其中,第一预测值分片为第二参与方将初始化后的预测值进行分片得到的多个预测值分片中任一个,第一标签分片为第二参与方将y标签进行分片得到的多个标签分片中任一个。

[0180] 参照图7,步骤S2205包括:

[0181] S22052、接收第二参与方发送的第一辅助计算矩阵分片。其中,第一辅助计算矩阵分片为第二参与方将构建的辅助计算矩阵进行分片得到的多个辅助计算矩阵分片中任一个。

[0182] 在步骤S2210之前,单方子模型还执行以下步骤:

[0183] 将第一直方图分片发送至第二参与方。

[0184] 在步骤S2211之前,单方子模型还执行以下步骤:

[0185] 接收第二参与方发送的最优分割点的信息。其中,最优分割点的信息即为节点分裂信息。

[0186] 在步骤S2211之后,单方子模型还执行以下步骤:

[0187] 接收第二参与方发送的下一级的节点信息。

[0188] 其中,非标签方不需要提供y标签数据。

[0189] 本实施例中,以MPC秘密共享为基础,标签方需要提供y标签数据并将标签分片分享给其他参与方,各参与方的求交信息都是以分片的形式存在,在全匿框架中进行数值计算时,中间数据也是以碎片形式计算,每一参与方的对齐特征矩阵分片包括了碎片密态化的共同用户的信息,并且矩阵的行高与参与方中数据最少的样本量相同;每一参与方的对齐特征矩阵分片使得本参与方(即第一参与方)中的共同用户对应的特征分片数据与其他参与方(即第二参与方)中的共同用户对应的特征分片数据对齐,并在密态计算中使得共同用户对应的特征分片数据相加为不变,使得非共同用户对应的特征分片数据相加为零,密态分片也避免了参与方识别出非交集用户的置零数据,从而使得整个流程的安全性大大提升,不会暴露任何数据。

[0190] 实施例2

[0191] 图9为本发明实施例2提供的一种电子设备的结构示意图。所述电子设备包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序,所述处理器执行所述程序时实现实施例1中的全匿联邦学习模型的训练方法。图9显示的电子设备30仅仅是一个示例,不应对本发明实施例的功能和使用范围带来任何限制。

[0192] 电子设备30可以以通用计算设备的形式表现,例如其可以为服务器设备。电子设备30的组件可以包括但不限于:上述至少一个处理器31、上述至少一个存储器32、连接不同系统组件(包括存储器32和处理器31)的总线33。

[0193] 总线33包括数据总线、地址总线和控制总线。

[0194] 存储器32可以包括易失性存储器，例如随机存取存储器(RAM)321和/或高速缓存存储器322，还可以进一步包括只读存储器(ROM)323。

[0195] 存储器32还可以包括具有一组(至少一个)程序模块324的程序/实用工具325，这样的程序模块324包括但不限于：操作系统、一个或者多个应用程序、其它程序模块以及程序数据，这些示例中的每一个或某种组合中可能包括网络环境的实现。

[0196] 处理器31通过运行存储在存储器32中的计算机程序，从而执行各种功能应用以及数据处理，例如本发明实施例1中的全匿联邦学习模型的训练方法。

[0197] 电子设备30也可以与一个或多个外部设备34(例如按键、指向设备等)通信。这种通信可以通过输入/输出(I/O)接口35进行。并且，模型生成的电子设备30还可以通过网络适配器36与一个或者多个网络(例如局域网(LAN)，广域网(WAN)和/或公共网络，例如因特网)通信。如图所示，网络适配器36通过总线33与模型生成的电子设备30的其它模块通信。应当明白，尽管图中未示出，可以结合模型生成的电子设备30使用其它硬件和/或软件模块，包括但不限于：微代码、设备驱动器、冗余处理器、外部磁盘驱动阵列、RAID(磁盘阵列)系统、磁带驱动器以及数据备份存储系统等。

[0198] 应当注意，尽管在上文详细描述中提及了电子设备的若干模块/模块或子模块/模块，但是这种划分仅仅是示例性的并非强制性的。实际上，根据本发明的实施方式，上文描述的两个或更多模块/模块的特征和功能可以在一个模块/模块中具体化；反之，上文描述的一个模块/模块的特征和功能可以进一步划分为由多个模块/模块来具体化。

[0199] 实施例3

[0200] 本实施例提供了一种计算机可读存储介质，其上存储有计算机程序，所述程序被处理器执行时实现实施例1中的全匿联邦学习模型的训练方法。

[0201] 其中，可读存储介质可以采用的更具体可以包括但不限于：便携式盘、硬盘、随机存取存储器、只读存储器、可擦拭可编程只读存储器、光存储器件、磁存储器件或上述的任意合适的组合。

[0202] 在可能的实施方式中，本发明还可以实现为一种程序产品的形式，其包括程序代码，当所述程序产品在终端设备上运行时，所述程序代码用于使所述终端设备执行实现实施例1中的全匿联邦学习模型的训练方法。

[0203] 其中，可以以一种或多种程序设计语言的任意组合来编写用于执行本发明的程序代码，所述程序代码可以完全地在用户设备上执行、部分地在用户设备上执行、作为一个独立的软件包执行、部分在用户设备上部分在远程设备上执行或完全在远程设备上执行。

[0204] 虽然以上描述了本发明的具体实施方式，但是本领域的技术人员应当理解，这仅是举例说明，本发明的保护范围是由所附权利要求书限定的。本领域的技术人员在不背离本发明的原理和实质的前提下，可以对这些实施方式做出多种变更或修改，但这些变更和修改均落入本发明的保护范围。

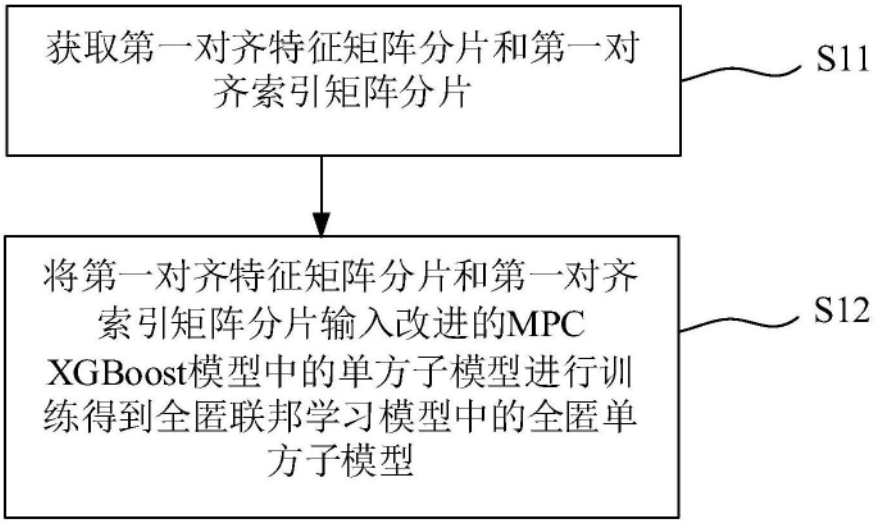

获取第一对齐特征矩阵分片和第一对齐索引矩阵分片
S11
将第一对齐特征矩阵分片和第一对齐索引矩阵分片输入改进的MPC XGBoost模型中的单方子模型进行训练得到全匿联邦学习模型中的全匿单方子模型
S12

图1

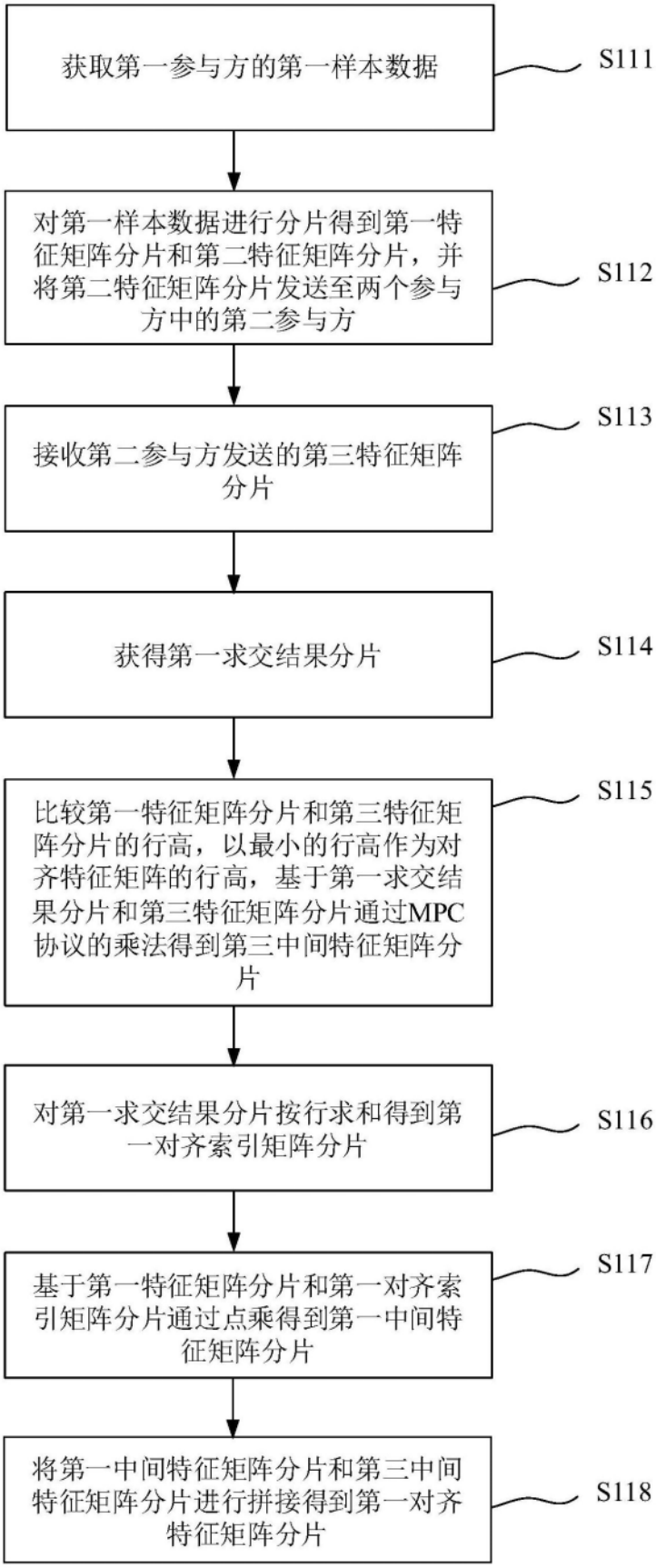

获取第一参与方的第一样本数据
S111
对第一样本数据进行分片得到第一特征矩阵分片和第二特征矩阵分片，并将第二特征矩阵分片发送至两个参与方中的第二参与方
S112
接收第二参与方发送的第三特征矩阵分片
S113
获得第一求交结果分片
S114
比较第一特征矩阵分片和第三特征矩阵分片的行高，以最小的行高作为对齐特征矩阵的行高，基于第一求交结果分片和第三特征矩阵分片通过MPC协议的乘法得到第三中间特征矩阵分片
S115
对第一求交结果分片按行求和得到第一对齐索引矩阵分片
S116
基于第一特征矩阵分片和第一对齐索引矩阵分片通过点乘得到第一中间特征矩阵分片
S117
将第一中间特征矩阵分片和第三中间特征矩阵分片进行拼接得到第一对齐特征矩阵分片
S118

图2A

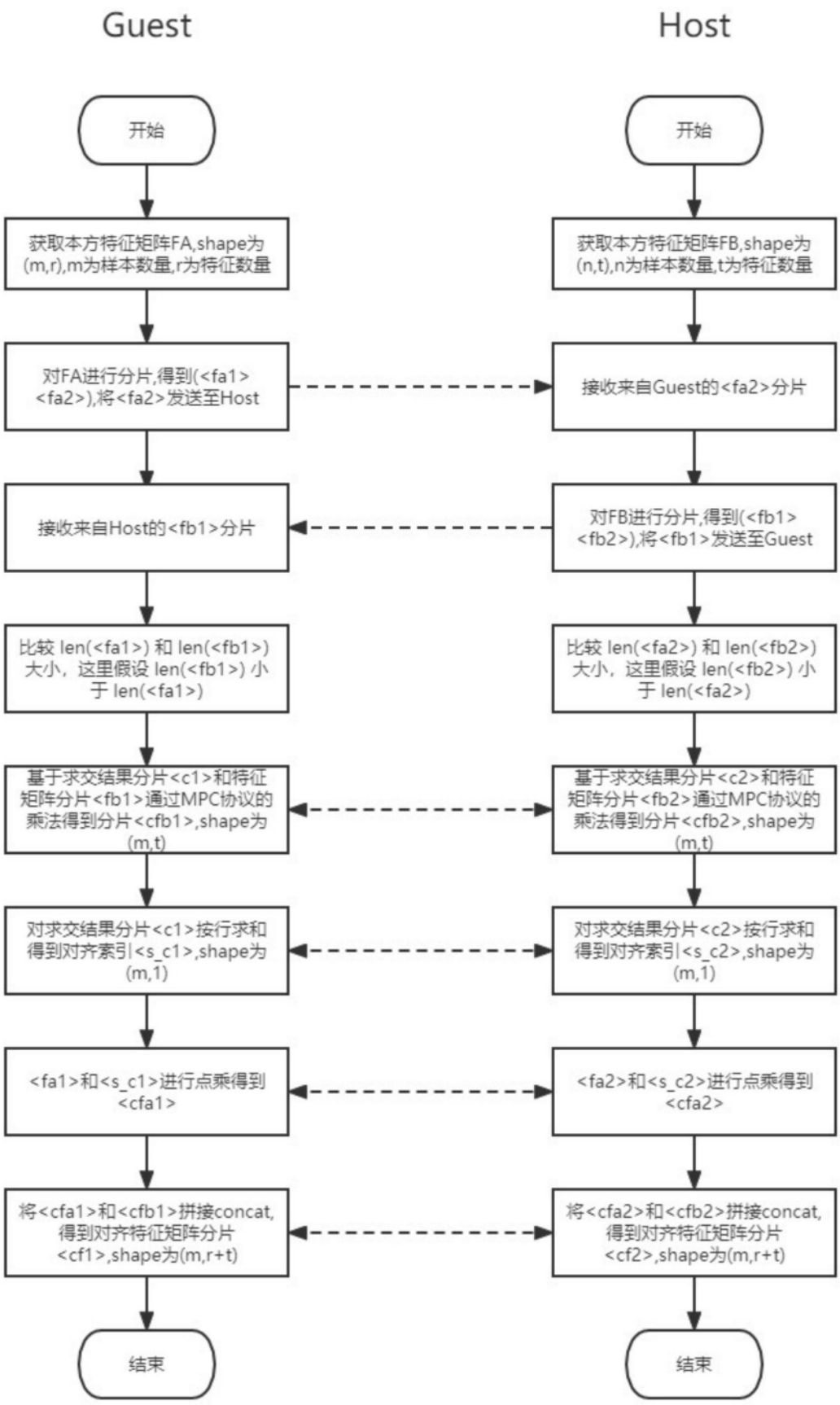

Guest
Host
开始
开始
获取本方特征矩阵FA,shape为(m,r),m为样本数量,r为特征数量
获取本方特征矩阵FB,shape为(n,t),n为样本数量,t为特征数量
对FA进行分片,得到(<fa1><fa2>),将<fa2>发送至Host
接收来自Guest的<fa2>分片
接收来自Host的<fb1>分片
对FB进行分片,得到(<fb1><fb2>),将<fb1>发送至Guest
比较 len(<fa1>) 和 len(<fb1>) 大小，这里假设 len(<fb1>) 小于 len(<fa1>)
比较 len(<fa2>) 和 len(<fb2>) 大小，这里假设 len(<fb2>) 小于 len(<fa2>)
基于求交结果分片<c1>和特征矩阵分片<fb1>通过MPC协议的乘法得到分片<cfb1>,shape为(m,t)
基于求交结果分片<c2>和特征矩阵分片<fb2>通过MPC协议的乘法得到分片<cfb2>,shape为(m,t)
对求交结果分片<c1>按行求和得到对齐索引<s_c1>,shape为(m,1)
对求交结果分片<c2>按行求和得到对齐索引<s_c2>,shape为(m,1)
<fa1>和<s_c1>进行点乘得到<cfa1>
<fa2>和<s_c2>进行点乘得到<cfa2>
将<cfa1>和<cfb1>拼接concat,得到对齐特征矩阵分片<cf1>,shape为(m,r+t)
将<cfa2>和<cfb2>拼接concat,得到对齐特征矩阵分片<cf2>,shape为(m,r+t)
结束
结束

图2B

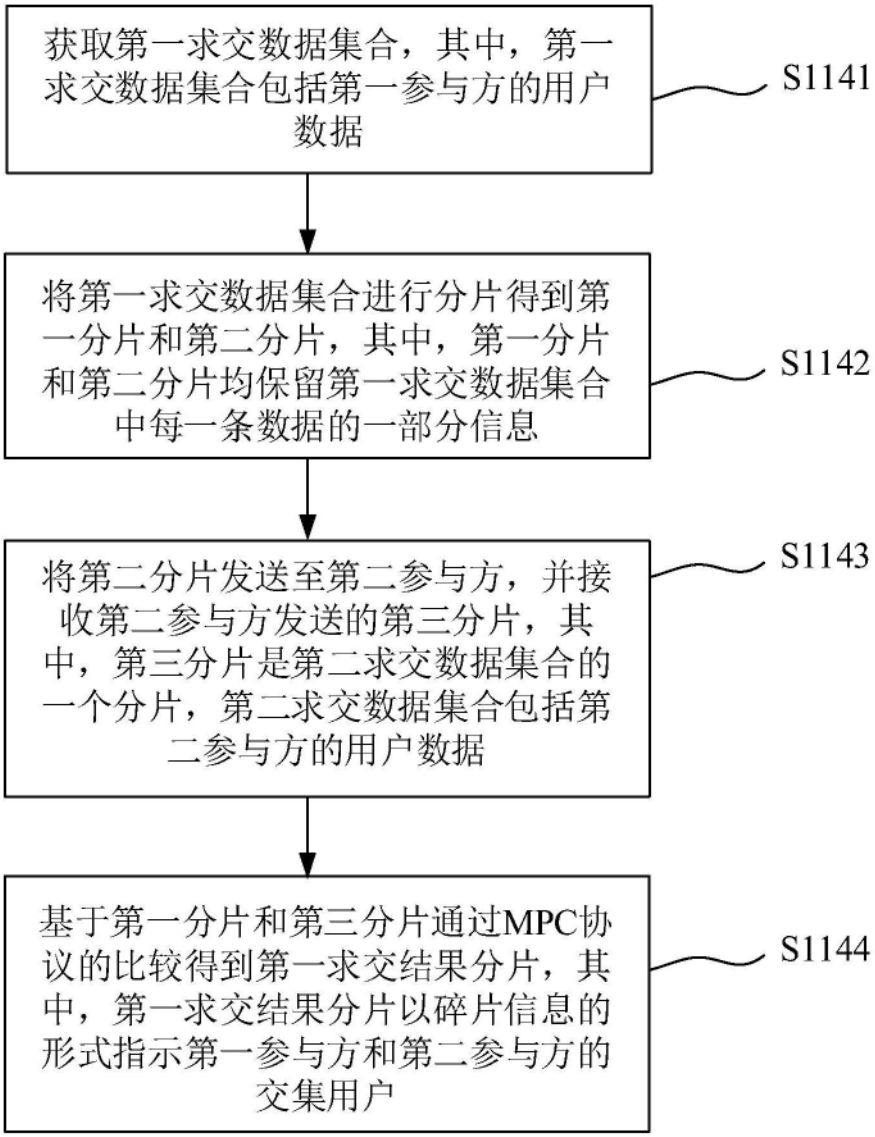

获取第一求交数据集合，其中，第一求交数据集合包括第一参与方的用户数据
S1141
将第一求交数据集合进行分片得到第一分片和第二分片，其中，第一分片和第二分片均保留第一求交数据集合中每一条数据的一部分信息
S1142
将第二分片发送至第二参与方，并接收第二参与方发送的第三分片，其中，第三分片是第二求交数据集合的一个分片，第二求交数据集合包括第二参与方的用户数据
S1143
基于第一分片和第三分片通过MPC协议的比较得到第一求交结果分片，其中，第一求交结果分片以碎片信息的形式指示第一参与方和第二参与方的交集用户
S1144

图3A

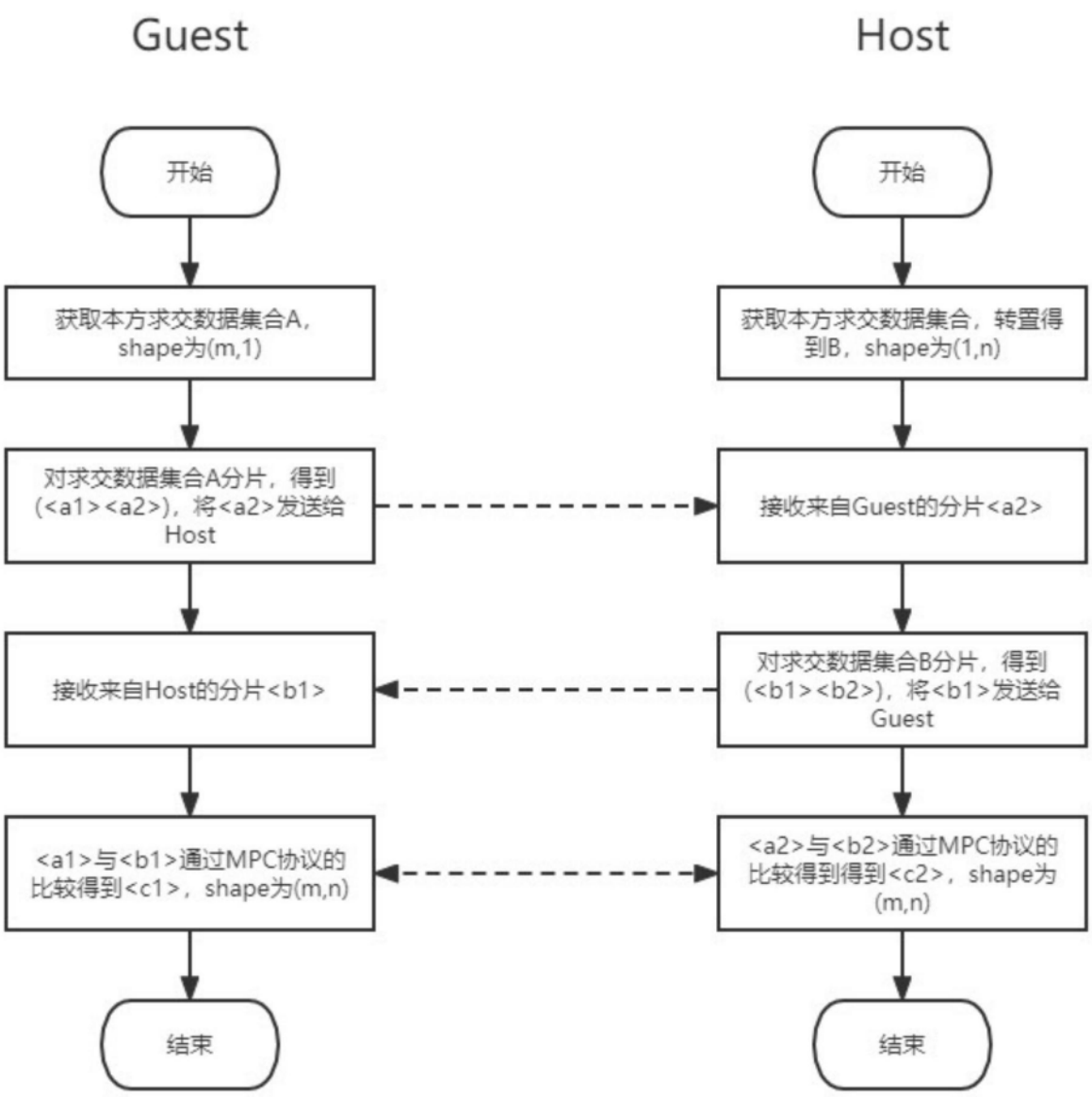

Guest
Host
开始
开始
获取本方求交数据集合A，shape为(m,1)
获取本方求交数据集合，转置得到B，shape为(1,n)
对求交数据集合A分片，得到(<a1><a2>)，将<a2>发送给Host
接收来自Guest的分片<a2>
接收来自Host的分片<b1>
对求交数据集合B分片，得到(<b1><b2>)，将<b1>发送给Guest
<a1>与<b1>通过MPC协议的比较得到<c1>，shape为(m,n)
<a2>与<b2>通过MPC协议的比较得到得到<c2>，shape为(m,n)
结束
结束

图3B

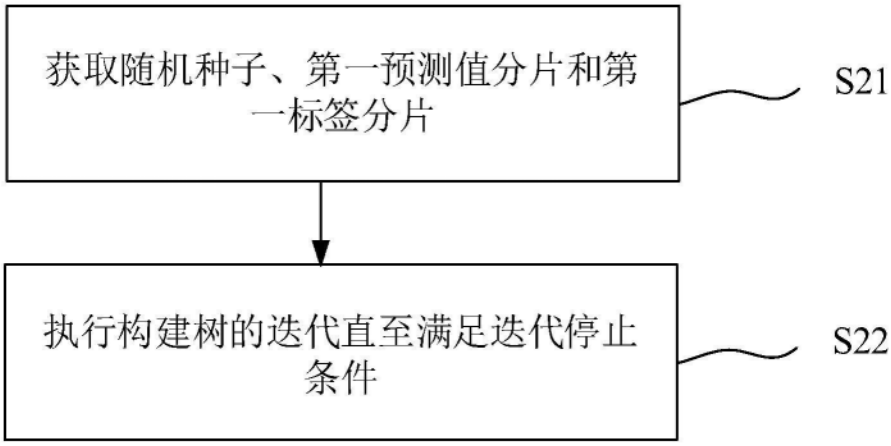

获取随机种子、第一预测值分片和第一标签分片
S21
执行构建树的迭代直至满足迭代停止条件
S22

图4

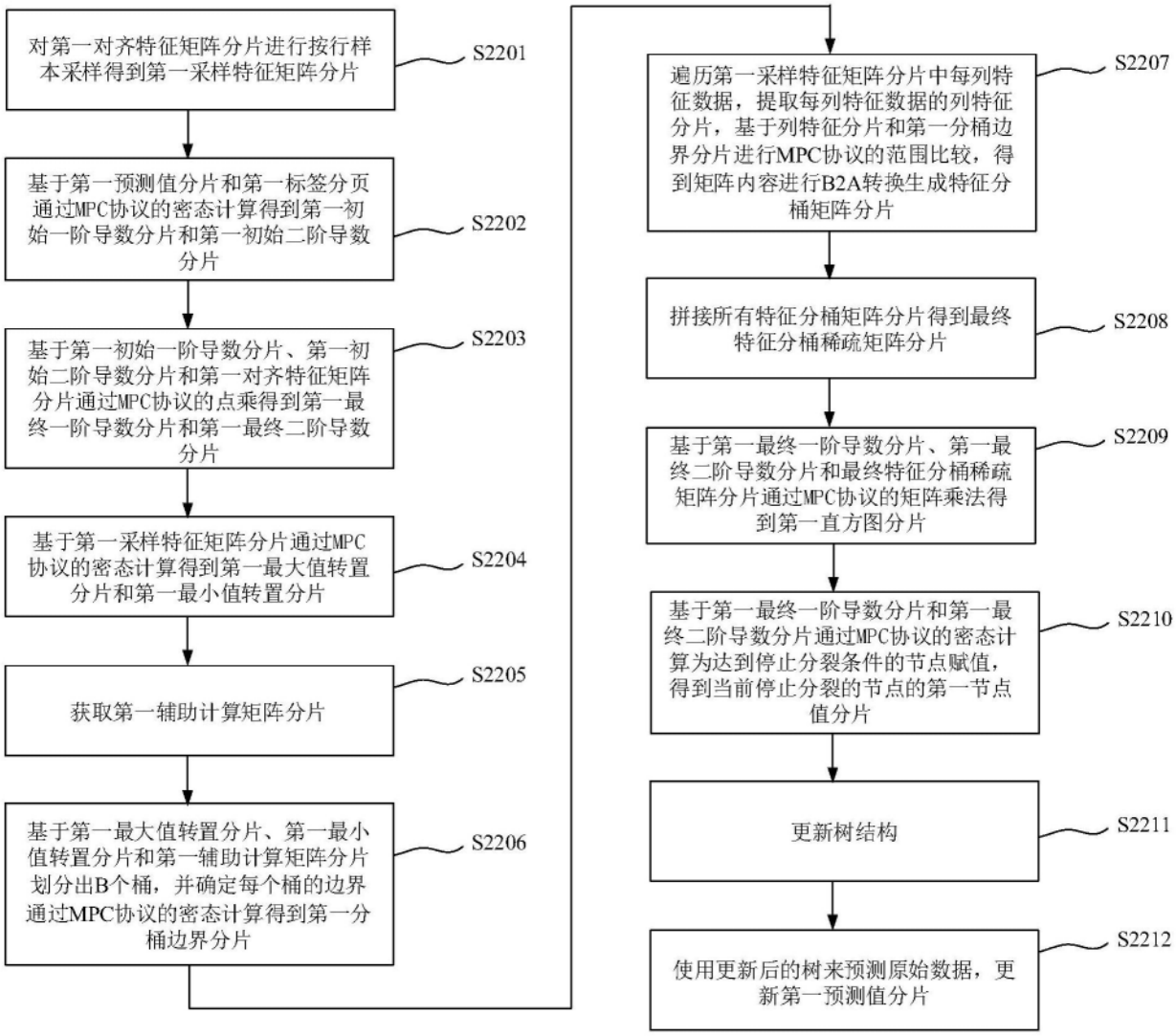

对第一对齐特征矩阵分片进行按行样本采样得到第一采样特征矩阵分片
S2201
基于第一预测值分片和第一标签分页通过MPC协议的密态计算得到第一初始一阶导数分片和第一初始二阶导数分片
S2202
基于第一初始一阶导数分片、第一初始二阶导数分片和第一对齐特征矩阵分片通过MPC协议的点乘得到第一最终一阶导数分片和第一最终二阶导数分片
S2203
基于第一采样特征矩阵分片通过MPC协议的密态计算得到第一最大值转置分片和第一最小值转置分片
S2204
获取第一辅助计算矩阵分片
S2205
基于第一最大值转置分片、第一最小值转置分片和第一辅助计算矩阵分片划分出B个桶，并确定每个桶的边界通过MPC协议的密态计算得到第一分桶边界分片
S2206
遍历第一采样特征矩阵分片中每列特征数据，提取每列特征数据的列特征分片，基于列特征分片和第一分桶边界分片进行MPC协议的范围比较，得到矩阵内容进行B2A转换生成特征分桶矩阵分片
S2207
拼接所有特征分桶矩阵分片得到最终特征分桶稀疏矩阵分片
S2208
基于第一最终一阶导数分片、第一最终二阶导数分片和最终特征分桶稀疏矩阵分片通过MPC协议的矩阵乘法得到第一直方图分片
S2209
基于第一最终一阶导数分片和第一最终二阶导数分片通过MPC协议的密态计算为达到停止分裂条件的节点赋值，得到当前停止分裂的节点的第一节点值分片
S2210
更新树结构
S2211
使用更新后的树来预测原始数据，更新第一预测值分片
S2212

图5A

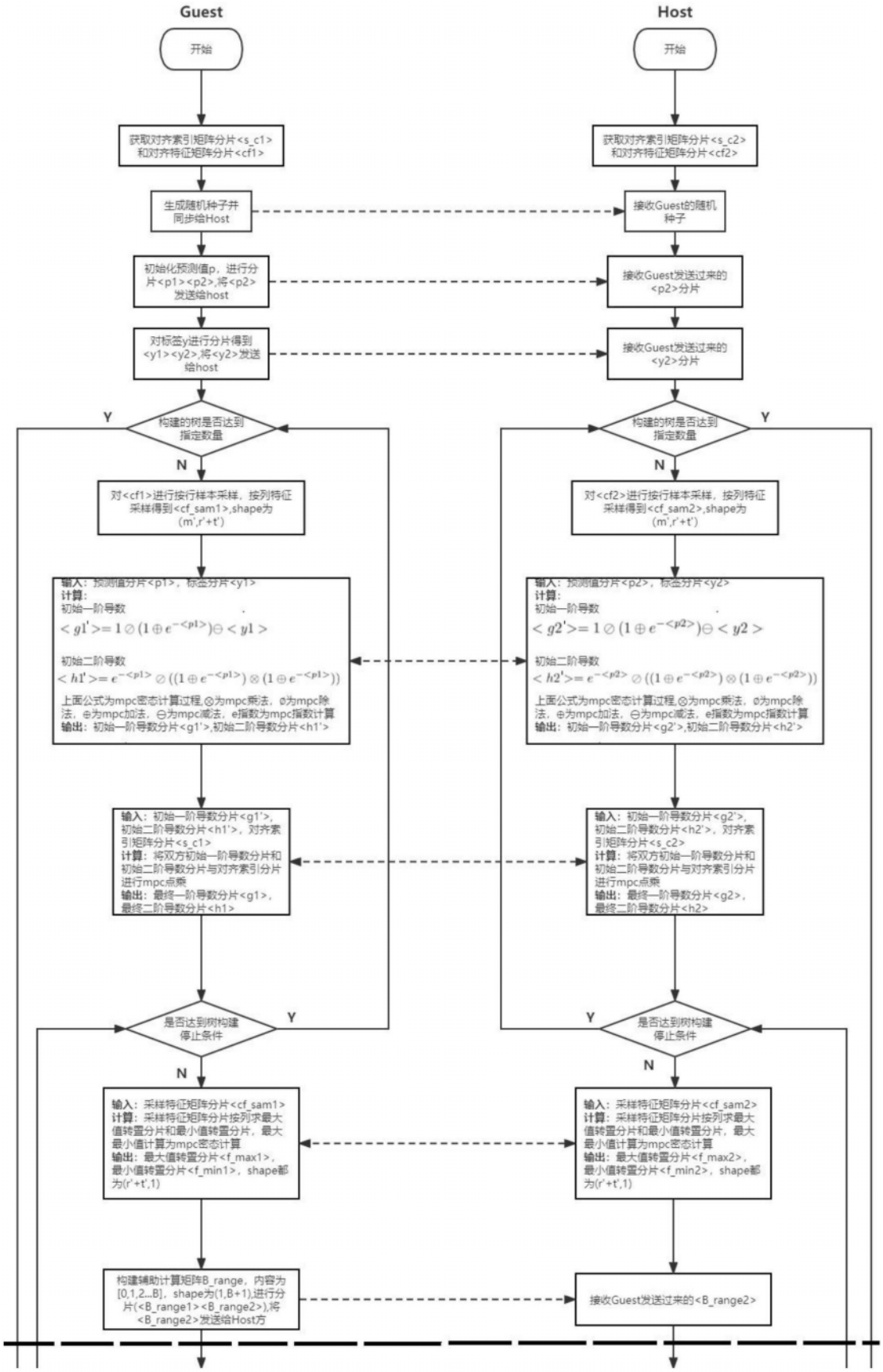

Guest
开始
获取对齐索引矩阵分片<s_c1>和对齐特征矩阵分片<cf1>
生成随机种子并同步给Host
初始化预测值p，进行分片<p1><p2>,将<p2>发送给host
对标签y进行分片得到<y1><y2>,将<y2>发送给host
构建的树是否达到指定数量
Y
N
对<cf1>进行按行样本采样，按列特征采样得到<cf_sam1>,shape为(m',r'+t')
输入：预测值分片<p1>，标签分片<y1>
计算：
初始一阶导数
$< g1' >= 1 \oslash (1 \oplus e^{-<p1>}) \ominus < y1 >$
初始二阶导数
$< h1' >= e^{-<p1>} \oslash ((1 \oplus e^{-<p1>}) \otimes (1 \oplus e^{-<p1>}))$
上面公式为mpc密态计算过程,⊗为mpc乘法，∅为mpc除法，⊕为mpc加法，⊖为mpc减法，e指数为mpc指数计算
输出：初始一阶导数分片<g1'>,初始二阶导数分片<h1'>
输入：初始一阶导数分片<g1'>，初始二阶导数分片<h1'>，对齐索引矩阵分片<s_c1>
计算：将双方初始一阶导数分片和初始二阶导数分片与对齐索引分片进行mpc点乘
输出：最终一阶导数分片<g1>，最终二阶导数分片<h1>
是否达到树构建停止条件
Y
N
输入：采样特征矩阵分片<cf_sam1>
计算：采样特征矩阵分片按列求最大值转置分片和最小值转置分片，最大最小值计算为mpc密态计算
输出：最大值转置分片<f_max1>，最小值转置分片<f_min1>，shape都为(r'+t',1)
构建辅助计算矩阵B_range，内容为[0,1,2...B]，shape为(1,B+1),进行分片(<B_range1><B_range2>),将<B_range2>发送给Host方
Host
开始
获取对齐索引矩阵分片<s_c2>和对齐特征矩阵分片<cf2>
接收Guest的随机种子
接收Guest发送过来的<p2>分片
接收Guest发送过来的<y2>分片
构建的树是否达到指定数量
Y
N
对<cf2>进行按行样本采样，按列特征采样得到<cf_sam2>,shape为(m',r'+t')
输入：预测值分片<p2>，标签分片<y2>
计算：
初始一阶导数
$< g2' >= 1 \oslash (1 \oplus e^{-<p2>}) \ominus < y2 >$
初始二阶导数
$< h2' >= e^{-<p2>} \oslash ((1 \oplus e^{-<p2>}) \otimes (1 \oplus e^{-<p2>}))$
上面公式为mpc密态计算过程,⊗为mpc乘法，∅为mpc除法，⊕为mpc加法，⊖为mpc减法，e指数为mpc指数计算
输出：初始一阶导数分片<g2'>,初始二阶导数分片<h2'>
输入：初始一阶导数分片<g2'>，初始二阶导数分片<h2'>，对齐索引矩阵分片<s_c2>
计算：将双方初始一阶导数分片和初始二阶导数分片与对齐索引分片进行mpc点乘
输出：最终一阶导数分片<g2>，最终二阶导数分片<h2>
是否达到树构建停止条件
Y
N
输入：采样特征矩阵分片<cf_sam2>
计算：采样特征矩阵分片按列求最大值转置分片和最小值转置分片，最大最小值计算为mpc密态计算
输出：最大值转置分片<f_max2>，最小值转置分片<f_min2>，shape都为(r'+t',1)
接收Guest发送过来的<B_range2>
输入：预测值分片<p1>，标签分片<y1>
输入：预测值分片<p2>，标签分片<y2>

图5B1

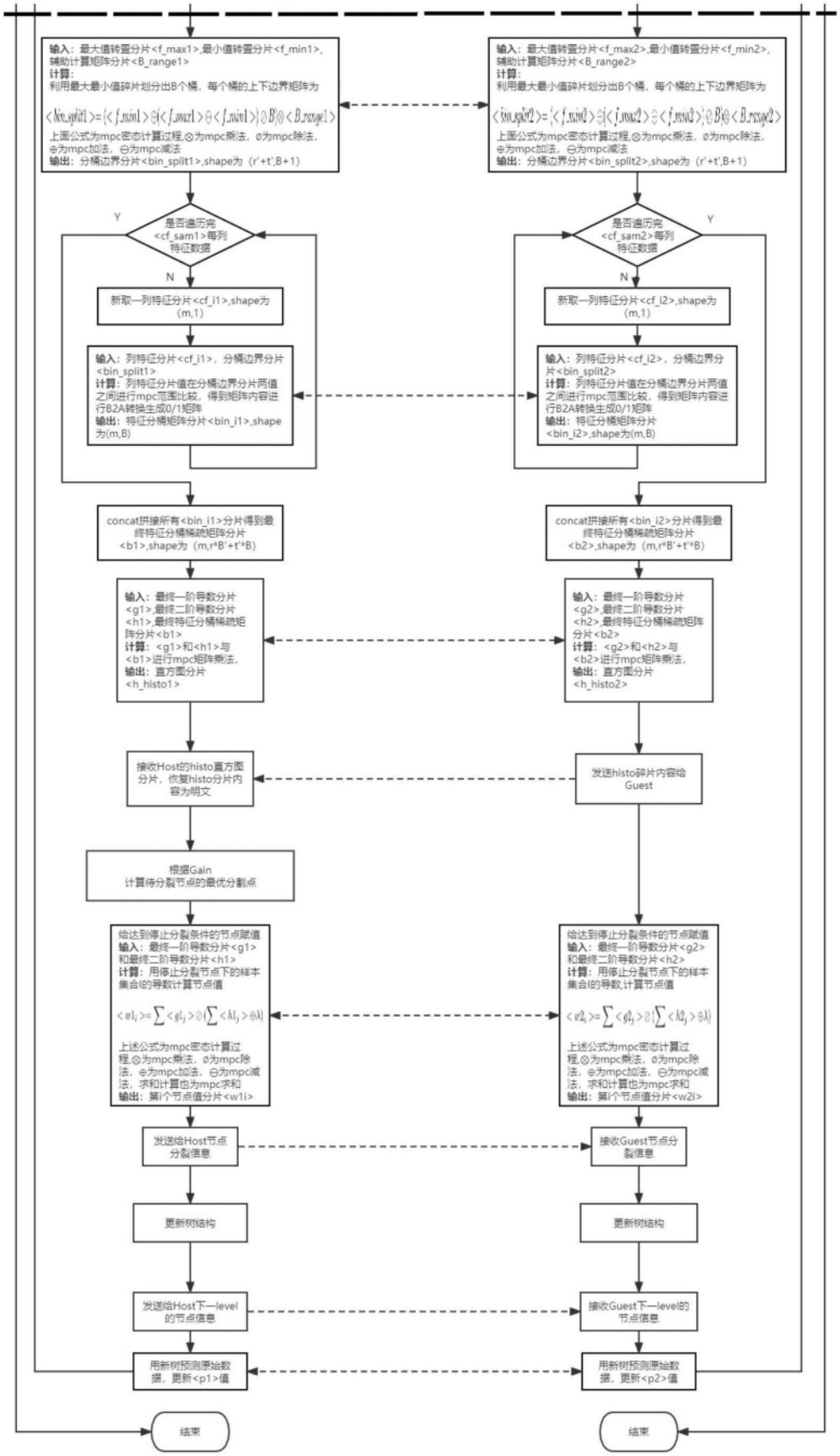

输入：最大值转置分片<f_max1>,最小值转置分片<f_min1>,辅助计算矩阵分片<B_range1>
计算：
利用最大最小值碎片划分出B个桶，每个桶的上下边界矩阵为
上面公式为mpc密态计算过程,⊗为mpc乘法，ø为mpc除法，⊕为mpc加法，⊖为mpc减法
输出：分桶边界分片<bin_split1>,shape为（r'+t',B+1）
输入：最大值转置分片<f_max2>,最小值转置分片<f_min2>,辅助计算矩阵分片<B_range2>
计算：
利用最大最小值碎片划分出B个桶，每个桶的上下边界矩阵为
上面公式为mpc密态计算过程,⊗为mpc乘法，ø为mpc除法，⊕为mpc加法，⊖为mpc减法
输出：分桶边界分片<bin_split2>,shape为（r'+t',B+1）
是否遍历完<cf_sam1>每列特征数据
Y
N
是否遍历完<cf_sam2>每列特征数据
Y
N
新取一列特征分片<cf_i1>,shape为(m,1)
新取一列特征分片<cf_i2>,shape为(m,1)
输入：列特征分片<cf_i1>，分桶边界分片<bin_split1>
计算：列特征分片值在分桶边界分片两值之间进行mpc范围比较，得到矩阵内容进行B2A转换生成0/1矩阵
输出：特征分桶矩阵分片<bin_i1>,shape为(m,B)
输入：列特征分片<cf_i2>，分桶边界分片<bin_split2>
计算：列特征分片值在分桶边界分片两值之间进行mpc范围比较，得到矩阵内容进行B2A转换生成0/1矩阵
输出：特征分桶矩阵分片<bin_i2>,shape为(m,B)
concat拼接所有<bin_i1>分片得到最终特征分桶稀疏矩阵分片<b1>,shape为（m,r*B'+t'*B）
concat拼接所有<bin_i2>分片得到最终特征分桶稀疏矩阵分片<b2>,shape为（m,r*B'+t'*B）
输入：最终一阶导数分片<g1>,最终二阶导数分片<h1>,最终特征分桶稀疏矩阵分片<b1>
计算：<g1>和<h1>与<b1>进行mpc矩阵乘法，
输出：直方图分片<h_histo1>
输入：最终一阶导数分片<g2>,最终二阶导数分片<h2>,最终特征分桶稀疏矩阵分片<b2>
计算：<g2>和<h2>与<b2>进行mpc矩阵乘法，
输出：直方图分片<h_histo2>
接收Host的histo直方图分片，恢复histo分片内容为明文
发送histo碎片内容给Guest
根据Gain计算待分裂节点的最优分割点
给达到停止分裂条件的节点赋值
输入：最终一阶导数分片<g1>和最终二阶导数分片<h1>
计算：用停止分裂节点下的样本集合i的导数计算节点值
上述公式为mpc密态计算过程,⊗为mpc乘法，ø为mpc除法，⊕为mpc加法，⊖为mpc减法，求和计算也为mpc求和
输出：第i个节点值分片<w1i>
给达到停止分裂条件的节点赋值
输入：最终一阶导数分片<g2>和最终二阶导数分片<h2>
计算：用停止分裂节点下的样本集合i的导数,计算节点值
上述公式为mpc密态计算过程,⊗为mpc乘法，ø为mpc除法，⊕为mpc加法，⊖为mpc减法，求和计算也为mpc求和
输出：第i个节点值分片<w2i>
发送给Host节点分裂信息
接收Guest节点分裂信息
更新树结构
更新树结构
发送给Host下一level的节点信息
接收Guest下一level的节点信息
用新树预测原始数据，更新<p1>值
用新树预测原始数据，更新<p2>值
结束
结束

图5B2

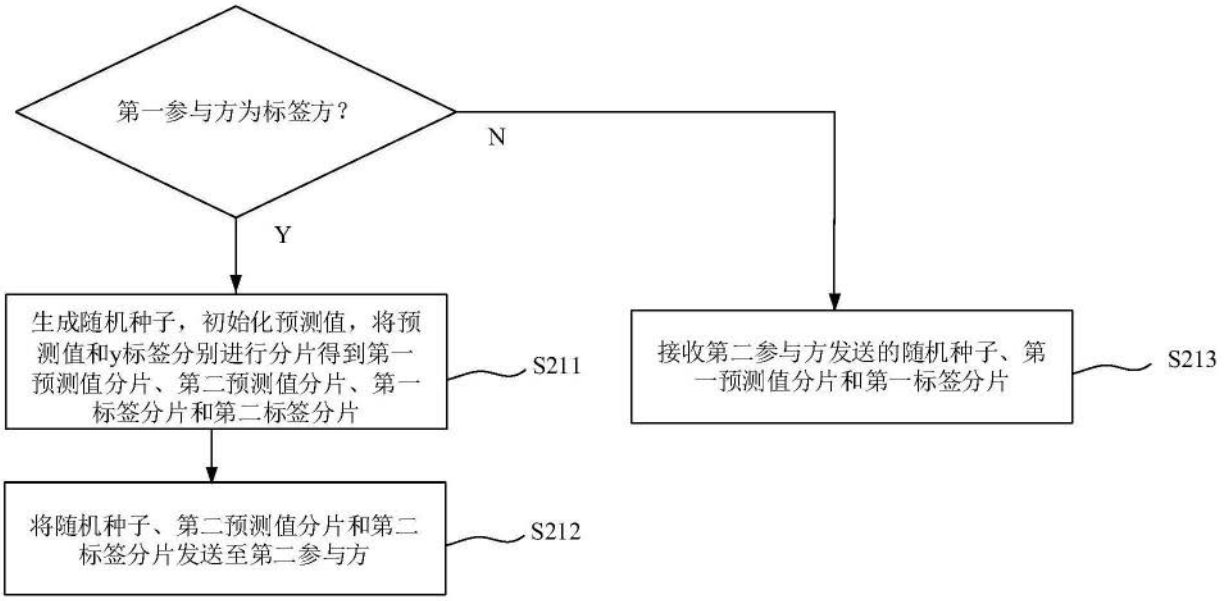

第一参与方为标签方？
N
Y
生成随机种子，初始化预测值，将预测值和y标签分别进行分片得到第一预测值分片、第二预测值分片、第一标签分片和第二标签分片
S211
将随机种子、第二预测值分片和第二标签分片发送至第二参与方
S212
接收第二参与方发送的随机种子、第一预测值分片和第一标签分片
S213

图6

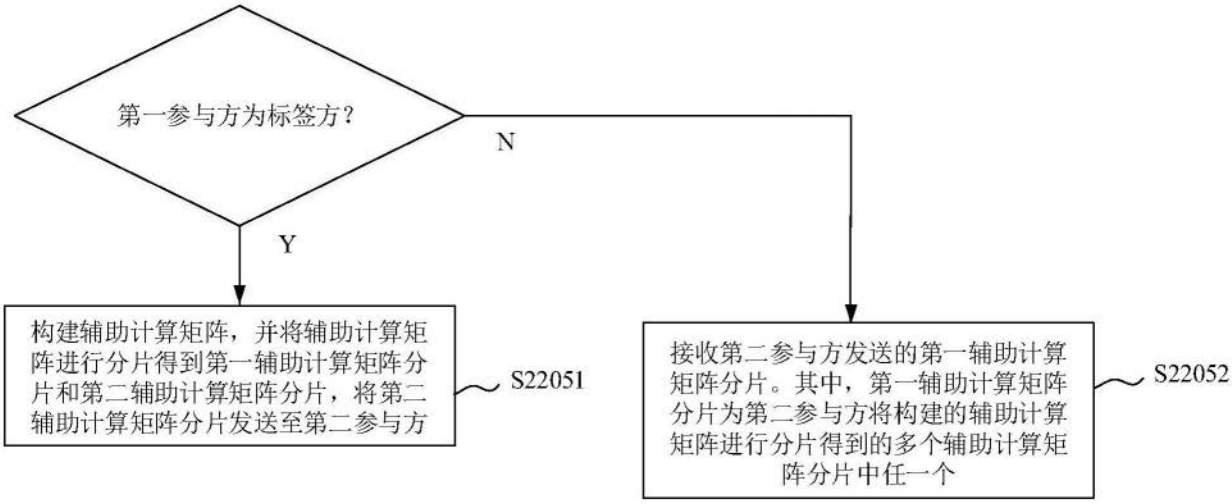

第一参与方为标签方？
N
Y
构建辅助计算矩阵，并将辅助计算矩阵进行分片得到第一辅助计算矩阵分片和第二辅助计算矩阵分片，将第二辅助计算矩阵分片发送至第二参与方
S22051
接收第二参与方发送的第一辅助计算矩阵分片。其中，第一辅助计算矩阵分片为第二参与方将构建的辅助计算矩阵进行分片得到的多个辅助计算矩阵分片中任一个
S22052

图7

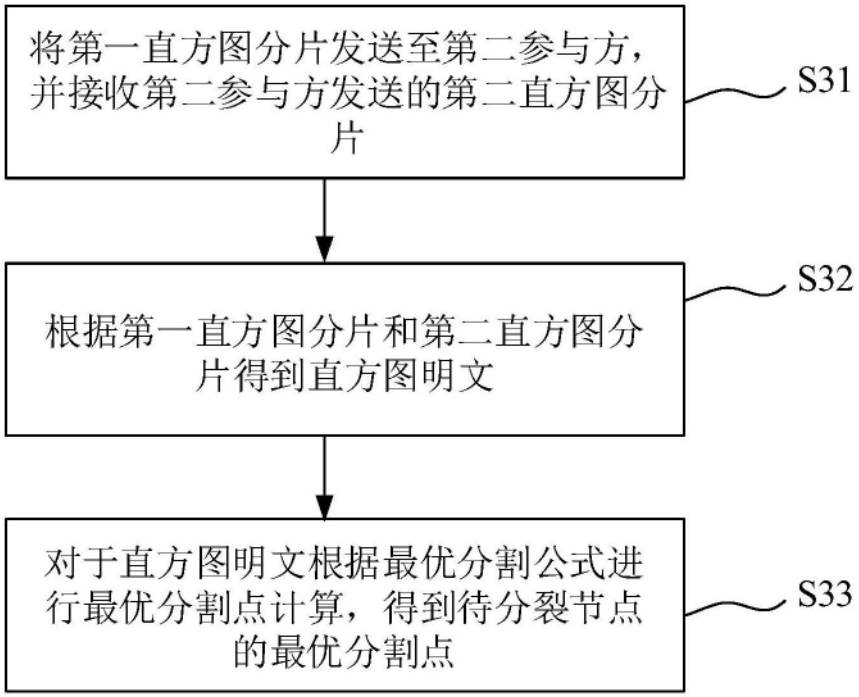

将第一直方图分片发送至第二参与方，并接收第二参与方发送的第二直方图分片
S31
根据第一直方图分片和第二直方图分片得到直方图明文
S32
对于直方图明文根据最优分割公式进行最优分割点计算，得到待分裂节点的最优分割点
S33

图8

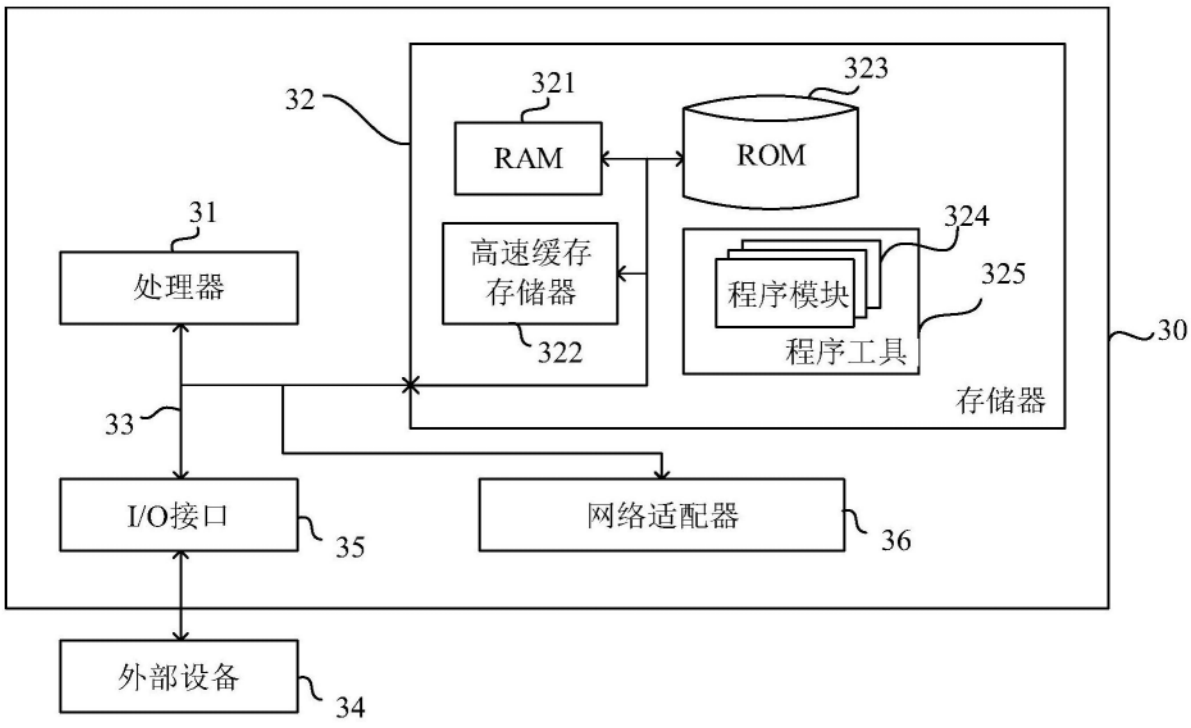

32
321
RAM
323
ROM
31
处理器
高速缓存存储器
322
324
程序模块
325
程序工具
存储器
30
33
I/O接口
35
网络适配器
36
外部设备
34

图9