(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116127495 A
(43)申请公布日 2023.05.16

(21)申请号 202310348364.8

(22)申请日 2023.04.04

(71)申请人 富算科技(上海)有限公司
地址 200135 上海市浦东新区自由贸易试验区浦东大道1200号2层A区

(72)发明人 尤志强 卞阳 王兆凯 赵东 陈立峰 张伟奇

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

(51)Int.Cl.
G06F 21/60(2013.01)
G06F 21/62(2013.01)
G06F 17/16(2006.01)
G06F 18/23(2023.01)
G06N 20/20(2019.01)

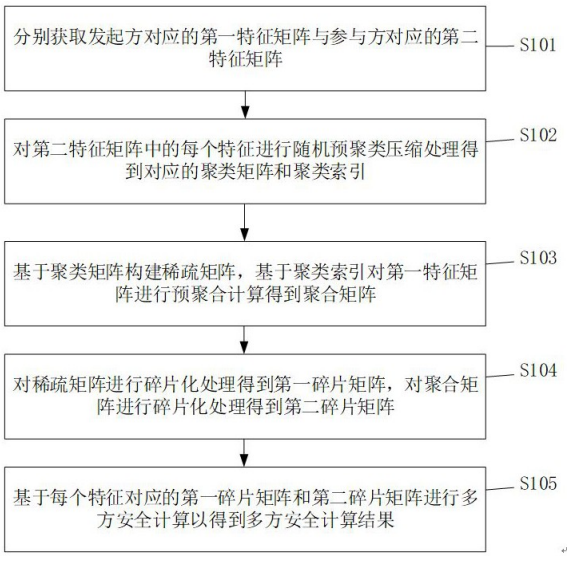
权利要求书2页 说明书12页 附图7页

(54)发明名称

多方安全计算、学习模型的训练方法、系统、设备及介质

(57)摘要

本发明公开了一种多方安全计算、学习模型的训练方法、系统、设备及介质，其中多方安全计算包括分别获取发起方对应的第一特征矩阵与参与方对应的第二特征矩阵；对第二特征矩阵中的每个特征进行随机预聚类压缩处理得到对应的聚类矩阵和聚类索引；基于聚类矩阵构建稀疏矩阵，基于聚类索引对第一特征矩阵进行预聚合计算得到聚合矩阵；对稀疏矩阵进行碎片化处理得到第一碎片矩阵，对聚合矩阵进行碎片化处理得到第二碎片矩阵；基于每个特征对应的第一碎片矩阵和第二碎片矩阵进行多方安全计算以得到多方安全计算结果，通过随机选取聚类中心的方式进行数据压缩，不仅在实现压缩数据的同时，还能够保证了数据的安全性，提高了计算效率。



CN 116127495 A

1.一种多方安全计算方法,其特征在于,应用于至少一个发起方与至少一个参与方之间数据共享场景中,所述方法包括:

分别获取所述发起方对应的第一特征矩阵与所述参与方对应的第二特征矩阵;

对所述第二特征矩阵中的每个特征进行随机预聚类压缩处理得到对应的聚类矩阵和聚类索引;

基于所述聚类矩阵构建稀疏矩阵,基于所述聚类索引对所述第一特征矩阵进行预聚合计算得到聚合矩阵;

对所述稀疏矩阵进行碎片化处理得到第一碎片矩阵,对所述聚合矩阵进行碎片化处理得到第二碎片矩阵;

基于每个所述特征对应的所述第一碎片矩阵和所述第二碎片矩阵进行多方安全计算以得到多方安全计算结果。

2.如权利要求1所述的多方安全计算方法,其特征在于,所述随机预聚类压缩处理的步骤,包括:

在每个所述特征中随机选取若干特征值作为该对应特征的聚类中心;

基于预设聚类规则将所述第二特征矩阵的其余特征值聚类至所述聚类中心处,以得到与各个所述聚类中心相对应的聚类矩阵和聚类索引。

3.如权利要求2所述的多方安全计算方法,其特征在于,所述基于所述聚类矩阵构建稀疏矩阵,包括:

根据所述聚类中心和预设分桶数对所述聚类矩阵中的特征进行分桶处理以得到子稀疏矩阵;

将全部的所述子稀疏矩阵进行拼接以得到所述稀疏矩阵。

4.如权利要求3所述的多方安全计算方法,其特征在于,所述第一特征矩阵包括一阶梯度特征矩阵和二阶梯度特征矩阵;

所述基于所述聚类索引对所述第一特征矩阵进行预聚合计算得到聚合矩阵,包括:

基于所述聚类索引对所述一阶梯度特征矩阵进行预聚合计算得到一阶聚合矩阵;

基于所述聚类索引对所述二阶梯度特征矩阵进行预聚合计算得到二阶聚合矩阵。

5.如权利要求4所述的多方安全计算方法,其特征在于,在所述基于所述聚类索引对所述第一特征矩阵进行预聚合计算得到聚合矩阵之后,所述方法还包括:

对所述一阶聚合矩阵进行扩展处理以得到第一聚合矩阵,对所述二阶聚合矩阵进行扩展处理以得到第二聚合矩阵;

所述对所述聚合矩阵进行碎片化处理得到第二碎片矩阵,包括:

对所述第一聚合矩阵进行碎片化处理得到第一子碎片矩阵,对所述第二聚合矩阵进行碎片化处理得到第二子碎片矩阵;

基于所述第一子碎片矩阵和所述第二子碎片矩阵得到所述第二碎片矩阵。

6.如权利要求5所述的多方安全计算方法,其特征在于,所述基于每个所述特征对应的所述第一碎片矩阵和所述第二碎片矩阵进行多方安全计算以得到多方安全计算结果,包括:

基于每个所述特征对应的所述第一碎片矩阵、第一子碎片矩阵和所述第二子碎片矩阵进行点乘计算,得到第一计算结果;

将每个所述特征对应的第一计算结果进行求和计算得到一阶梯度直方图和二阶梯度直方图；

将所述一阶梯度直方图和所述二阶梯度直方图进行拼接处理以得到所述多方安全计算结果。

7.一种联邦学习模型的训练方法，其特征在于，所述训练方法包括：

分别获取发起方和参与方的求交特征数据集；

利用所述求交特征数据集构建XGBoost树模型；

通过如权利要求1-6中任一项所述的多方安全计算方法得到的聚合矩阵以计算所述XGBoost树模型的最优分割点；

利用所述最优分割点更新所述XGBoost树模型；

将待预测的数据输入至更新后的所述XGBoost树模型进行预测，得到预测结果。

8.一种多方安全计算系统，其特征在于，应用于至少一个发起方与至少一个参与方之间数据共享场景中，所述系统包括：

获取模块，分别获取所述发起方对应的第一特征矩阵与所述参与方对应的第二特征矩阵；

预聚类压缩模块，用于对所述第二特征矩阵中的每个特征进行随机预聚类压缩处理得到对应的聚类矩阵和聚类索引；

矩阵计算模块，用于基于所述聚类矩阵构建稀疏矩阵，基于所述聚类索引对所述第一特征矩阵进行预聚合计算得到聚合矩阵；

碎片化模块，用于对所述稀疏矩阵进行碎片化处理得到第一碎片矩阵，对所述聚合矩阵进行碎片化处理得到第二碎片矩阵；

计算模块，用于基于每个所述特征对应的所述第一碎片矩阵和所述第二碎片矩阵进行多方安全计算以得到多方安全计算结果。

9.一种联邦学习模型的训练系统，其特征在于，所述训练系统包括：

数据集获取模块，用于分别获取发起方和参与方的求交特征数据集；

模型构建模块，用于利用所述求交特征数据集构建XGBoost树模型；

分割点计算模块，用于通过如权利要求1-6中任一项所述的多方安全计算方法得到的聚合矩阵以计算所述XGBoost树模型的最优分割点；

模型更新模块，用于利用所述最优分割点更新所述XGBoost树模型；

模型预测模块，用于将待预测的数据输入至更新后的所述XGBoost树模型进行预测，得到预测结果。

10.一种电子设备，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如权利要求1-6中任一项所述的多方安全计算方法；或，实现如权利要求7所述的联邦学习模型的训练方法。

11.一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如权利要求1-6中任一项所述的多方安全计算方法；或，实现如权利要求7所述的联邦学习模型的训练方法。

多方安全计算、学习模型的训练方法、系统、设备及介质

技术领域

[0001] 本发明涉及多方安全计算的技术领域，特别涉及一种多方安全计算、学习模型的训练方法、系统、设备及介质。

背景技术

[0002] 随着人工智能技术的发展，人们为解决数据孤岛的问题，提出了“联邦学习”的概念，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。它的核心思想是在多个数据源共同参与模型训练时，不需要进行原始数据流转的前提下，仅通过交互模型中间参数进行模型联合训练，原始数据可以不出本地。这种方式实现数据隐私保护和数据共享分析的平衡，即“数据可用不可见”的数据应用模式。

[0003] 联邦学习中的发起方和参与方作为成员方，在不用给出己方数据的情况下，也可进行模型训练得到模型参数，并且可以避免数据隐私泄露的问题。由于联邦学习过程需要大量的数据来支持，而数据又大都分布于不同的数据持有方，所以需要联合各个数据持有方来进行模型构建。

[0004] XGBoost(Exterme Gradient Boosting)全称为极限梯度提升树模型，是一种基于决策树的集成机器学习算法，因其模型预测能力强，在工业界被广泛使用，比如应用在广告推荐、金融风控等业务场景。然而将该算法应用于联邦学习场景，由于对该算法的联邦化设计目前还处于不成熟阶段，普遍来说当前的联邦学习XGBoost算法训练的通信量比较大，导致训练耗时过长，不能满足业界需求。

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的联邦学习模型训练耗时长、效率低的缺陷，提供一种多方安全计算、学习模型的训练方法、系统、设备及介质。

[0006] 本发明是通过下述技术方案来解决上述技术问题：

[0007] 本发明提供一种多方安全计算方法，应用于至少一个发起方与至少一个参与方之间数据共享场景中，所述方法包括：

[0008] 分别获取所述发起方对应的第一特征矩阵与所述参与方对应的第二特征矩阵；

[0009] 对所述第二特征矩阵中的每个特征进行随机预聚类压缩处理得到对应的聚类矩阵和聚类索引；

[0010] 基于所述聚类矩阵构建稀疏矩阵，基于所述聚类索引对所述第一特征矩阵进行预聚合计算得到聚合矩阵；

[0011] 对所述稀疏矩阵进行碎片化处理得到第一碎片矩阵，对所述聚合矩阵进行碎片化处理得到第二碎片矩阵；

[0012] 基于每个所述特征对应的所述第一碎片矩阵和所述第二碎片矩阵进行多方安全计算以得到多方安全计算结果。

[0013] 较佳地,所述随机预聚类压缩处理的步骤,包括:

[0014] 在每个所述特征中随机选取若干特征值作为该对应特征的聚类中心;

[0015] 基于预设聚类规则将所述第二特征矩阵的其余特征值聚类至所述聚类中心处,以得到与各个所述聚类中心相对应的聚类矩阵和聚类索引。

[0016] 较佳地,所述基于所述聚类矩阵构建稀疏矩阵,包括:

[0017] 根据所述聚类中心和预设分桶数对所述聚类矩阵中的特征进行分桶处理以得到子稀疏矩阵;

[0018] 将全部的所述子稀疏矩阵进行拼接以得到所述稀疏矩阵。

[0019] 较佳地,所述第一特征矩阵包括一阶梯度特征矩阵和二阶梯度特征矩阵;

[0020] 所述基于所述聚类索引对所述第一特征矩阵进行预聚合计算得到聚合矩阵,包括:

[0021] 基于所述聚类索引对所述一阶梯度特征矩阵进行预聚合计算得到一阶聚合矩阵;

[0022] 基于所述聚类索引对所述二阶梯度特征矩阵进行预聚合计算得到二阶聚合矩阵。

[0023] 较佳地,在所述基于所述聚类索引对所述第一特征矩阵进行预聚合计算得到聚合矩阵之后,所述方法还包括:

[0024] 对所述一阶聚合矩阵进行扩展处理以得到第一聚合矩阵,对所述二阶聚合矩阵进行扩展处理以得到第二聚合矩阵;

[0025] 所述对所述聚合矩阵进行碎片化处理得到第二碎片矩阵,包括:

[0026] 对所述第一聚合矩阵进行碎片化处理得到第一子碎片矩阵,对所述第二聚合矩阵进行碎片化处理得到第二子碎片矩阵;

[0027] 基于所述第一子碎片矩阵和所述第二子碎片矩阵得到所述第二碎片矩阵。

[0028] 较佳地,所述基于每个所述特征对应的所述第一碎片矩阵和所述第二碎片矩阵进行多方安全计算以得到多方安全计算结果,包括:

[0029] 基于每个所述特征对应的所述第一碎片矩阵、第一子碎片矩阵和所述第二子碎片矩阵进行点乘计算,得到第一计算结果;

[0030] 将每个所述特征对应的第一计算结果进行求和计算得到一阶梯度直方图和二阶梯度直方图;

[0031] 将所述一阶梯度直方图和所述二阶梯度直方图进行拼接处理以得到所述多方安全计算结果。

[0032] 本发明还提供一种联邦学习模型的训练方法,所述训练方法包括:

[0033] 分别获取发起方和参与方的求交特征数据集;

[0034] 利用所述求交特征数据集构建XGBoost树模型;

[0035] 通过如上所述的多方安全计算方法得到的聚合矩阵以计算所述XGBoost树模型的最优分割点;

[0036] 利用所述最优分割点更新所述XGBoost树模型;

[0037] 将待预测的数据输入至更新后的所述XGBoost树模型进行预测,得到预测结果。

[0038] 本发明还提供一种多方安全计算系统,其特征在于,应用于至少一个发起方与至少一个参与方之间数据共享场景中,所述系统包括:

[0039] 获取模块,分别获取所述发起方对应的第一特征矩阵与所述参与方对应的第二特

征矩阵；

[0040] 预聚类压缩模块，用于对所述第二特征矩阵中的每个特征进行随机预聚类压缩处理得到对应的聚类矩阵和聚类索引；

[0041] 矩阵计算模块，用于基于所述聚类矩阵构建稀疏矩阵，基于所述聚类索引对所述第一特征矩阵进行预聚合计算得到聚合矩阵；

[0042] 碎片化模块，用于对所述稀疏矩阵进行碎片化处理得到第一碎片矩阵，对所述聚合矩阵进行碎片化处理得到第二碎片矩阵；

[0043] 计算模块，用于基于每个所述特征对应的所述第一碎片矩阵和所述第二碎片矩阵进行多方安全计算以得到多方安全计算结果。

[0044] 较佳地，所述预聚类压缩模块，还用于在每个所述特征中随机选取若干特征值作为该对应特征的聚类中心；

[0045] 基于预设聚类规则将所述第二特征矩阵的其余特征值聚类至所述聚类中心处，以得到与各个所述聚类中心相对应的聚类矩阵和聚类索引。

[0046] 较佳地，所述矩阵计算模块，还用于根据所述聚类中心和预设分桶数对所述聚类矩阵中的特征进行分桶处理以得到子稀疏矩阵；

[0047] 将全部的所述子稀疏矩阵进行拼接以得到所述稀疏矩阵。

[0048] 较佳地，所述第一特征矩阵包括一阶梯度特征矩阵和二阶梯度特征矩阵；

[0049] 所述矩阵计算模块，还用于基于所述聚类索引对所述一阶梯度特征矩阵进行预聚合计算得到一阶聚合矩阵；

[0050] 基于所述聚类索引对所述二阶梯度特征矩阵进行预聚合计算得到二阶聚合矩阵。

[0051] 较佳地，所述系统还包括扩展模块，所述扩展模块用于对所述一阶聚合矩阵进行扩展处理以得到第一聚合矩阵，对所述二阶聚合矩阵进行扩展处理以得到第二聚合矩阵；

[0052] 所述碎片化模块，还用于对所述聚合矩阵进行碎片化处理得到第二碎片矩阵，包括：

[0053] 对所述第一聚合矩阵进行碎片化处理得到第一子碎片矩阵，对所述第二聚合矩阵进行碎片化处理得到第二子碎片矩阵；

[0054] 基于所述第一子碎片矩阵和所述第二子碎片矩阵得到所述第二碎片矩阵。

[0055] 较佳地，所述计算模块，还用于基于每个所述特征对应的所述第一碎片矩阵、第一子碎片矩阵和所述第二子碎片矩阵进行点乘计算，得到第一计算结果；

[0056] 将每个所述特征对应的第一计算结果进行求和计算得到一阶梯度直方图和二阶梯度直方图；

[0057] 将所述一阶梯度直方图和所述二阶梯度直方图进行拼接处理以得到所述多方安全计算结果。

[0058] 本发明还提供一种联邦学习模型的训练系统，所述训练系统包括：

[0059] 数据集获取模块，用于分别获取发起方和参与方的求交特征数据集；

[0060] 模型构建模块，用于利用所述求交特征数据集构建XGBoost树模型；

[0061] 分割点计算模块，用于通过如上所述的多方安全计算方法得到的聚合矩阵以计算所述XGBoost树模型的最优分割点；

[0062] 模型更新模块，用于利用所述最优分割点更新所述XGBoost树模型；

[0063] 模型预测模块，用于将待预测的数据输入至更新后的所述XGBoost树模型进行预测，得到预测结果。

[0064] 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如上所述的多方安全计算方法；或，实现如上所述的联邦学习模型的训练方法。

[0065] 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如上所述的多方安全计算方法；或，实现如上所述的联邦学习模型的训练方法。

[0066] 本发明的积极进步效果在于：分别获取发起方对应的第一特征矩阵与参与方对应的第二特征矩阵；对第二特征矩阵中的每个特征进行随机预聚类压缩处理得到对应的聚类矩阵和聚类索引；基于聚类矩阵构建稀疏矩阵，基于聚类索引对第一特征矩阵进行预聚合计算得到聚合矩阵；对稀疏矩阵进行碎片化处理得到第一碎片矩阵，对聚合矩阵进行碎片化处理得到第二碎片矩阵；基于每个特征对应的第一碎片矩阵和第二碎片矩阵进行多方安全计算以得到多方安全计算结果，通过随机选取聚类中心的方式进行数据压缩，不仅在实现压缩数据的同时，还能够保证了数据的安全性，进而降低发起方和参与方的数据传输量，提高了计算效率。

附图说明

[0067] 图1为现有XGBoost在Host方的直方图碎片态矩阵计算方法。

[0068] 图2为本发明实施例提供的多方安全计算方法的第一流程示意图。

[0069] 图3为本发明实施例提供的多方安全计算方法的第二流程示意图。

[0070] 图4为本发明实施例提供的多方安全计算方法的第三流程示意图。

[0071] 图5为本发明实施例提供的多方安全计算方法的第四流程示意图。

[0072] 图6为本发明实施例提供的多方安全计算方法的第五流程示意图。

[0073] 图7为本发明实施例提供的多方安全计算方法的第六流程示意图。

[0074] 图8为本发明实施例提供的多方安全计算系统的模块示意图。

[0075] 图9为本发明实施例提供的联邦学习模型的训练方法的第一流程示意图。

[0076] 图10a为本发明实施例提供的联邦学习模型的训练方法的第二流程示意图的第一部分。

[0077] 图10b为本发明实施例提供的联邦学习模型的训练方法的第二流程示意图的第二部分。

[0078] 图11为本发明实施例提供的联邦学习模型的训练系统的模块示意图。

[0079] 图12为本发明实施例提供的实现多方安全计算方法或联邦学习模型的训练方法的电子设备的结构示意图。

具体实施方式

[0080] 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。

[0081] 为了更清楚地说明本说明书实施例的技术方案，下面将对实施例描述中所需要使

用的附图作简单的介绍。显而易见地，下面描述中的附图仅仅是本说明书的一些示例或实施例，对于本领域的普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图将本说明书应用于其它类似情景。除非从语言环境中显而易见或另做说明，图中相同标号代表相同结构或操作。

[0082] 在本申请中提及“实施例”意味着，结合实施例描述的特定特征、结构或特性可以包含在本申请的至少一个实施例中。在说明书中的各个位置出现该短语并不一定均是指相同的实施例，也不是与其它实施例互斥的独立的或备选的实施例。本领域技术人员显式地和隐式地理解的是，本申请所描述的实施例可以与其它实施例相结合。

[0083] 应当理解，本文使用的“系统”、“装置”、“单元”和/或“模组”是用于区分不同级别的不同组件、元件、部件、部分或装配的一种方法。然而，如果其他词语可实现相同的目的，则可通过其他表达来替换所述词语。

[0084] 如本说明书中所示，除非上下文明确提示例外情形，“一”、“一个”、“一种”和/或“该”等词并非特指单数，也可包括复数。一般说来，术语“包括”与“包含”仅提示包括已明确标识的步骤和元素，而这些步骤和元素不构成一个排它性的罗列，方法或者设备也可能包含其它的步骤或元素。

[0085] 本说明书中使用了流程图用来说明根据本说明书的实施例的系统所执行的操作。应当理解的是，前面或后面操作不一定按照顺序来精确地执行。相反，可以按照倒序或同时处理各个步骤。同时，也可以将其他操作添加到这些过程中，或从这些过程移除某一步或数步操作。

[0086] XGBoost是一种基于决策树的集成机器学习算法，因其模型预测能力强，在工业界被广泛使用，比如应用在广告推荐、金融风控等业务场景。然而将该算法应用于联邦学习场景中时，由于对该算法的联邦化设计目前还处于不成熟阶段，普遍来说当前的联邦学习XGBoost算法训练存在以下几个缺点：

[0087] (1)交互数据过大

[0088] 在MPC XGBoost中假设一个数据集有40万样本数据，每条数据有600个特征，那么在直方图计算过程中会有255G的数据传输量，如图1示出了一个原始MPC XGBoost 在Host方直方图碎片态矩阵计算过程。这还仅仅是构建一张直方图的，这种数据级别的数据交互使得难以在大样本上进行高效的模型训练。

[0089] (2)计算性能不高

[0090] 原始MPC XGBoost 中核心交互部分采用全碎片态进行计算，使得计算性能过低。主要是为了保证数据安全性，在矩阵乘法时没有利用稀疏化矩阵可加速计算的特性。还是以40万训练样本集数据为例，其中每条数据有600个特征，每个特征分50个桶，整个过程需要执行240亿次乘法运算和239亿9994万次加法运算。

[0091] 由此可见，当前的联邦学习XGBoost算法训练的交互数据过大且计算性能不高，导致训练耗时过长，从而满足不了业界的需求。

[0092] 基于上述原因，如图2所示，本实施例提供一种多方安全计算方法，应用于至少一个发起方与至少一个参与方之间数据共享场景中，在本实施例中，以一个发起方(Guest)和一个参与方(Host)为例进行说明，本实施例的多方安全计算方法包括：

[0093] S101、分别获取发起方对应的第一特征矩阵与参与方对应的第二特征矩阵。

[0094] 在一种可选地实施方式，第一特征矩阵包括一阶梯度特征矩阵和二阶梯度特征矩阵，本实施例对此不作限制。

[0095] S102、对第二特征矩阵中的每个特征进行随机预聚类压缩处理得到对应的聚类矩阵和聚类索引。

[0096] 需要说明的是，特征即可以是连续性特征，也可以是离散性特征，本实施例对此不作限制，本实施例以连续性特征为例。

[0097] 作为本实施例的一个可选地实施方式，为了减少最后MPC计算的数据量，可以对特征进行聚类压缩，其中，如图3所示，随机预聚类压缩处理的步骤，包括：

[0098] S201、在每个特征中随机选取若干特征值作为该对应特征的聚类中心。

[0099] 可选地，随机选择对应特征中若干特征值或者生成随机数作为该特征的聚类中心。

[0100] S202、基于预设聚类规则将第二特征矩阵的其余特征值聚类至聚类中心处，以得到与各个聚类中心相对应的聚类矩阵和聚类索引。

[0101] 每一个特征都会执行，在每一个特征中随机选取若干特征值作为该对应特征的聚类中心。也就是如果有100个特征，是会对这100个特征分别都做一次随机聚类中心的选取并执行预聚类，每一个特征的操作是独立的事件。

[0102] 在一个具体例子中，对f个参与方(Host)方的特征依次随机选取k个聚类中心，将样本的特征聚到近似的聚类中心上。如第一个特征内容为[10.5,12.34,2.66,9.5 ... 10.13]，其中随机选择的聚类中心有9.5, 11, 2.66等。所以聚类结果可表达为{9.5:[3,...],11:[0,1,...],... 2.66:[2,...]}，给Guest方发送聚类索引，还是上面那个特征为例最终发送过去的内容为{0:[3,...], 1:[0,1,...],... k-1:[2,...]}，即会隐去实际数据，仅保留位置次序。

[0103] 由于预聚类中心点是随机选取，即对每个特征使用了随机数据作为预聚类中心点，这样的好处是即使发送给Host聚类中心的索引(非实际数据内容)也不会暴露特征分布，也就是说即使拿到了聚类索引，其实是无法推知数据特征的分布信息，是具有完全随机性的，而且发送的索引内容对Guest方是没有实际意义的，其不知道特征意义并且也不知道聚类的实际分箱结果，从而能够有效保证数据的安全性。

[0104] S103、基于聚类矩阵构建稀疏矩阵，基于聚类索引对第一特征矩阵进行预聚合计算得到聚合矩阵。

[0105] 作为本实施例的一个可选地实施方式，如图4所示，步骤S103包括：

[0106] S103a1、根据聚类中心和预设分桶数对聚类矩阵中的特征进行分桶处理以得到子稀疏矩阵。

[0107] S103a2、将全部的子稀疏矩阵进行拼接以得到稀疏矩阵。

[0108] 可选地，Host方基于聚类后的特征内容构建稀疏矩阵，用来表示对聚合中心的分桶，即将不同特征的聚合中心数据落到不同桶中，如上面case的9.5,11,...,2.66进行分桶，最终表示为0/1稀疏矩阵histo，这里一个特征的shape为(k, b)，Host方所有特征的结果组合后，shape为(k, f*b)。

[0109] 作为本实施例的一个可选地实施方式，如图5所示，步骤S103还包括：

[0110] S103b1、基于聚类索引对一阶梯度特征矩阵进行预聚合计算得到一阶聚合矩阵。

[0111] S103b1、基于聚类索引对二阶梯度特征矩阵进行预聚合计算得到二阶聚合矩阵。

[0112] 可选地，根据Host方发送过来的聚类索引，Guest方事先对一阶梯度特征矩阵(一阶梯度g)和二阶梯度特征矩阵(二阶梯度h)进行聚类计算。如根据{0:[3,...], 1:[0,1,...], ... k-1:[2,...]}，会生成[g3+...,g0+g1+...,...,g2+...]一阶聚合矩阵(一阶梯度聚合结果clu_g)，和[h3+...,h0+h1+...,...,h2+...]二阶聚合矩阵(二阶梯度聚合结果clu_h)。所以Host方所有特征会生成shape为(f,k)的一阶梯度结果和二阶梯度结果，f为特征数量，k为聚合中心数。为了后续计算表示方便会对结果进行转置，shape为(k,f)。

[0113] 作为本实施例的一个可选地实施方式，在步骤S103之后，本实施例的多方安全计算方法还包括：

[0114] S1031、对一阶聚合矩阵进行扩展处理以得到第一聚合矩阵，对二阶聚合矩阵进行扩展处理以得到第二聚合矩阵。

[0115] 可选地，Guest方对一阶聚合矩阵(一阶梯度聚合结果clu_g)和二阶聚合矩阵(二阶梯度聚合结果clu_h)，第二维内容复制b次，以扩展得到shape为(k, f*b)的clu_g和clu_h。

[0116] S104、对稀疏矩阵进行碎片化处理得到第一碎片矩阵，对聚合矩阵进行碎片化处理得到第二碎片矩阵。

[0117] 可选地，对稀疏矩阵进行碎片化处理得到第一碎片矩阵后进行秘密共享，对聚合矩阵进行碎片化处理得到第二碎片矩阵进行秘密共享。

[0118] 作为本实施例的一个可选地实施方式，如图6所示，步骤S104，包括：

[0119] S1041、对第一聚合矩阵进行碎片化处理得到第一子碎片矩阵，对第二聚合矩阵进行碎片化处理得到第二子碎片矩阵。

[0120] S1042、基于第一子碎片矩阵和第二子碎片矩阵得到第二碎片矩阵。

[0121] S105、基于每个特征对应的第一碎片矩阵和第二碎片矩阵进行多方安全计算以得到多方安全计算结果。

[0122] 作为本实施例的一个可选地实施方式，如图7所示，步骤S105包括：

[0123] S1051、基于每个特征对应的第一碎片矩阵、第一子碎片矩阵和第二子碎片矩阵进行点乘计算，得到第一计算结果。

[0124] S1052、将每个特征对应的第一计算结果进行求和计算得到一阶梯度直方图和二阶梯度直方图。

[0125] S1053、将一阶梯度直方图和二阶梯度直方图进行拼接处理以得到多方安全计算结果。

[0126] 可选地，Guest方的clu_g和clu_h分片内容(第二碎片矩阵)和Host方的histo分片内容(第一碎片矩阵)进行MPC点乘，点乘结果的shape为(k, f*b)。

[0127] 对每个特征的点乘结果进行相同桶聚类内容的MPC求和，即将shape(k, f*b)的第一维度进行就和得到shape 为(1, f*b)的一阶梯度直方图和二阶梯度直方图，这里b为分桶数量。最后将一阶梯度直方图和二阶梯度直方图按第一维拼接得到多方安全计算结果，即最终的Host方直方图，shape为(2,f*b)，并同步至发起方。

[0128] 从而，本实施例通过随机选取聚类中心的方式进行数据压缩，不仅在实现压缩数据的同时，还能够保证了数据的安全性，进而降低发起方和参与方的数据传输量，提高了计

算效率。

[0129] 与上面介绍的多方安全计算对应地，本实施例还提供了一种多方安全计算系统。下面将分别进行介绍。具体地，如图8所示，本实施例还提供一种多方安全计算系统，该系统包括：

[0130] 获取模块1，分别获取发起方对应的第一特征矩阵与参与方对应的第二特征矩阵。

[0131] 在一种可选地实施方式，第一特征矩阵包括一阶梯度特征矩阵和二阶梯度特征矩阵，本实施例对此不作限制。

[0132] 预聚类压缩模块2，用于对第二特征矩阵中的每个特征进行随机预聚类压缩处理得到对应的聚类矩阵和聚类索引。

[0133] 需要说明的是，特征即可以是连续性特征，也可以是离散性特征，本实施例对此不作限制，本实施例以连续性特征为例。

[0134] 作为本实施例的一个可选地实施方式，为了减少最后MPC计算的数据量，可以对特征进行聚类压缩。

[0135] 预聚类压缩模块2，还用于在每个特征中随机选取若干特征值作为该对应特征的聚类中心。

[0136] 基于预设聚类规则将第二特征矩阵的其余特征值聚类至聚类中心处，以得到与各个聚类中心相对应的聚类矩阵和聚类索引。

[0137] 在一个具体例子中，对f个参与方(Host)方的特征依次随机选取k个聚类中心，将样本的特征聚到近似的聚类中心上。如第一个特征内容为[10.5,12.34,2.66,9.5 ... 10.13]，其中随机选择的聚类中心有9.5, 11, 2.66等。所以聚类结果可表达为{9.5:[3,...],11:[0,1,...],... 2.66:[2,...]}，给Guest方发送聚类索引，还是上面那个特征为例最终发送过去的内容为{0:[3,...], 1:[0,1,...],... k-1:[2,...]}，即会隐去实际数据，仅保留位置次序。

[0138] 由于预聚类中心点是随机选取，即对每个特征使用了随机数据作为预聚类中心点，这样的好处是即使发送给Host聚类中心的索引(非实际数据内容)也不会暴露特征分布，也就是说即使拿到了聚类索引，其实是无法推知数据特征的分布信息，是具有完全随机性的，而且发送的索引内容对Guest方是没有实际意义的，其不知道特征意义并且也不知道聚类的实际分箱结果，从而能够有效保证数据的安全性。

[0139] 矩阵计算模块3，用于基于聚类矩阵构建稀疏矩阵，基于聚类索引对第一特征矩阵进行预聚合计算得到聚合矩阵。

[0140] 作为本实施例的一个可选地实施方式，矩阵计算模块3，还用于根据聚类中心和预设分桶数对聚类矩阵中的特征进行分桶处理以得到子稀疏矩阵。

[0141] 将全部的子稀疏矩阵进行拼接以得到稀疏矩阵。

[0142] 可选地，Host方基于聚类后的特征内容构建稀疏矩阵，用来表示对聚合中心的分桶，即将不同特征的聚合中心数据落到不同桶中，如上面case的9.5,11,...,2.66进行分桶，最终表示为0/1稀疏矩阵histo，这里一个特征的shape为(k, b)，Host方所有特征的结果组合后，shape为(k, f*b)。

[0143] 作为本实施例的一个可选地实施方式，矩阵计算模块3，还用于基于聚类索引对一阶梯度特征矩阵进行预聚合计算得到一阶聚合矩阵。

[0144] 基于聚类索引对二阶梯度特征矩阵进行预聚合计算得到二阶聚合矩阵。

[0145] 可选地,根据Host方发送过来的聚类索引,Guest方事先对一阶梯度特征矩阵(一阶梯度g)和二阶梯度特征矩阵(二阶梯度h)进行聚类计算。如根据{0:[3,...], 1:[0,1,...], ... k-1:[2,...]},会生成[g3+...,g0+g1+...,...,g2+...]一阶聚合矩阵(一阶梯度聚合结果clu_g),和[h3+...,h0+h1+...,...,h2+...]二阶聚合矩阵(二阶梯度聚合结果clu_h)。所以Host方所有特征会生成shape为(f,k)的一阶梯度结果和二阶梯度结果,f为特征数量,k为聚合中心数。为了后续计算表示方便会对结果进行转置,shape为(k,f)。

[0146] 作为本实施例的一个可选地实施方式,本实施例的多方安全计算系统还包括扩展模块6。

[0147] 扩展模块6,用于对一阶聚合矩阵进行扩展处理以得到第一聚合矩阵,对二阶聚合矩阵进行扩展处理以得到第二聚合矩阵。

[0148] 可选地,Guest方对一阶聚合矩阵(一阶梯度聚合结果clu_g)和二阶聚合矩阵(二阶梯度聚合结果clu_h),第二维内容复制b次,以扩展得到shape为(k, f*b)的clu_g和clu_h。

[0149] 碎片化模块4,用于对稀疏矩阵进行碎片化处理得到第一碎片矩阵,对聚合矩阵进行碎片化处理得到第二碎片矩阵。

[0150] 可选地,对稀疏矩阵进行碎片化处理得到第一碎片矩阵后进行秘密共享,对聚合矩阵进行碎片化处理得到第二碎片矩阵进行秘密共享。

[0151] 作为本实施例的一个可选地实施方式,碎片化模块4,还用于对聚合矩阵进行碎片化处理得到第二碎片矩阵,包括:

[0152] 对第一聚合矩阵进行碎片化处理得到第一子碎片矩阵,对第二聚合矩阵进行碎片化处理得到第二子碎片矩阵。

[0153] 基于第一子碎片矩阵和第二子碎片矩阵得到第二碎片矩阵。

[0154] 计算模块5,用于基于每个特征对应的第一碎片矩阵和第二碎片矩阵进行多方安全计算以得到多方安全计算结果。

[0155] 作为本实施例的一个可选地实施方式,计算模块5,还用于基于每个特征对应的第一碎片矩阵、第一子碎片矩阵和第二子碎片矩阵进行点乘计算,得到第一计算结果。

[0156] 将每个特征对应的第一计算结果进行求和计算得到一阶梯度直方图和二阶梯度直方图。

[0157] 将一阶梯度直方图和二阶梯度直方图进行拼接处理以得到多方安全计算结果。

[0158] 可选地,Guest方的clu_g和clu_h分片内容(第二碎片矩阵)和Host方的histo分片内容(第一碎片矩阵)进行MPC点乘,点乘结果的shape为(k, f*b)。

[0159] 对每个特征的点乘结果进行相同桶聚类内容的MPC求和,即将shape(k, f*b)的第一维度进行就和得到shape 为(1, f*b)的一阶梯度直方图和二阶梯度直方图,这里b为分桶数量。最后将一阶梯度直方图和二阶梯度直方图按第一维拼接得到多方安全计算结果,即最终的Host方直方图,shape为(2,f*b),并同步至发起方。

[0160] 从而,本实施例通过随机选取聚类中心的方式进行数据压缩,不仅在实现压缩数据的同时,还能够保证了数据的安全性,进而降低发起方和参与方的数据传输量,提高了计算效率。

[0161] 通过上述过程，可以降低训练过程中数据量，联邦学习xgboost主要瓶颈在于Host方直方图的计算，之所以为瓶颈是因为保证Guest方梯度矩阵和Host方特征数据矩阵不被泄露的情况下进行矩阵乘法，原始做法是将Host方特征数据矩阵分桶稀疏化然后直接与Guest方进行mpc乘法，但这种方式会有巨大的数据传输开销，不可在大数据样本进行执行。而本方案在保证数据不泄露的前提下大大下降数据传输量。

[0162] 正如前述所说，当前的联邦学习XGBoost算法训练的通信量比较大，导致训练耗时过长，不能满足业界需求。因此，本实施例还提供一种联邦学习模型的训练方法，如图9所示，该训练方法包括：

[0163] S1、分别获取发起方和参与方的求交特征数据集。

[0164] S2、利用求交特征数据集构建XGBoost树模型。

[0165] S3、通过如上的多方安全计算方法得到的聚合矩阵以计算XGBoost树模型的最优分割点。

[0166] S4、利用最优分割点更新XGBoost树模型。

[0167] S5、将待预测的数据输入至更新后的XGBoost树模型进行预测，得到预测结果。

[0168] 在一个例子中，如图10a和图10b所示，Guest发起方根据uid获取Guest方求交特征数据集，接收Host参与方的特征基本信息，生成随机种子并同步给Host方，开始初始化预测值p为0，并判断构建的树是否达到指定数量，若是则随机采样训练样本和训练特征，以计算一阶梯度和二阶梯度，随后判断是否达到树构建停止条件，若是则初始化Guest的特征直方图histo，并计算Guest方特征数据的分桶边界数值，计算Guest方本地直方图g_hist，同时接收特征聚类索引，接收host方发来的<histo1>，根据特征聚类索引计算不同特征的聚合一阶梯度clu_g和二阶梯度clu_h，将clu_g分片为(<clu_g1>,<clu_g2>)，将clu_h分片为(<clu_h1>,<clu_h2>)，发送<clu_g2>和<clu_h2>给host方，将<clu_g><clu_h>分别和<histo>进行mpc矩阵点乘得到<sum_g1><sum_h1>，shape为(k,f*b)，k为聚类中心数，f为特征数，b为分桶数，对<sum_g><sum_h>每个特征同桶聚类进行mpc求和得到新的<sum_g1><sum_h1>，shape为(1,f*b)，接收host方发送过来的<sum_g2>和<sum_h2>，恢复生成sum_g和sum_h，sum_g和sum_h进行拼接得到h_histo，shape为(2,f*b)，进而获得Guest和Host所有histo内容，根据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，发送给Host节点分裂信息，更新树结构，发送给Host下一level的节点信息，最后利用新树预测原始数据，更新p值。

[0169] 与上述Guest发起方进行步骤相对应的Host参与方的步骤如下，Host参与方根据uid获取host方求交特征数据集，发送给Guest特征基本信息，接收Guest的随机种子，并判断构建的树是否达到指定数量，若是则随机采样训练样本和训练特征，随后判断是否达到树构建停止条件，若否则对每个特征随机选取聚类中心，将样本特征聚类到各自特征的聚类中心上，发送特征聚类索引，初始化Host的特征直方图histo，shape为(k,f*b)，k为聚类中心数，f为特征数，b为分桶数，将histo分片为(<histo1><histo2>)，发送<histo1>给guest方，接收guest发送过来的<clu_g2>和<clu_h2>，将<clu_g><clu_h>分别和<histo>进行mpc矩阵乘法得到<sum_g2><sum_h2>，shape为(k,f*b)，k为聚类中心数，f为特征数，b为分桶数，对<sum_g><sum_h>每个特征同桶聚类求和得到新的<sum_g2><sum_h2>，shape为(1,f*b)，发送<sum_g2>和<sum_h2>给guest方，接收Guest节点分裂信息，以更新树结构，接

收Guest下一level的节点信息，最后用新树预测原始数据。

[0170] 在现有的XGBoost算法训练过程中最大的瓶颈在计算Host直方图，Guest方的梯度矩阵和Host的特征矩阵都不能相互泄露，而通过本实施例的多方安全计算方法所得到的Host直方图，不仅可以保证数据不相互泄露，又可以保证多方安全计算机制在精度没有损失的前提下，大幅度降低计算耗时。

[0171] 与上面介绍的联邦学习模型的训练方法对应地，本实施例还提供了一种联邦学习模型的训练系统。下面将分别进行介绍。具体地，如图11所示，本实施例还提供一种联邦学习模型的训练系统，该系统包括：

[0172] 数据集获取模块101，用于分别获取发起方和参与方的求交特征数据集。

[0173] 模型构建模块102，用于利用求交特征数据集构建XGBoost树模型。

[0174] 分割点计算模块103，用于通过如上的多方安全计算系统得到的聚合矩阵以计算XGBoost树模型的最优分割点。

[0175] 模型更新模块104，用于利用最优分割点更新XGBoost树模型。

[0176] 模型预测模块105，用于将待预测的数据输入至更新后的XGBoost树模型进行预测，得到预测结果。

[0177] 在现有的XGBoost算法训练过程中最大的瓶颈在计算Host直方图，Guest方的梯度矩阵和Host的特征矩阵都不能相互泄露，而通过本实施例的多方安全计算方法所得到的聚合矩阵(Host直方图)，不仅可以保证数据不相互泄露，又可以通过混合态多方安全计算机制在精度没有损失的前提下，大幅度降低计算耗时。

[0178] 还需要说明的是，本实施例的多方安全计系统或联邦学习模型的训练系统，例如可以是：单独的芯片、芯片模组或电子设备，也可以是集成于电子设备内的芯片或者芯片模组。关于上述实施例中描述的各个装置、产品包含的各个模块/单元，其可以是软件模块/单元，也可以是硬件模块/单元，或者也可以部分是软件模块/单元，部分是硬件模块/单元。例如，对于应用于或集成于芯片的各个装置、产品，其包含的各个模块/单元可以都采用电路等硬件的方式实现，或者，至少部分模块/单元可以采用软件程序的方式实现，该软件程序运行于芯片内部集成的处理器，剩余的(如果有)部分模块/单元可以采用电路等硬件方式实现；对于应用于或集成于芯片模组的各个装置、产品，其包含的各个模块/单元可以都采用电路等硬件的方式实现，不同的模块/单元可以位于芯片模组的同一组件(例如芯片、电路模块等)或者不同组件中，或者，至少部分模块/单元可以采用软件程序的方式现，该软件程序运行于芯片模组内部集成的处理器，剩余的(如果有)部分模块/单元可以采用电路等硬件方式实现；对于应用于或集成于终端的各个装置、产品，其包含的各个模块/单元可以都采用电路等硬件的方式实现，不同的模块/单元可以位于终端内同一组件(例如，芯片、电路模块等)或者不同组件中，或者，至少部分模块/单元可以采用软件程序的方式实现，该软件程序运行于终端内部集成的处理器，剩余的(如果有)部分模块/单元可以采用电路等硬件方式实现。

[0179] 图12为本实施例提供的一种电子设备的结构示意图。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，处理器执行程序时实现上述实施例中的方法。图12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。

[0180] 如图12所示，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。

[0181] 总线33包括数据总线、地址总线和控制总线。

[0182] 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。

[0183] 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。

[0184] 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，例如本发明如上所述的方法。

[0185] 电子设备30也可以与一个或多个外部设备34(例如键盘、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12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。

[0186] 应当注意，尽管在上文详细描述中提及了电子设备的若干单元/模块或子单元/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单元/模块的特征和功能可以在一个单元/模块中具体化。反之，上文描述的一个单元/模块的特征和功能可以进一步划分为由多个单元/模块来具体化。

[0187] 本实施例还提供了一种计算机可读存储介质，其上存储有计算机程序，程序被处理器执行时实现如上述实施例的方法中的步骤。

[0188] 其中，可读存储介质可以采用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。

[0189] 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，当程序产品在终端设备上运行时，程序代码用于使终端设备执行实现如上所述的方法中的步骤。

[0190] 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，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。

[0191] 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。

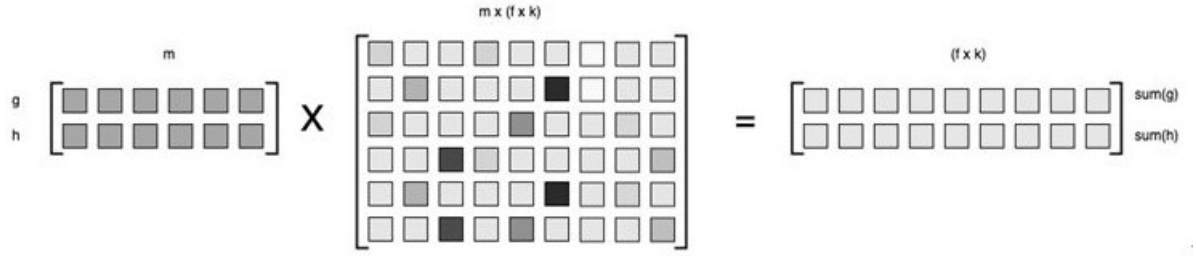

m x (f x k)
m
g
h
X
=
(f x k)
sum(g)
sum(h)

图 1

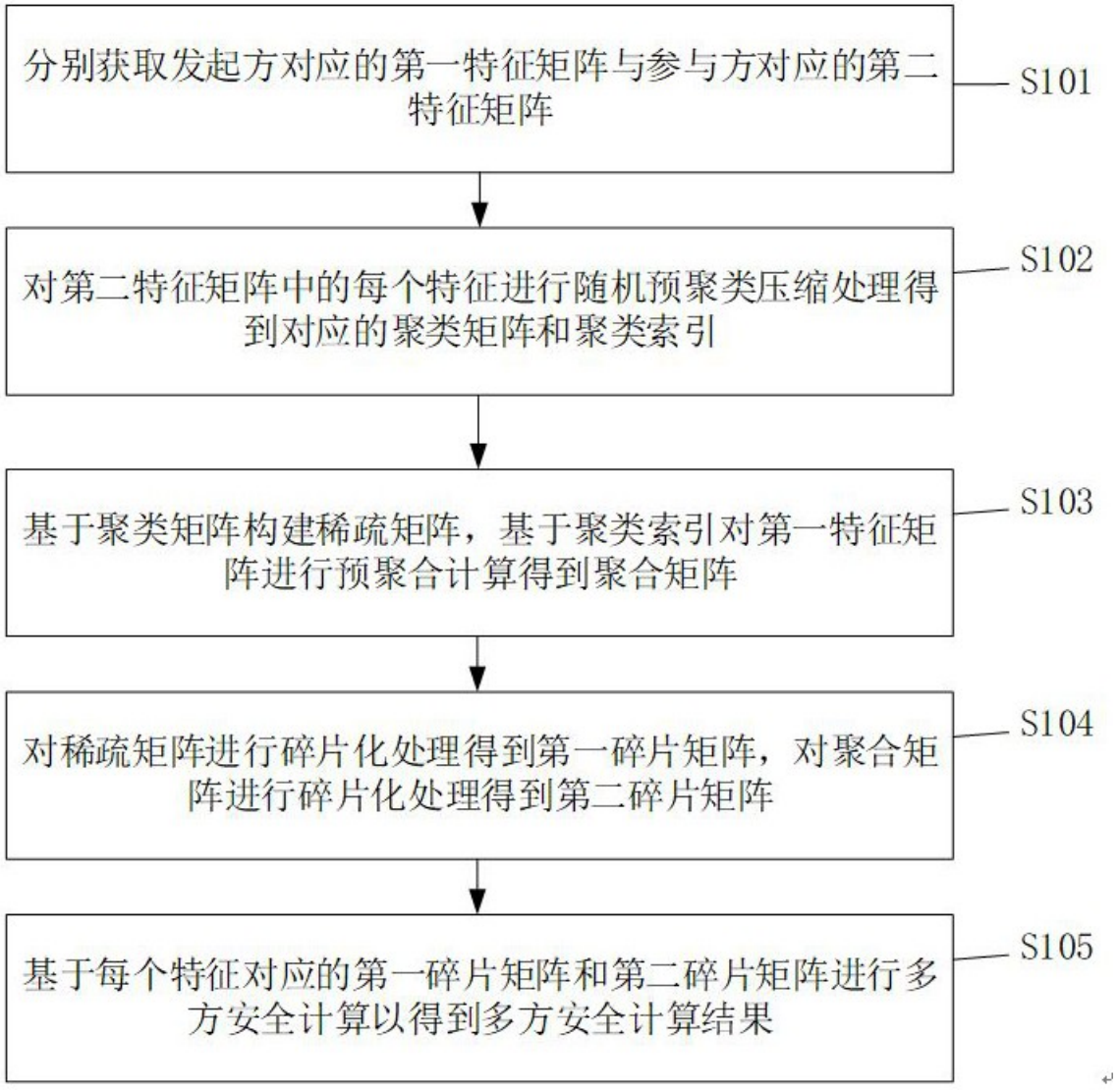

分别获取发起方对应的第一特征矩阵与参与方对应的第二特征矩阵
S101
对第二特征矩阵中的每个特征进行随机预聚类压缩处理得到对应的聚类矩阵和聚类索引
S102
基于聚类矩阵构建稀疏矩阵，基于聚类索引对第一特征矩阵进行预聚合计算得到聚合矩阵
S103
对稀疏矩阵进行碎片化处理得到第一碎片矩阵，对聚合矩阵进行碎片化处理得到第二碎片矩阵
S104
基于每个特征对应的第一碎片矩阵和第二碎片矩阵进行多方安全计算以得到多方安全计算结果
S105

图 2

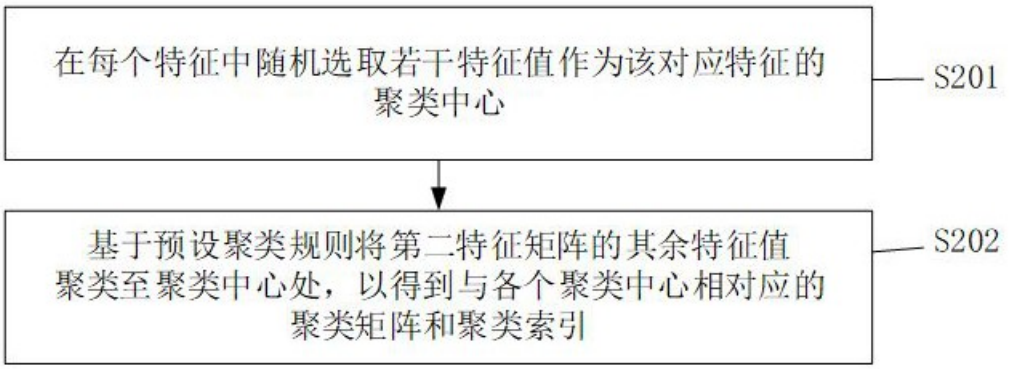

在每个特征中随机选取若干特征值作为该对应特征的聚类中心
S201
基于预设聚类规则将第二特征矩阵的其余特征值聚类至聚类中心处，以得到与各个聚类中心相对应的聚类矩阵和聚类索引
S202

图 3

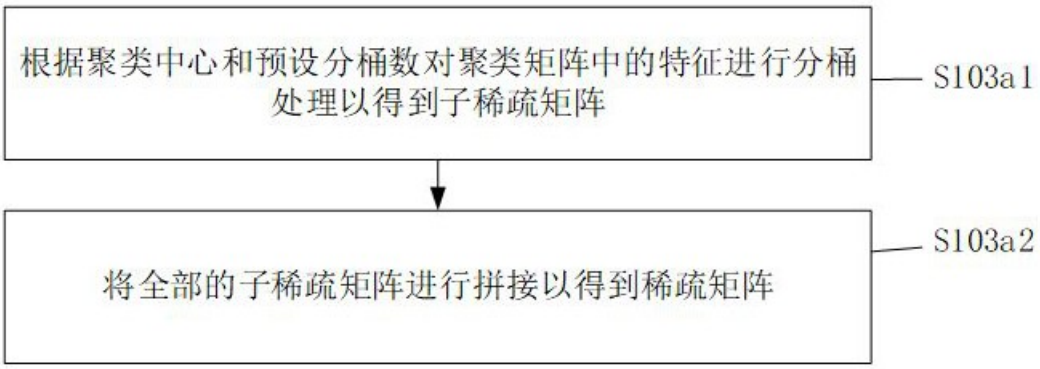

根据聚类中心和预设分桶数对聚类矩阵中的特征进行分桶处理以得到子稀疏矩阵
S103a1
将全部的子稀疏矩阵进行拼接以得到稀疏矩阵
S103a2

图 4

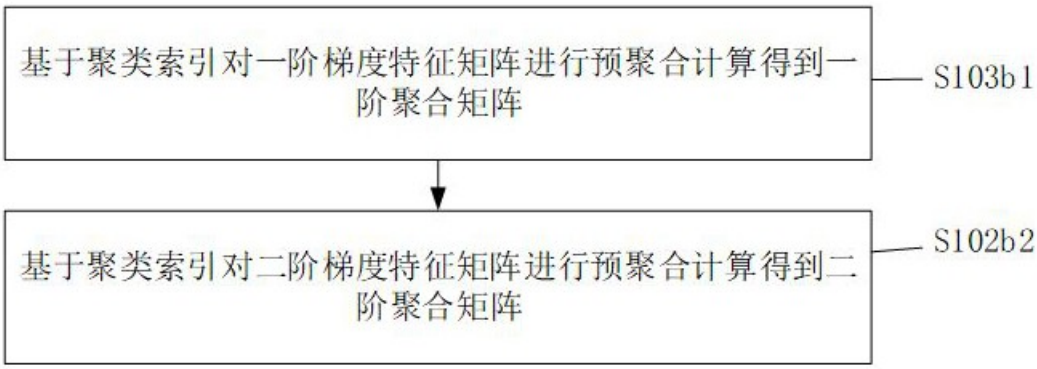

基于聚类索引对一阶梯度特征矩阵进行预聚合计算得到一阶聚合矩阵
S103b1
基于聚类索引对二阶梯度特征矩阵进行预聚合计算得到二阶聚合矩阵
S102b2

图 5

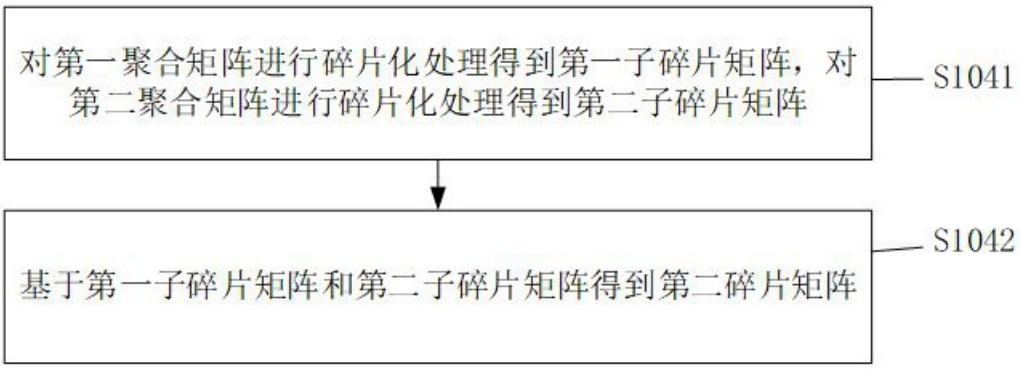

对第一聚合矩阵进行碎片化处理得到第一子碎片矩阵，对第二聚合矩阵进行碎片化处理得到第二子碎片矩阵
S1041
基于第一子碎片矩阵和第二子碎片矩阵得到第二碎片矩阵
S1042

图 6

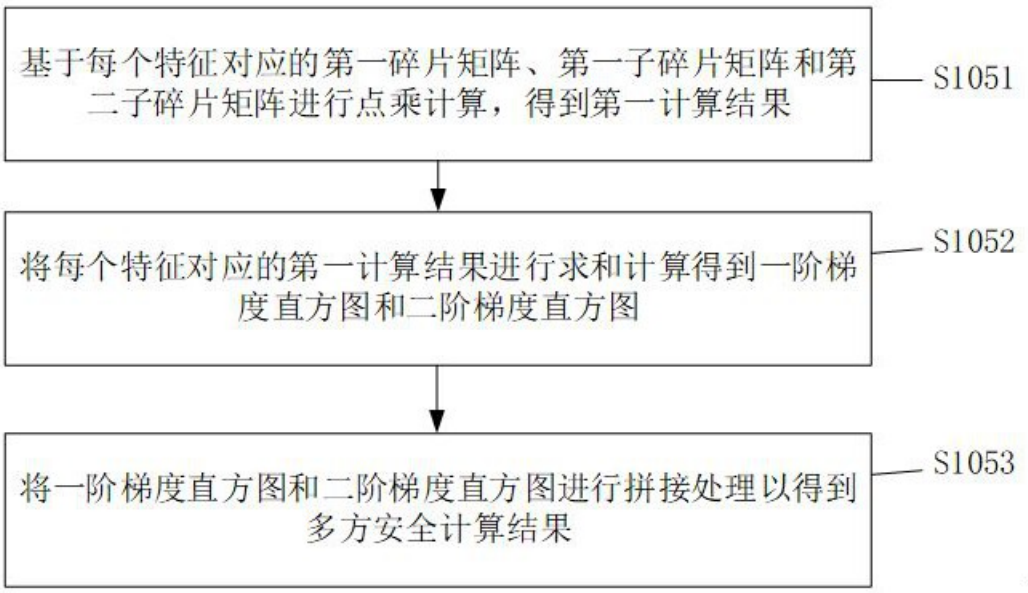

基于每个特征对应的第一碎片矩阵、第一子碎片矩阵和第二子碎片矩阵进行点乘计算，得到第一计算结果
S1051
将每个特征对应的第一计算结果进行求和计算得到一阶梯度直方图和二阶梯度直方图
S1052
将一阶梯度直方图和二阶梯度直方图进行拼接处理以得到多方安全计算结果
S1053

图 7

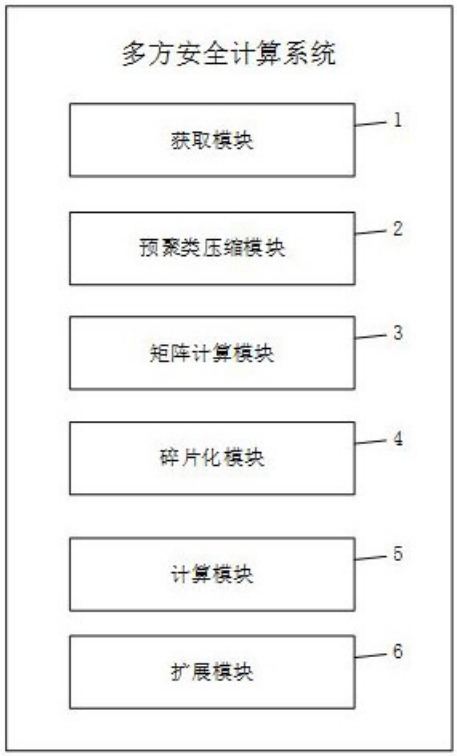

多方安全计算系统
获取模块
1
预聚类压缩模块
2
矩阵计算模块
3
碎片化模块
4
计算模块
5
扩展模块
6

图 8

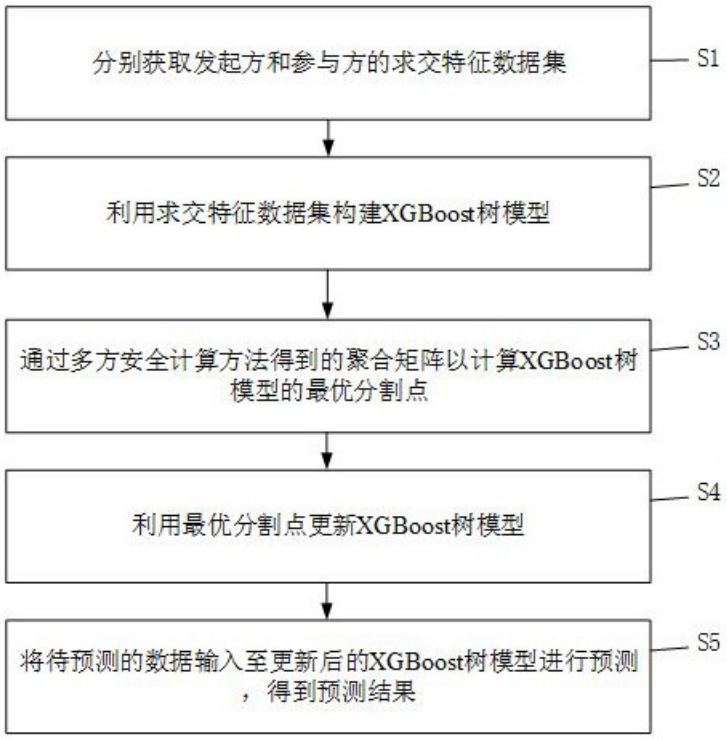

分别获取发起方和参与方的求交特征数据集
S1
利用求交特征数据集构建XGBoost树模型
S2
通过多方安全计算方法得到的聚合矩阵以计算XGBoost树模型的最优分割点
S3
利用最优分割点更新XGBoost树模型
S4
将待预测的数据输入至更新后的XGBoost树模型进行预测，得到预测结果
S5

图 9

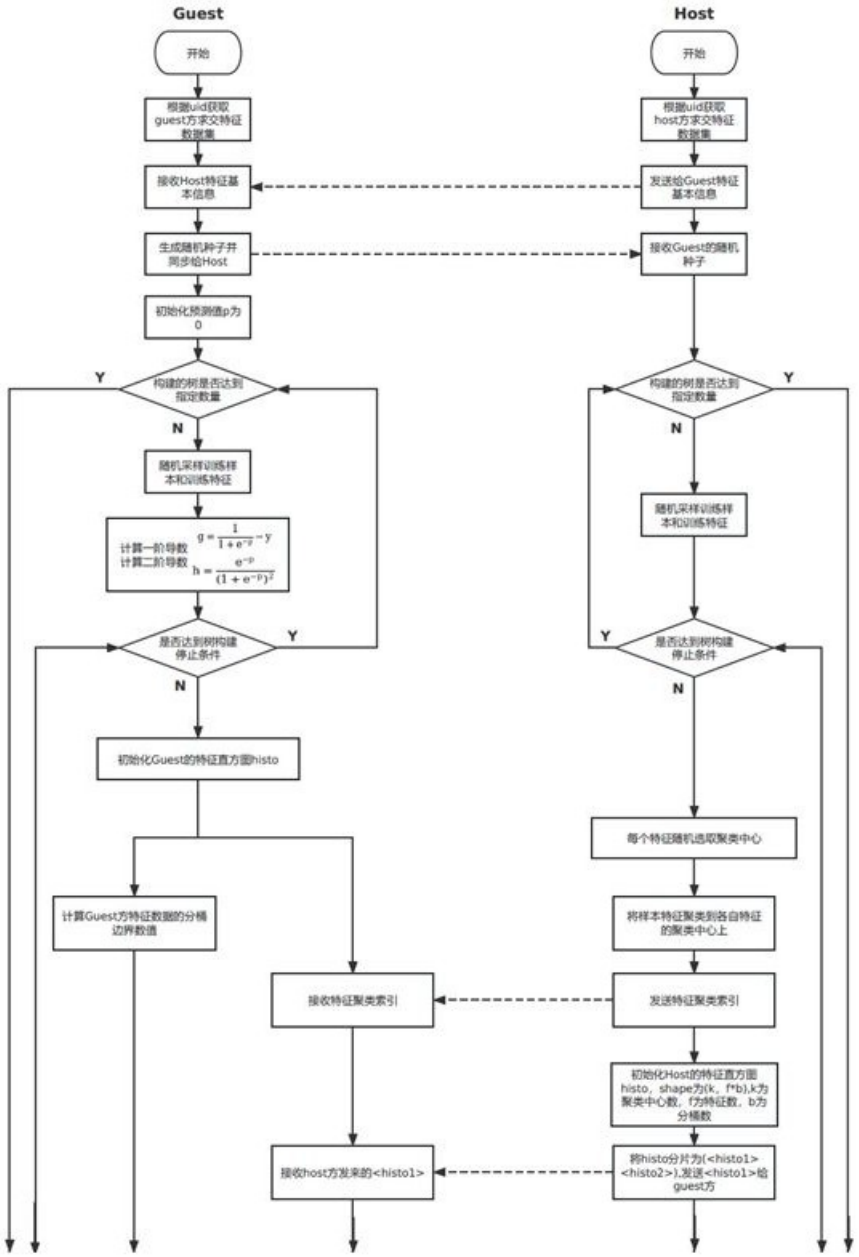

Guest
开始
根据uid获取guest方求交特征数据集
接收Host特征基本信息
生成随机种子并同步给Host
初始化预测值p为0
构建的树是否达到指定数量
Y
N
随机采样训练样本和训练特征
计算一阶导数 $g=\frac{1}{1+e^{-p}}-y$
计算二阶导数 $h=\frac{e^{-p}}{(1+e^{-p})^2}$
是否达到树构建停止条件
Y
N
初始化Guest的特征直方图histo
计算Guest方特征数据的分桶边界数值
接收特征聚类索引
接收host方发来的<histo1>
Host
开始
根据uid获取host方求交特征数据集
发送给Guest特征基本信息
接收Guest的随机种子
构建的树是否达到指定数量
Y
N
随机采样训练样本和训练特征
是否达到树构建停止条件
Y
N
每个特征随机选取聚类中心
将样本特征聚类到各自特征的聚类中心上
发送特征聚类索引
初始化Host的特征直方图histo，shape为(k，f*b),k为聚类中心数，f为特征数，b为分桶数
将histo分片为(<histo1><histo2>),发送<histo1>给guest方

图 10a

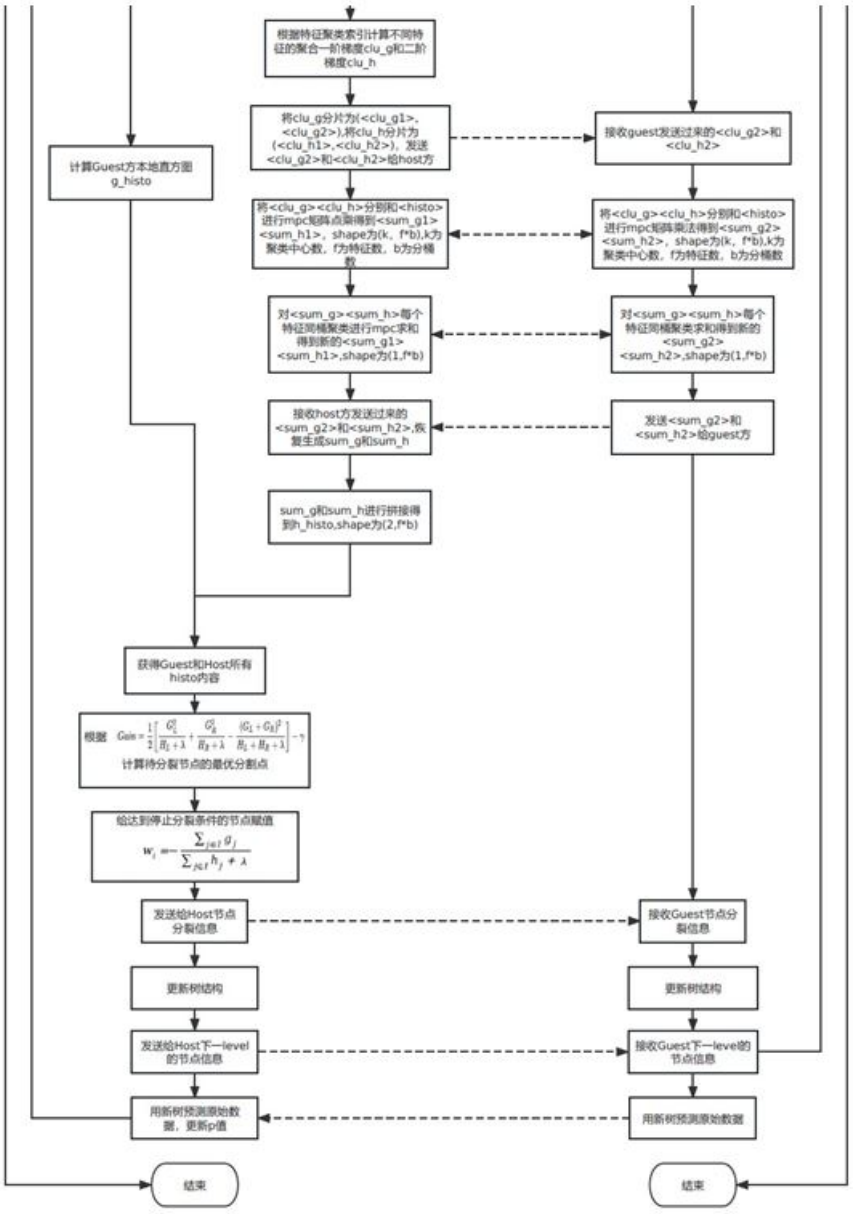

根据特征聚类索引计算不同特征的聚合一阶梯度clu_g和二阶梯度clu_h
将clu_g分片为(<clu_g1>,<clu_g2>),将clu_h分片为(<clu_h1>,<clu_h2>)，发送<clu_g2>和<clu_h2>给host方
接收guest发送过来的<clu_g2>和<clu_h2>
计算Guest方本地直方图g_histo
将<clu_g><clu_h>分别和<histo>进行mpc矩阵点乘得到<sum_g1><sum_h1>，shape为(k，f*b),k为聚类中心数，f为特征数，b为分桶数
将<clu_g><clu_h>分别和<histo>进行mpc矩阵乘法得到<sum_g2><sum_h2>，shape为(k，f*b),k为聚类中心数，f为特征数，b为分桶数
对<sum_g><sum_h>每个特征同桶聚类进行mpc求和得到新的<sum_g1><sum_h1>,shape为(1,f*b)
对<sum_g><sum_h>每个特征同桶聚类求和得到新的<sum_g2><sum_h2>,shape为(1,f*b)
接收host方发送过来的<sum_g2>和<sum_h2>,恢复生成sum_g和sum_h
发送<sum_g2>和<sum_h2>给guest方
sum_g和sum_h进行拼接得到h_histo,shape为(2,f*b)
获得Guest和Host所有histo内容
根据 $Gain=\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right]-\gamma$ 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 $w_i=-\frac{\sum_{j\in I}g_j}{\sum_{j\in I}h_j+\lambda}$
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leve的节点信息
用新树预测原始数据，更新p值
用新树预测原始数据
结束
结束

图 10b

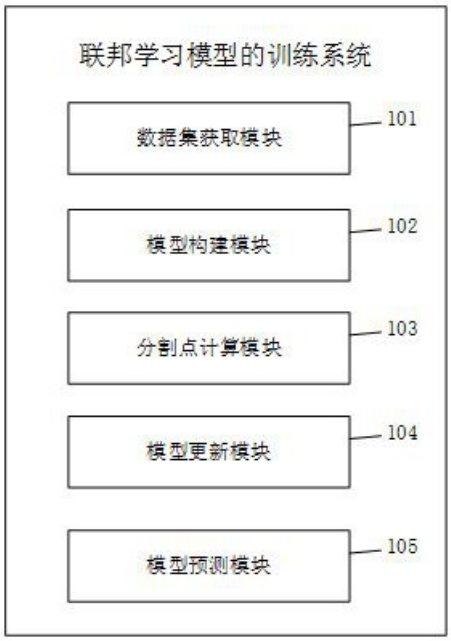

联邦学习模型的训练系统
数据集获取模块
101
模型构建模块
102
分割点计算模块
103
模型更新模块
104
模型预测模块
105

图 11

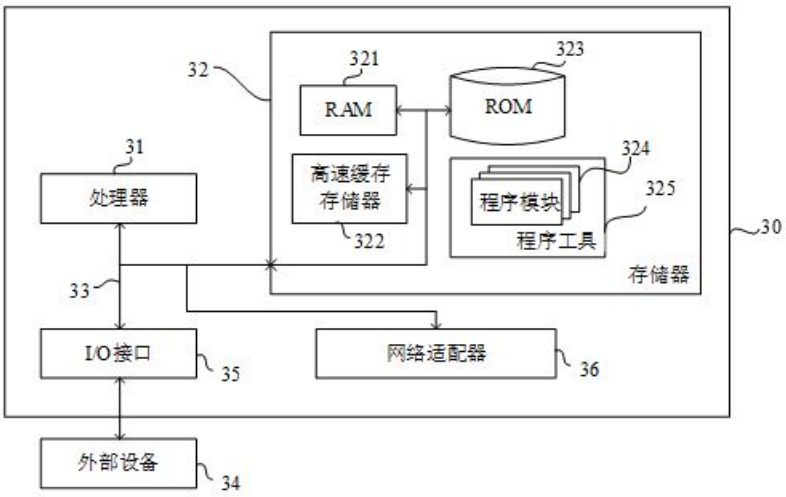

32
321
RAM
323
ROM
31
处理器
高速缓存
存储器
322
324
程序模块
325
程序工具
存储器
30
33
I/O接口
35
网络适配器
36
外部设备
34

图 12