(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117648992 A
(43)申请公布日 2024.03.05

(21)申请号 202311436932.6

(22)申请日 2023.10.31

(66)本国优先权数据
202310348364.8 2023.04.04 CN

(71)申请人 富算科技(上海)有限公司
地址 200120 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 王兆凯 陈立峰 赵华宇 卞阳 张伟奇

(74)专利代理机构 北京慧加伦知识产权代理有限公司 16035
专利代理师 郝聪慧

(51)Int.Cl.
G06N 20/00(2019.01)
G06F 18/232(2023.01)

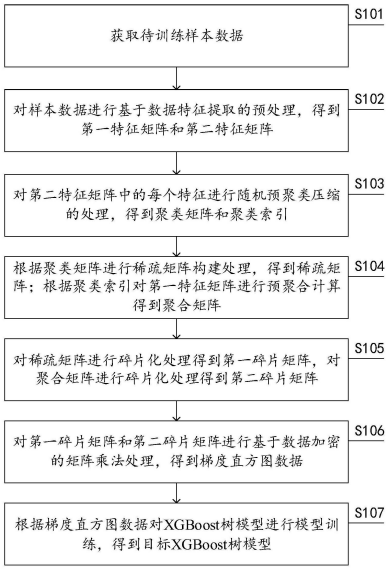
权利要求书3页 说明书11页 附图7页

(54)发明名称

用于XGBoost联邦学习模型训练的数据处理方法和装置

(57)摘要

本申请公开了一种用于XGBoost联邦学习模型训练的数据处理方法和装置。应用于联邦学习场景,获取多方的训练样本数据,根据多方样本数据联邦训练XGBoost树模型;对样本数据进行基于数据特征的聚类压缩得到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;基于第一碎片矩阵和第二碎片矩阵进行基于数据加密的矩阵乘法处理,同桶特征聚类中心求和,得到梯度直方图数据;根据梯度直方图数据进行模型训练得到目标XGBoost树模型。通过对特征进行聚类压缩实现在稀疏矩阵上的运算加速,降低模型训练过程中的数据计算量及传输量,提高效率。



CN 117648992 A

1.一种用于XGBoost联邦学习模型训练的数据处理方法,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,所述数据处理方法包括:

获取待训练样本数据,其中,所述待训练样本数据为所述至少一个发起方和所述至少一个参与方的样本数据;

对所述样本数据进行基于数据特征提取的预处理,得到第一特征矩阵和第二特征矩阵,其中,所述第一特征矩阵为用于表示发起方样本数据特征的矩阵,所述第二特征矩阵为用于表示参与方样本数据特征对应的矩阵;

对所述第二特征矩阵中的每个特征进行随机预聚类压缩的处理,得到聚类矩阵和聚类索引;

根据所述聚类矩阵进行稀疏矩阵构建处理,得到稀疏矩阵;根据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;

对所述第一碎片矩阵和所述第二碎片矩阵进行基于数据加密的矩阵乘法处理,得到梯度直方图数据;

根据所述梯度直方图数据对XGBoost树模型进行模型训练,得到目标XGBoost树模型。

2.根据所述权利要求1所述的数据处理方法,其特征在于,对所述第二特征矩阵中的每个特征进行随机预聚类压缩的处理,得到聚类矩阵和聚类索引包括:

在所述第二特征矩阵中的每个特征中随机选取预设数量的特征值作为特征对应的聚类中心,并得到待聚类特征数据,其中,所述待聚类特征数据为所述第二特征矩阵中除聚类中心外的特征值;

根据所述聚类中心和预设聚类规则对所述待聚类特征数据进行聚类处理,得到与所述聚类中心对应的聚类矩阵和聚类索引。

3.根据权利要求1所述的数据处理方法,其特征在于,根据所述聚类矩阵进行稀疏矩阵构建处理,得到稀疏矩阵包括:

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,得到子稀疏矩阵;

将所述子稀疏矩阵进行拼接处理,得到所述稀疏矩阵。

4.根据权利要求1所述的数据处理方法,其特征在于,根据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:

对所述第一特征矩阵进行识别处理,得到一阶梯度矩阵和二阶梯度矩阵;

根据所述聚类索引对所述一阶梯度矩阵进行预聚合处理,得到一阶聚合矩阵;

根据所述聚类索引对所述二阶梯度矩阵进行预聚合处理,得到二阶聚合矩阵;

对所述一阶聚合矩阵和所述二阶聚合矩阵进行组合优化处理,得到所述聚合矩阵。

5.根据权利要求4所述的数据处理方法,其特征在于,对所述一阶聚合矩阵和所述二阶聚合矩阵进行组合优化处理,得到所述聚合矩阵包括:

对一阶聚合矩阵进行扩展处理,得到第一聚合矩阵;对二阶聚合矩阵进行扩展处理,得到第二聚合矩阵;

对所述第一聚合矩阵和所述第二聚合矩阵进行梯度合并的组合处理,得到过程聚合矩

阵数据;

对所述过程聚合矩阵数据进行密态化处理,得到所述聚合矩阵。

6.根据权利要求1所述的数据处理方法,其特征在于,对所述第一碎片矩阵和所述第二碎片矩阵进行基于数据加密的矩阵乘法处理,得到梯度直方图数据包括:

基于每个特征对应的第一碎片矩阵和第二碎片矩阵进行点乘处理,得到过程梯度直方图数据;

对所述过程梯度直方图数据进行基于同桶特征聚类中心求和的梯度直方图计算处理,得到一阶梯度直方图数据和二阶梯度直方图数据;

对所述一阶梯度直方图数据和所述二阶梯度直方图数据进行拼接处理,得到所述梯度直方图数据。

7.根据权利要求1所述的数据处理方法,其特征在于,根据所述梯度直方图数据对XGBoost树模型进行模型训练,得到目标XGBoost树模型包括:

根据所述梯度直方图数据进行XGBoost树模型的最优分割点计算处理,得到最优分割点数据,其中,所述最优分割点数据为用于表示XGBoost树模型最优分割点的数据;

根据所述最优分割点数据对所述XGBoost树模型进行基于树结构更新的模型训练处理,得到所述目标XGBoost树模型。

8.根据权利要求1所述的数据处理方法,其特征在于,在根据所述梯度直方图数据对XGBoost树模型进行模型训练,得到目标XGBoost树模型之后,所述数据处理方法包括:

获取待预测样本数据,其中,所述待预测样本数据为所述至少一个发起方和所述至少一个参与方需要进行样本预测的数据;

根据所述的目标XGBoost树模型对所述待预测样本数据进行预测处理,得到预测结果数据。

9.一种用于XGBoost联邦学习模型训练的数据处理装置,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,所述数据处理装置包括:

训练样本获取模块,用于获取待训练样本数据,其中,所述待训练样本数据为所述至少一个发起方和所述至少一个参与方的样本数据;

预处理模块,用于对所述样本数据进行基于数据特征提取的预处理,得到第一特征矩阵和第二特征矩阵,其中,所述第一特征矩阵为用于表示发起方样本数据特征的矩阵,所述第二特征矩阵为用于表示参与方样本数据特征对应的矩阵;

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的处理,得到聚类矩阵和聚类索引;

矩阵计算模块,用于根据所述聚类矩阵进行稀疏矩阵构建处理,得到稀疏矩阵;根据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;

梯度直方图计算模块,用于对所述第一碎片矩阵和所述第二碎片矩阵进行基于数据加密的矩阵乘法处理,得到梯度直方图数据;

模型训练模块,用于根据所述梯度直方图数据对XGBoost树模型进行模型训练,得到目标XGBoost树模型。

10.一种计算机可读存储介质,其特征在于,所述计算机可读存储介质存储有计算机指令,所述计算机指令用于使所述计算机执行权利要求1-8任意一项所述的用于XGBoost联邦学习模型训练的数据处理方法。

用于XGBoost联邦学习模型训练的数据处理方法和装置

技术领域

[0001] 本申请涉及计算机领域,具体而言,涉及一种用于XGBoost联邦学习模型训练的数据处理方法和装置。

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。

[0005] 发明人发现,在进行联邦学习的多方XGBoost模型训练时,为了保证数据隐私安全,采用对发起方特征数据矩阵分桶稀疏化然后直接与参与方进行多方安全计算乘法,在计算过程中产生大量的数据计算开销,且效率较低,难以在大数据样本场景下执行。

[0006] 因此,在机器学习时,如何在确保数据隐私安全的情况下,同时降低模型训练过程中数据计算开销,成为本领域技术人员亟待解决的问题。

发明内容

[0007] 本申请的主要目的在于提供一种用于XGBoost联邦学习模型训练的数据处理方法和装置,以解决现有技术如何在确保数据隐私安全的情况下,同时降低模型训练过程中数据计算开销的问题,降低模型训练过程中的数据计算开销,提高模型训练效率。

[0008] 为了实现上述目的,本申请的第一方面,提出了一种用于XGBoost联邦学习模型训练的数据处理方法,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,所述数据处理方法包括:

[0009] 获取待训练样本数据,其中,所述待训练样本数据为所述至少一个发起方和所述至少一个参与方的样本数据;

[0010] 对所述样本数据进行基于数据特征提取的预处理,得到第一特征矩阵和第二特征矩阵,其中,所述第一特征矩阵为用于表示发起方样本数据特征的矩阵,所述第二特征矩阵为用于表示参与方样本数据特征对应的矩阵;

[0011] 对所述第二特征矩阵中的每个特征进行随机预聚类压缩的处理,得到聚类矩阵和聚类索引;

[0012] 根据所述聚类矩阵进行稀疏矩阵构建处理,得到稀疏矩阵;根据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;

[0013] 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;

[0014] 对所述第一碎片矩阵和所述第二碎片矩阵进行基于数据加密的矩阵乘法处理,得到梯度直方图数据;

[0015] 根据所述梯度直方图数据对XGBoost树模型进行模型训练,得到目标XGBoost树模型。

[0016] 进一步地,对所述第二特征矩阵中的每个特征进行随机预聚类压缩的处理,得到聚类矩阵和聚类索引包括:

[0017] 在所述第二特征矩阵中的每个特征中随机选取预设数量的特征值作为特征对应的聚类中心,并得到待聚类特征数据,其中,所述待聚类特征数据为所述第二特征矩阵中除聚类中心外的特征值;

[0018] 根据所述聚类中心和预设聚类规则对所述待聚类特征数据进行聚类处理,得到与所述聚类中心对应的聚类矩阵和聚类索引。

[0019] 进一步地,根据所述聚类矩阵进行稀疏矩阵构建处理,得到稀疏矩阵包括:

[0020] 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,得到子稀疏矩阵;

[0021] 将所述子稀疏矩阵进行拼接处理,得到所述稀疏矩阵。

[0022] 进一步地,根据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:

[0023] 对所述第一特征矩阵进行识别处理,得到一阶梯度矩阵和二阶梯度矩阵;

[0024] 根据所述聚类索引对所述一阶梯度矩阵进行预聚合处理,得到一阶聚合矩阵;

[0025] 根据所述聚类索引对所述二阶梯度矩阵进行预聚合处理,得到二阶聚合矩阵;

[0026] 对所述一阶聚合矩阵和所述二阶聚合矩阵进行组合优化处理,得到所述聚合矩阵。

[0027] 进一步地,对所述一阶聚合矩阵和所述二阶聚合矩阵进行组合优化处理,得到所述聚合矩阵包括:

[0028] 对一阶聚合矩阵进行扩展处理,得到第一聚合矩阵;对二阶聚合矩阵进行扩展处理,得到第二聚合矩阵;

[0029] 对所述第一聚合矩阵和所述第二聚合矩阵进行梯度合并的组合处理,得到过程聚合矩阵数据;

[0030] 对所述过程聚合矩阵数据进行密态化处理,得到所述聚合矩阵。

[0031] 进一步地,对所述第一碎片矩阵和所述第二碎片矩阵进行基于数据加密的矩阵乘法处理,得到梯度直方图数据包括:

[0032] 基于每个特征对应的第一碎片矩阵和第二碎片矩阵进行点乘处理,得到过程梯度直方图数据;

[0033] 对所述过程梯度直方图数据进行基于同桶特征聚类中心求和的梯度直方图计算处理,得到一阶梯度直方图数据和二阶梯度直方图数据;

[0034] 对所述一阶梯度直方图数据和所述二阶梯度直方图数据进行拼接处理,得到所述梯度直方图数据。

[0035] 进一步地,根据所述梯度直方图数据对XGBoost树模型进行模型训练,得到目标XGBoost树模型包括:

[0036] 根据所述梯度直方图数据进行XGBoost树模型的最优分割点计算处理,得到最优分割点数据,其中,所述最优分割点数据为用于表示XGBoost树模型最优分割点的数据;

[0037] 根据所述最优分割点数据对所述XGBoost树模型进行基于树结构更新的模型训练处理,得到所述目标XGBoost树模型。

[0038] 进一步地,在根据所述梯度直方图数据对XGBoost树模型进行模型训练,得到目标XGBoost树模型之后,所述数据处理方法包括:

[0039] 获取待预测样本数据,其中,所述待预测样本数据为所述至少一个发起方和所述至少一个参与方需要进行样本预测的数据;

[0040] 根据所述的目标XGBoost树模型对所述待预测样本数据进行预测处理,得到预测结果数据。

[0041] 根据本申请的第二方面,提出了一种用于XGBoost联邦学习模型训练的数据处理装置,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,所述数据处理装置包括:

[0042] 训练样本获取模块,用于获取待训练样本数据,其中,所述待训练样本数据为所述至少一个发起方和所述至少一个参与方的样本数据;

[0043] 预处理模块,用于对所述样本数据进行基于数据特征提取的预处理,得到第一特征矩阵和第二特征矩阵,其中,所述第一特征矩阵为用于表示发起方样本数据特征的矩阵,所述第二特征矩阵为用于表示参与方样本数据特征对应的矩阵;

[0044] 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的处理,得到聚类矩阵和聚类索引;

[0045] 矩阵计算模块,用于根据所述聚类矩阵进行稀疏矩阵构建处理,得到稀疏矩阵;根据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;

[0046] 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;

[0047] 梯度直方图计算模块,用于对所述第一碎片矩阵和所述第二碎片矩阵进行基于数据加密的矩阵乘法处理,得到梯度直方图数据;

[0048] 模型训练模块,用于根据所述梯度直方图数据对XGBoost树模型进行模型训练,得到目标XGBoost树模型。

[0049] 根据本申请的第三方面,提出了一种计算机可读存储介质,所述计算机可读存储介质存储有计算机指令,所述计算机指令用于使所述计算机执行上述的用于XGBoost联邦学习模型训练的数据处理方法。

[0050] 根据本申请的第四方面,提出了一种电子设备,包括:至少一个处理器;以及与所述至少一个处理器通信连接的存储器;其中,所述存储器存储有可被所述至少一个处理器

执行的计算机程序,所述计算机程序被所述至少一个处理器执行,以使所述至少一个处理器执行上述的用于XGBoost联邦学习模型训练的数据处理方法。

[0051] 本申请的实施例提供的技术方案可以包括以下有益效果:

[0052] 在本申请中,通过在至少一个发起方与至少一个参与方之间数据共享场景中,获取待训练样本数据,其中,所述待训练样本数据为所述至少一个发起方和所述至少一个参与方的样本数据;对所述样本数据进行基于数据特征提取的预处理,得到第一特征矩阵和第二特征矩阵,其中,所述第一特征矩阵为用于表示发起方样本数据特征的矩阵,所述第二特征矩阵为用于表示参与方样本数据特征对应的矩阵;对所述第二特征矩阵中的每个特征进行随机预聚类压缩的处理,得到聚类矩阵和聚类索引;根据所述聚类矩阵进行稀疏矩阵构建处理,得到稀疏矩阵;根据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;对所述第一碎片矩阵和所述第二碎片矩阵进行基于数据加密的矩阵乘法处理,得到梯度直方图数据;根据所述梯度直方图数据对XGBoost树模型进行模型训练,得到目标XGBoost树模型。通过对联邦学习模型训练计算梯度直方图的过程中,对特征进行聚类压缩实现在稀疏矩阵上的运算加速,降低模型训练过程中的数据计算量及传输量等,在保证联邦学习数据隐私安全的情况下,降低模型训练过程中数据计算开销,提高模型训练的效率。

附图说明

[0053] 构成本申请的一部分的附图用来提供对本申请的进一步理解,使得本申请的其它特征、目的和优点变得更明显。本申请的示意性实施例附图及其说明用于解释本申请,并不构成对本申请的不当限定。在附图中:

[0054] 图1为现有XGBoost模型训练过程中在Host方的直方图碎片态矩阵计算方法;

[0055] 图2为本申请提供的一种用于XGBoost联邦学习模型训练的数据处理方法的流程图;

[0056] 图3为本申请提供的一种用于XGBoost联邦学习模型训练的数据处理方法的流程图;

[0057] 图4为本申请提供的一种用于XGBoost联邦学习模型训练的数据处理方法的流程图;

[0058] 图5为本申请提供的一种用于XGBoost联邦学习模型训练的数据处理方法的流程图;

[0059] 图6为本申请提供的一种用于XGBoost联邦学习模型训练的数据处理方法的流程图;

[0060] 图7a和7b为本申请实施例提供的一种用于XGBoost联邦学习模型训练的数据处理方法的流程示意图;

[0061] 图8为本申请提供的一种用于XGBoost联邦学习模型训练的数据处理装置的结构示意图。

具体实施方式

[0062] 为了使本技术领域的人员更好地理解本申请方案,下面将结合本申请实施例中的附图,对本申请实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本申请一部分的实施例,而不是全部的实施例。基于本申请中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都应当属于本申请保护的范围。

[0063] 需要说明的是,本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象,而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换,以便这里描述的本申请的实施例。此外,术语“包括”和“具有”以及他们的任何变形,意图在于覆盖不排他的包含,例如,包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元,而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

[0064] 在本申请中,术语“上”、“下”、“左”、“右”、“前”、“后”、“顶”、“底”、“内”、“外”、“中”、“竖直”、“水平”、“横向”、“纵向”等指示的方位或位置关系为基于附图所示的方位或位置关系。这些术语主要是为了更好地描述本申请及其实施例,并非用于限定所指示的装置、元件或组成部分必须具有特定方位,或以特定方位进行构造和操作。

[0065] 并且,上述部分术语除了可以用于表示方位或位置关系以外,还可能用于表示其他含义,例如术语“上”在某些情况下也可能用于表示某种依附关系或连接关系。对于本领域普通技术人员而言,可以根据具体情况理解这些术语在本申请中的具体含义。

[0066] 此外,术语“安装”、“设置”、“设有”、“连接”、“相连”、“套接”应做广义理解。例如,“连接”可以是固定连接,可拆卸连接,或整体式构造;可以是机械连接,或电连接;可以是直接相连,或者是通过中间媒介间接相连,又或者是两个装置、元件或组成部分之间内部的连通。对于本领域普通技术人员而言,可以根据具体情况理解上述术语在本申请中的具体含义。

[0067] 联邦学习是一种分布式计算框架,为了协同各方联合完成计算任务,参与方需要在多方安全协议调度下,传输计算过程中的中间结果。

[0068] 安全多方计算(MPC)是实现联邦学习的隐私保护方式之一,其相比半同态、差分隐私具有高性能且高精度的计算优势。在实现数据价值和保证数据不出门的前提下,安全多方计算可以在许多领域创造价值,如金融风控,医疗科研,广告推进等。并且在数据安全越来越被重视的当下环境,安全多方计算技术将会成为数据交互必不可少的底层技术之一。

[0069] Guest方是发起方(协调方)和标签拥有方,Host方是计算参与方,红色方框内为Guest和Host主要数据交互部分。Guest和Host都是数据拥有方,双方的数据对象有不同维度的特征。

[0070] 因此,在联邦学习系统中,参与方之间需要进行大量的数据通信来交互模型更新信息,举例说明,如假设一个数据集有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示出了一个现有XGBoost模型训练过程中在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,模型训练过程中产生大量数据传输量,数据计算开销较大;如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,模型训练过程耗时较长,效率较低。

[0071] 在本申请的可选实施例中,提出了一种用于XGBoost联邦学习模型训练的数据处理方法,通过在模型训练计算host方直方图的过程中,对特征进行聚类压缩实现在稀疏矩阵上的运算加速,降低模型训练过程中的数据计算量及传输量等,在保证联邦学习数据隐私安全的情况下,降低模型训练过程中数据计算开销,提高模型训练的效率,该用于XGBoost联邦学习模型训练的数据处理方法,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,图2为本申请提供的一种用于XGBoost联邦学习模型训练的数据处理方法的流程图,如图2所示,该方法包括以下步骤:

[0072] S101:获取待训练样本数据;

[0073] 待训练样本数据为至少一个发起方和至少一个参与方的样本数据;

[0074] 该XGBoost联邦学习模型应用于发起方与参与方之间的数据共享场景中,样本数据分别为发起方和参与方的样本数据,如,应用于患者理赔、投保风控等预测场景下,获取Guest方拥有的保险公司的投保数据,Host方拥有的医院的患者数据,根据双方用户数据求交集所训练的模型应用于上述患者理赔、投保风控场景;如应用于金融风控等场景下,根据各银行中的用户数据进行联邦学习XGBoost树模型构建,构建的模型应用于金融风控场景下;如应用在广告推荐,获取广告商、广告投放平台等的用户数据进行联邦学习XGBoost树模型构建,构建的模型应用于广告投放场景下,以实现在广告投放场景下进行预测。

[0075] S102:对样本数据进行基于数据特征提取的预处理,得到第一特征矩阵和第二特征矩阵;

[0076] 第一特征矩阵为用于表示发起方样本数据特征的矩阵,第二特征矩阵为用于表示参与方样本数据特征对应的矩阵;用于分别获取发起方和参与方的求交特征数据集,得到对应于发起方的第一特征矩阵和对应于参与方的第二特征矩阵。其中,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。

[0077] S103:对第二特征矩阵中的每个特征进行随机预聚类压缩的处理,得到聚类矩阵和聚类索引;

[0078] 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。

[0079] 在本申请的另一可选实施例中,提供了一种用于XGBoost联邦学习模型训练的数据处理方法,图3为本申请提供的一种用于XGBoost联邦学习模型训练的数据处理方法,如图3所示,该方法包括以下步骤:

[0080] S201:在第二特征矩阵中的每个特征中随机选取预设数量的特征值作为特征对应的聚类中心,并得到待聚类特征数据;

[0081] 其中,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,待聚类特征数据为第二特征矩阵中除聚类中心外的特征值。

[0082] S202:根据聚类中心和预设聚类规则对待聚类特征数据进行聚类处理,得到与聚类中心对应的聚类矩阵和聚类索引。

[0083] 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。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。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。

[0084] 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。

[0085] 由于预聚类中心点是随机选取,且随机聚类中心点数量远大于分桶数,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。

[0086] S104:根据聚类矩阵进行稀疏矩阵构建处理,得到稀疏矩阵;根据聚类索引对第一特征矩阵进行预聚合计算得到聚合矩阵;

[0087] 作为本实施例的一个可选地实施方式,在本申请的另一可选实施例中,提供了一种用于XGBoost联邦学习模型训练的数据处理方法,图4为本申请提供的一种用于XGBoost联邦学习模型训练的数据处理方法的流程图,如图4所示,该方法包括以下步骤:

[0088] S301: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,得到子稀疏矩阵。

[0089] S302:将子稀疏矩阵进行拼接处理,得到稀疏矩阵。

[0090] 将全部的子稀疏矩阵进行拼接以得到稀疏矩阵。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,如上述具体例子中的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)。

[0091] 作为本实施例的一个可选地实施方式,图5为本申请提供的一种用于XGBoost联邦学习模型训练的数据处理方法的流程图,如图5所示,该方法包括以下步骤:

[0092] S401:对第一特征矩阵进行识别处理,得到一阶梯度矩阵和二阶梯度矩阵;

[0093] S402:根据聚类索引对一阶梯度矩阵进行预聚合处理,得到一阶聚合矩阵;

[0094] S403:根据聚类索引对二阶梯度矩阵进行预聚合处理,得到二阶聚合矩阵;

[0095] 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)。

[0096] S404:对一阶聚合矩阵和二阶聚合矩阵进行组合优化处理,得到聚合矩阵。

[0097] 作为本实施例的一个可选地实施方式,提供了用于XGBoost联邦学习模型训练的

数据处理方法,包括:

[0098] 对一阶聚合矩阵进行扩展处理,得到第一聚合矩阵;对二阶聚合矩阵进行扩展处理,得到第二聚合矩阵。其中,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。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。

[0099] 对所述第一聚合矩阵和所述第二聚合矩阵进行梯度合并的组合处理,得到过程聚合矩阵数据;对所述过程聚合矩阵数据进行密态化处理,得到所述聚合矩阵。其中,通过梯度打包,将原先一阶梯度与二阶梯度两列合并为单列,再进行直方图的计算,将第一聚合矩阵对应的一阶梯度和第二聚合矩阵对应的二阶梯度组合到一起,形成一个数字进行密态化处理,每个数字用固定位数来进行保存,固定位数计算方式为:

[0100] 先计算一阶梯度g和二阶梯度h的求和后的最大可能范围,

[0101] $g_{imax}=n_i*(g_{max}+g_{off})*2^r$

[0102] $h_{imax}=n_i*h_{max}*2^r$

[0103] 用g和h求和的最大可能范围计算位数间隔,

[0104] $b_g=BitLength(g_{imax})$

[0105] $b_h=BitLength(h_{imax})$

[0106] 其中,上述公式中是n_i实例特征数量,g_{max}和h_{max}分别表示最大一阶梯度值和最大二阶梯度值。由于一阶梯度值的取值范围为[-1,1],二阶梯度的取值范围为[0,1],所以一阶梯度需要一个偏置量g_{off}用于将负数转正数,这个值的计算方式是abs(min(g))。梯度范围因子r可设置为53,即在float中实际使用到的位数个数。

[0107] 可以将单个样本的G和H进行拼接成一个对象GH,通过上述预估方式算出大致的数量级方式最终结果越位。一阶梯度g的取值范围为[-1,1],二阶梯度h的取值范围为[0,1]。64bit长度下,如果考虑浮点不损失精度情况,则小数取2**53表示,假设有1亿的训练数据,一个数值的位数估算为:

[0108] gimax=100000000*(1+1)*(2**53)=180143985094819840000000

[0109] himax=100000000*1*(2**53)=900719925474099200000000

[0110] $b_g=log2(gimax)=81$,一阶梯度需要81位,

[0111] $b_h=log2(himax)=80$,二阶梯度需要80位,

[0112] 因此合并和的数值范围控制在161个bit。

[0113] 对于将一阶梯度g和二阶梯度h进行拼接计算,计算gh,则gh=g*(2**r)+h,其中r为倍数扩大因子。

[0114] 在本申请实施例中,对于需要保持高精度计算的多方安全计算场景下,,如对于原始碎片随机数长度在128bit的场景,会带来通信量的降低,且减少通信次数,来达到模型训练效率及模型性能的效果。

[0115] S105: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;

[0116] 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。

[0117] S106:对第一碎片矩阵和第二碎片矩阵进行基于数据加密的矩阵乘法处理,得到梯度直方图数据;

[0118] 作为本实施例的一个可选地实施方式,图6为本申请提供的一种用于XGBoost联邦学习模型训练的数据处理方法的流程图,如图6所示,该方法包括以下步骤:

[0119] S501:基于每个特征对应的第一碎片矩阵和第二碎片矩阵进行点乘处理,得到过程梯度直方图数据;

[0120] S502:对过程梯度直方图数据进行基于同桶特征聚类中心求和的梯度直方图计算处理,得到一阶梯度直方图数据和二阶梯度直方图数据;

[0121] S503:对一阶梯度直方图数据和二阶梯度直方图数据进行拼接处理,得到梯度直方图数据。

[0122] 基于每个特征的第一碎片矩阵和第二碎片矩阵进行基于数据加密的矩阵乘法处理,并对同桶特征聚类中心求和,得到梯度直方图数据。

[0123] 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)。

[0124] 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。

[0125] S107:根据梯度直方图数据对XGBoost树模型进行模型训练,得到目标XGBoost树模型。

[0126] 在本申请的另一可选实施例中,提供了一种用于XGBoost联邦学习模型训练的数据处理方法,根据梯度直方图数据对XGBoost树模型进行模型训练,得到目标XGBoost树模型,包括以下步骤:

[0127] 根据所述直方图数据进行XGBoost树模型的最优分割点计算处理,得到最优分割点数据,其中,最优分割点数据为用于表示XGBoost树模型最优分割点的数据;根据最优分割点数据对XGBoost树模型进行基于树结构更新的模型训练处理,得到目标XGBoost树模型。

[0128] 在一个例子中,图7a和7b为本申请实施例提供的一种用于XGBoost联邦学习模型训练的数据处理方法的流程示意图,如图7a和7b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,以计算一阶梯度和计算二阶梯度,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。

[0129] 与上述Guest发起方进行步骤想对应的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。

[0130] 在本申请的另一可选实施例中,提供了一种用于XGBoost联邦学习模型训练的数据处理方法,在根据所述梯度直方图数据对XGBoost树模型进行模型训练,得到目标XGBoost树模型之后,该方法还包括:获取待预测样本数据,其中,所述待预测样本数据为至少一个发起方和至少一个参与方需要进行样本预测的数据;根据的目标XGBoost树模型对待预测样本数据进行预测处理,得到预测结果数据。

[0131] 举例说明如上述目标XGBoost树模型为用于患者理赔的预测模型,获取待预测的样本数据,根据该目标XGBoost树模型对待预测的样本数据进行预测处理,得到患者理赔预测结果;举例说明如上述目标XGBoost树模型用于贷款违约的预测模型,获取待预测的样本数据,根据该目标XGBoost树模型对待预测的样本数据进行预测处理,得到贷款违约预测结果;举例说明如上述目标XGBoost树模型用于广告推广的预测模型,获取待预测的样本数据,根据该目标XGBoost树模型对待预测的样本数据进行预测处理,得到广告点击率预测结果。

[0132] 在本申请的另一可选实施例中,提供了一种用于XGBoost联邦学习模型训练的数据处理装置,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,图8为本申请提供的一种用于XGBoost联邦学习模型训练的数据处理装置的示意图,如图8所示,该装置包括:

[0133] 训练样本获取模块81,用于获取待训练样本数据,其中,待训练样本数据为至少一个发起方和至少一个参与方的样本数据;

[0134] 预处理模块82,用于对样本数据进行基于数据特征提取的预处理,得到第一特征矩阵和第二特征矩阵,其中,第一特征矩阵为用于表示发起方样本数据特征的矩阵,第二特征矩阵为用于表示参与方样本数据特征对应的矩阵;

[0135] 预聚类压缩模块83,用于对第二特征矩阵中的每个特征进行随机预聚类压缩的处理,得到聚类矩阵和聚类索引;

[0136] 矩阵计算模块84,用于根据聚类矩阵进行稀疏矩阵构建处理,得到稀疏矩阵;根据

聚类索引对第一特征矩阵进行预聚合计算得到聚合矩阵；

[0137] 碎片化模块85，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；

[0138] 梯度直方图计算模块86，用于对第一碎片矩阵和第二碎片矩阵进行基于数据加密的矩阵乘法处理，得到梯度直方图数据；

[0139] 模型训练模块87，用于根据梯度直方图数据对XGBoost树模型进行模型训练，得到目标XGBoost树模型。

[0140] 关于上述实施例中各单元的执行操作的具体方式已经在有关该方法的实施例中进行了详细描述，此处将不做详细阐述说明。

[0141] 综上所述，在本申请中，通过在至少一个发起方与至少一个参与方之间数据共享场景中，获取待训练样本数据，其中，所述待训练样本数据为所述至少一个发起方和所述至少一个参与方的样本数据；对所述样本数据进行基于数据特征提取的预处理，得到第一特征矩阵和第二特征矩阵，其中，所述第一特征矩阵为用于表示发起方样本数据特征的矩阵，所述第二特征矩阵为用于表示参与方样本数据特征对应的矩阵；对所述第二特征矩阵中的每个特征进行随机预聚类压缩的处理，得到聚类矩阵和聚类索引；根据所述聚类矩阵进行稀疏矩阵构建处理，得到稀疏矩阵；根据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；对所述第一碎片矩阵和所述第二碎片矩阵进行基于数据加密的矩阵乘法处理，得到梯度直方图数据；根据所述梯度直方图数据对XGBoost树模型进行模型训练，得到目标XGBoost树模型。通过对联邦学习模型训练计算梯度直方图的过程中，对特征进行聚类压缩实现在稀疏矩阵上的运算加速，降低模型训练过程中的数据计算量及传输量等，在保证联邦学习数据隐私安全的情况下，降低模型训练过程中数据计算开销，提高模型训练的效率。

[0142] 需要说明的是，在附图的流程图示出的步骤可以在诸如一组计算机可执行指令的计算机系统中执行，并且，虽然在流程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。

[0143] 显然，本领域的技术人员应该明白，上述的本申请的各单元或各步骤可以用通用的计算装置来实现，它们可以集中在单个的计算装置上，或者分布在多个计算装置所组成的网络上，可选地，它们可以用计算装置可执行的程序代码来实现，从而，可以将它们存储在存储装置中由计算装置来执行，或者将它们分别制作成各个集成电路模块，或者将它们中的多个模块或步骤制作成单个集成电路模块来实现。这样，本申请不限制于任何特定的硬件和软件结合。

[0144] 以上所述仅为本申请的优选实施例而已，并不用于限制本申请，对于本领域的技术人员来说，本申请可以有各种更改和变化。凡在本申请的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

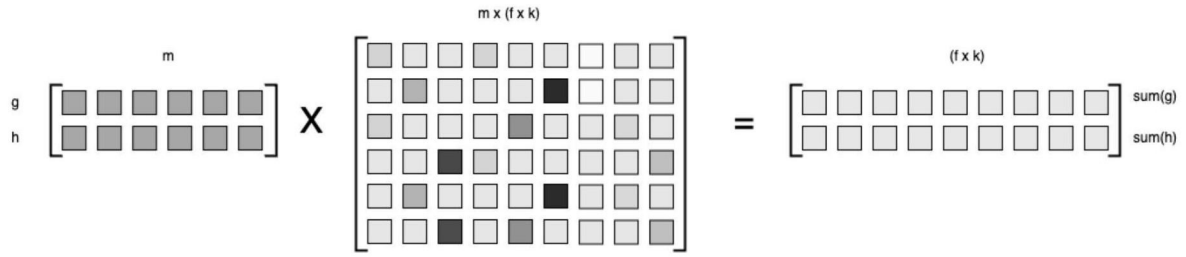


图1

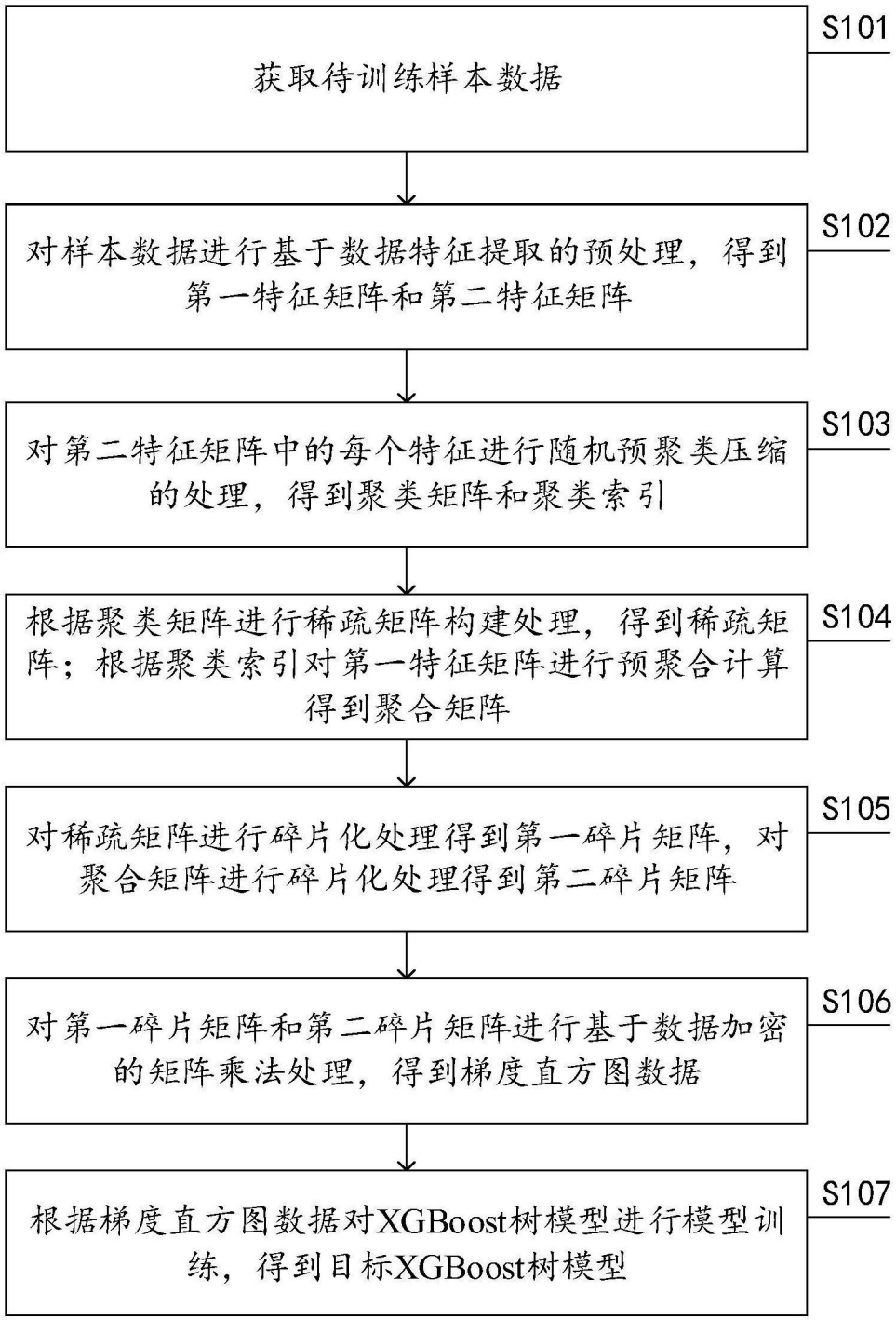

S101
获取待训练样本数据
S102
对样本数据进行基于数据特征提取的预处理，得到第一特征矩阵和第二特征矩阵
S103
对第二特征矩阵中的每个特征进行随机预聚类压缩的处理，得到聚类矩阵和聚类索引
S104
根据聚类矩阵进行稀疏矩阵构建处理，得到稀疏矩阵；根据聚类索引对第一特征矩阵进行预聚合计算得到聚合矩阵
S105
对稀疏矩阵进行碎片化处理得到第一碎片矩阵，对聚合矩阵进行碎片化处理得到第二碎片矩阵
S106
对第一碎片矩阵和第二碎片矩阵进行基于数据加密的矩阵乘法处理，得到梯度直方图数据
S107
根据梯度直方图数据对XGBoost树模型进行模型训练，得到目标XGBoost树模型

图2

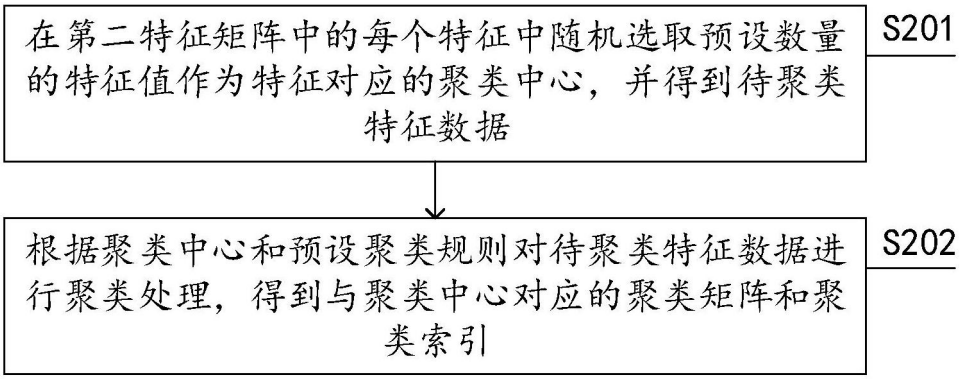

在第二特征矩阵中的每个特征中随机选取预设数量的特征值作为特征对应的聚类中心，并得到待聚类特征数据
S201
根据聚类中心和预设聚类规则对待聚类特征数据进行聚类处理，得到与聚类中心对应的聚类矩阵和聚类索引
S202

图3

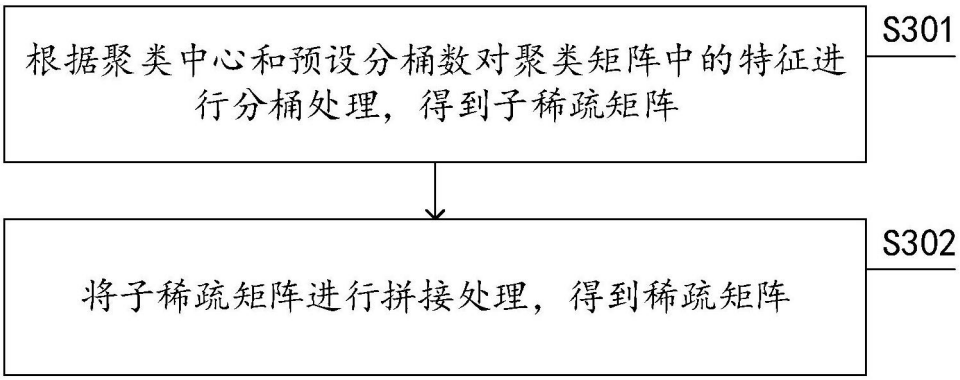

根据聚类中心和预设分桶数对聚类矩阵中的特征进行分桶处理，得到子稀疏矩阵
S301
将子稀疏矩阵进行拼接处理，得到稀疏矩阵
S302

图4

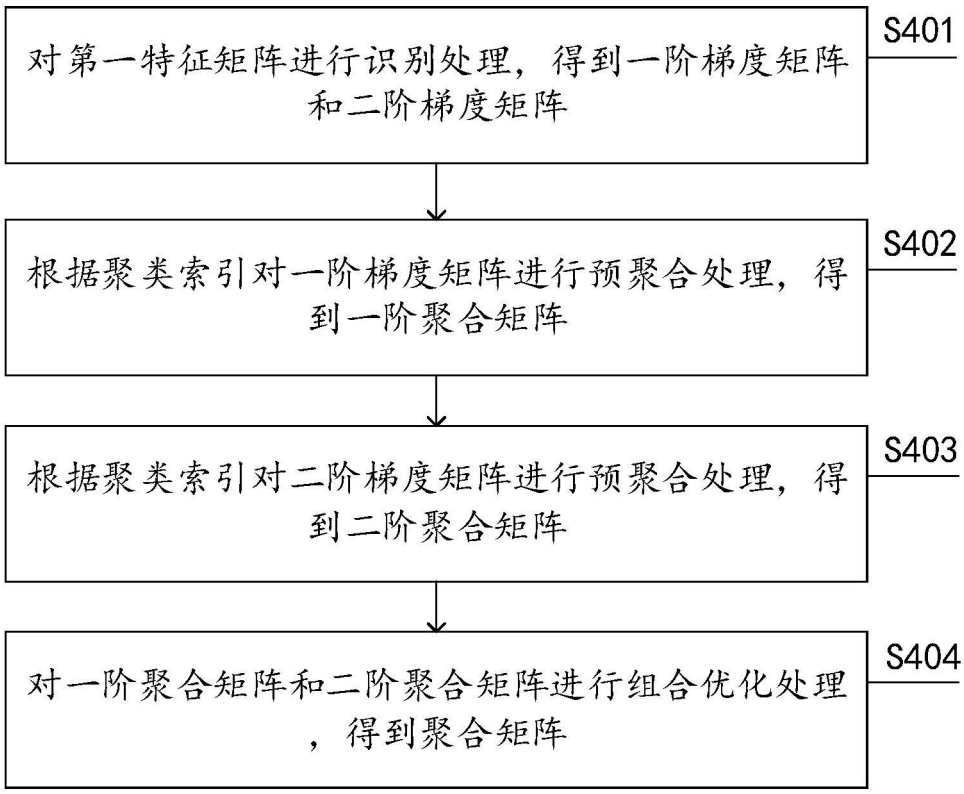

对第一特征矩阵进行识别处理，得到一阶梯度矩阵和二阶梯度矩阵
S401
根据聚类索引对一阶梯度矩阵进行预聚合处理，得到一阶聚合矩阵
S402
根据聚类索引对二阶梯度矩阵进行预聚合处理，得到二阶聚合矩阵
S403
对一阶聚合矩阵和二阶聚合矩阵进行组合优化处理，得到聚合矩阵
S404

图5

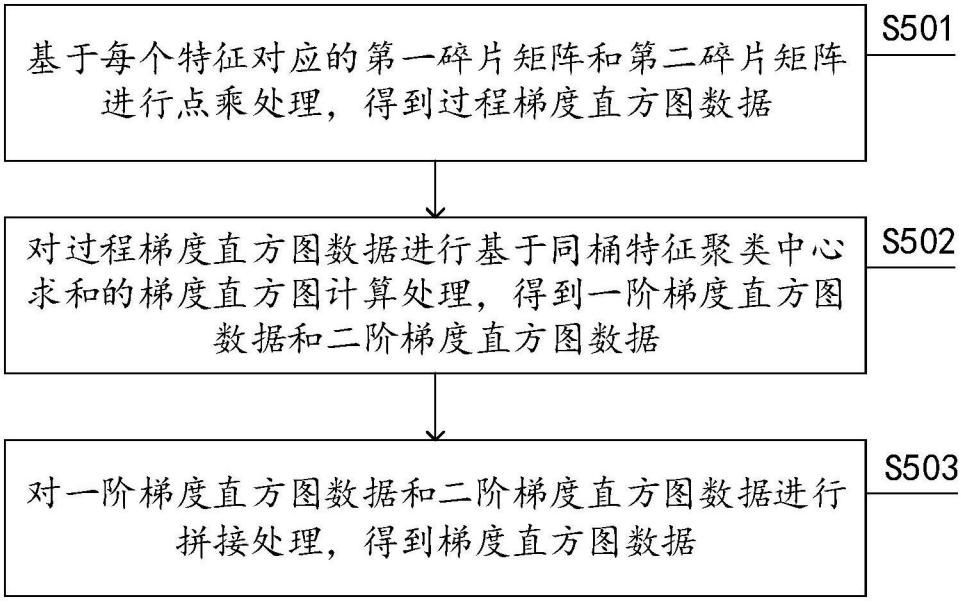

基于每个特征对应的第一碎片矩阵和第二碎片矩阵进行点乘处理，得到过程梯度直方图数据
S501
对过程梯度直方图数据进行基于同桶特征聚类中心求和的梯度直方图计算处理，得到一阶梯度直方图数据和二阶梯度直方图数据
S502
对一阶梯度直方图数据和二阶梯度直方图数据进行拼接处理，得到梯度直方图数据
S503

图6

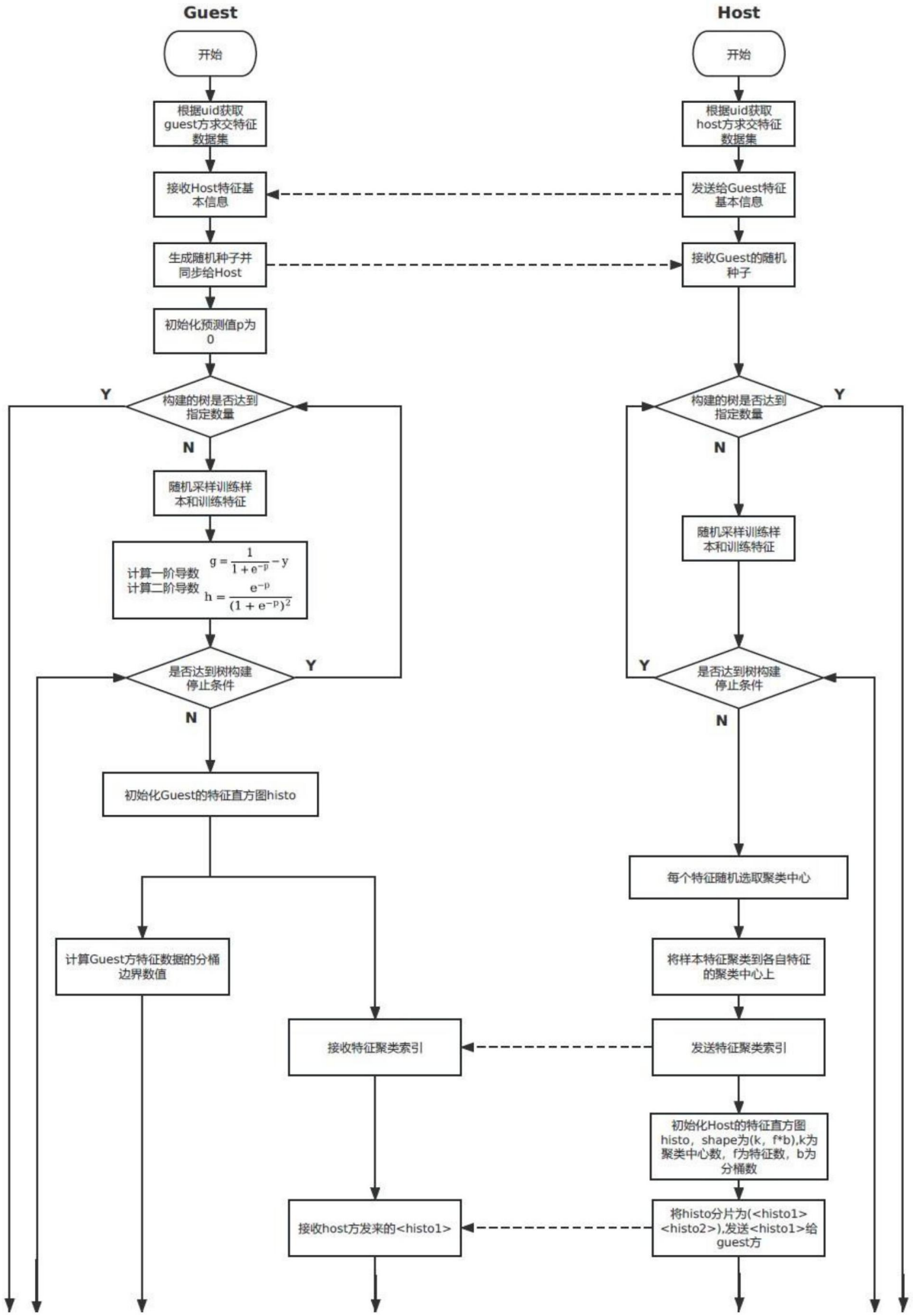

Guest
开始
根据uid获取guest方求交特征数据集
接收Host特征基本信息
生成随机种子并同步给Host
初始化预测值p为0
构建的树是否达到指定数量
Y
N
随机采样训练样本和训练特征
计算一阶导数 $g=\frac{1}{1+e^{-p}}-y$
计算二阶导数 $h=\frac{e^{-p}}{(1+e^{-p})^2}$
是否达到树构建停止条件
Y
N
初始化Guest的特征直方图histo
计算Guest方特征数据的分桶边界数值
接收特征聚类索引
接收host方发来的<histo1>
Host
开始
根据uid获取host方求交特征数据集
发送给Guest特征基本信息
接收Guest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7a

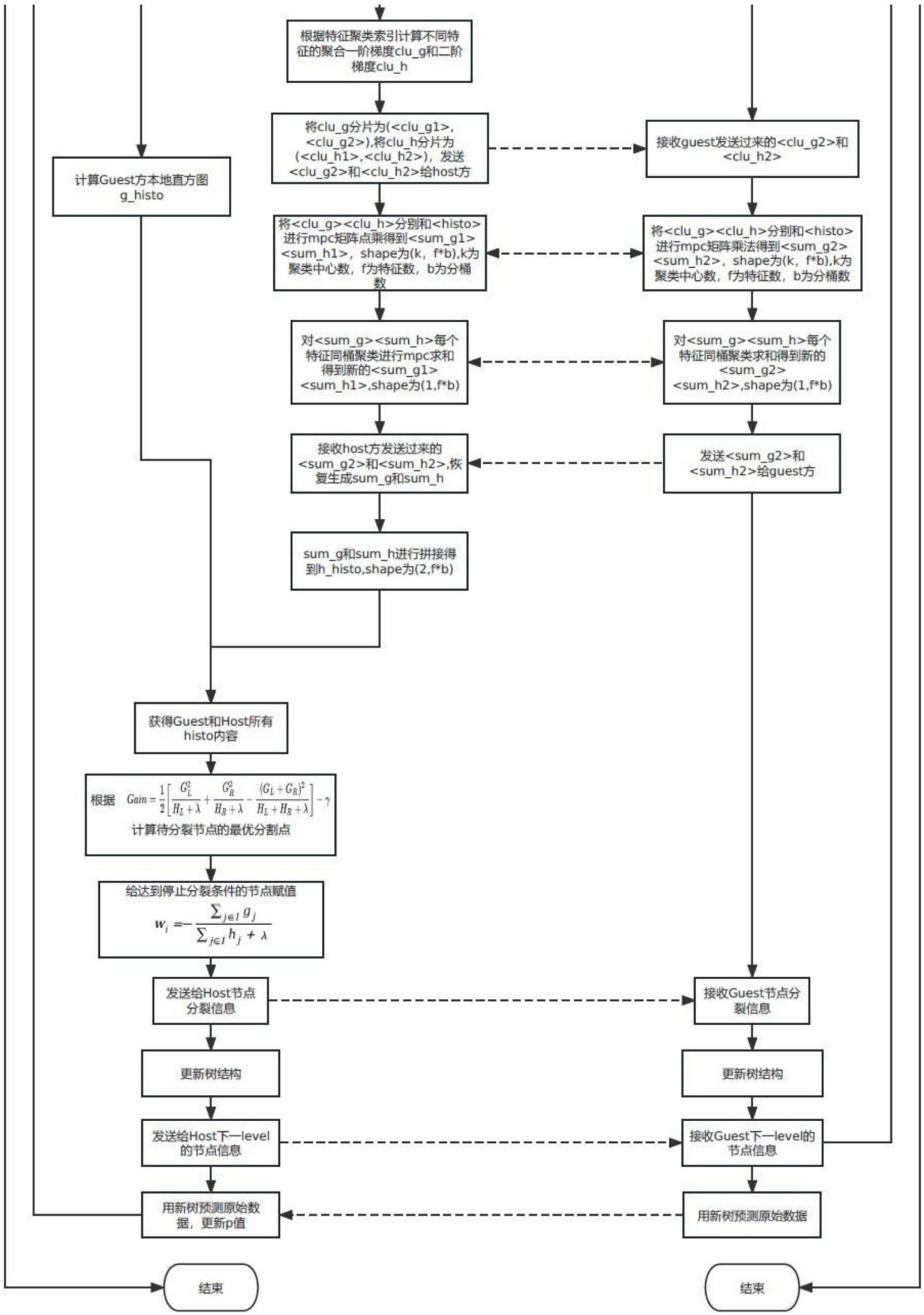

根据特征聚类索引计算不同特征的聚合一阶梯度clu_g和二阶梯度clu_h
将clu_g分片为(<clu_g1>,<clu_g2>),将clu_h分片为(<clu_h1>,<clu_h2>)，发送<clu_g2>和<clu_h2>给host方
接收guest发送过来的<clu_g2>和<clu_h2>
计算Guest方本地直方图g_histo
将<clu_g><clu_h>分别和<histo>进行mpc矩阵点乘得到<sum_g1><sum_h1>，shape为(k，f*b),k为聚类中心数，f为特征数，b为分桶数
将<clu_g><clu_h>分别和<histo>进行mpc矩阵乘法得到<sum_g2><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level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7b

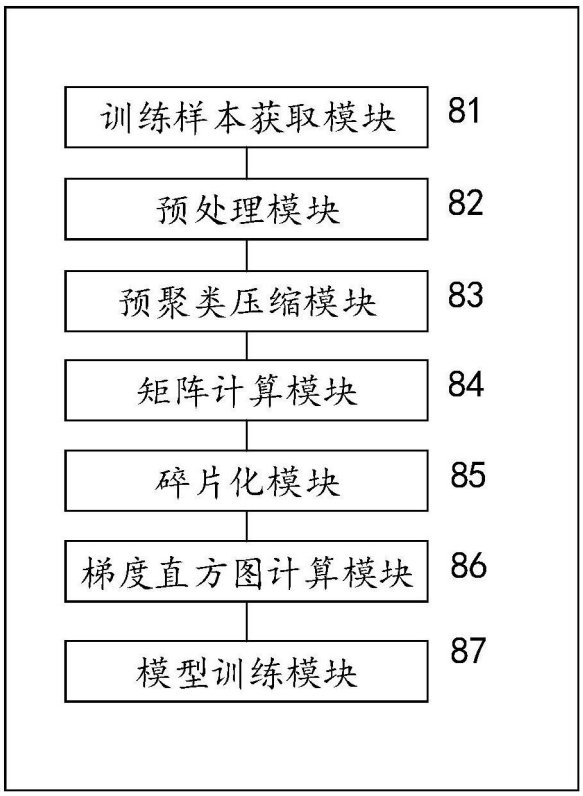

训练样本获取模块
81
预处理模块
82
预聚类压缩模块
83
矩阵计算模块
84
碎片化模块
85
梯度直方图计算模块
86
模型训练模块
87

图8