

权 利 要 求 书

1. 一种基于随机贪心算法的横向联邦梯度提升树优化方法，其特征在于：其步骤如下：

步骤一：协调方设置梯度提升树模型相关参数，包括决策树最大数量 T 、树最大深度 L 、初始预测值 $base$ 等，并下发到各个参与方 p_i ；

5 步骤二：令树计数器 $t=1$ ；

步骤三：对每个参与方 p_i ，初始化第 k 棵树训练目标 $y_k = y_{k-1} - \hat{y}_{k-1}$ ；其中 $y_0 = y$ ， $\hat{y}_0 = base$ ；

步骤四：令树层数计数器 $l=1$ ；

步骤五：令当前层节点计数器 $n=1$ ；

10 步骤六：对每个参与方 p_i ，根据本地当前节点 n 的数据，根据最优分割点算法，确定当前节点的分割点，并将分割点信息发送给协调方；

步骤七：协调方统计全部参与方的切割点信息，根据 ϵ -贪心算法，确定分割特征 f 和分割值 v ；

15 步骤八：协调方将最终确定的分割信息，包括确定分割特征 f 和分割值 v ，下发给各个参与方；

步骤九：各个参与方根据分割特征 f 和分割值 v 分割当前节点数据集，并将新的分割数据分配给子节点；

步骤十：令 $n=n+1$ ，如果 n 小于或等于当前层最大节点数，继续步骤三；反之，继续下一步；

20 步骤十一：根据 l 层节点的子节点重置当前层节点信息，令 $l=l+1$ ，如果 l 小于或等于树最大深度 L ，继续步骤五；反之，继续下一步；

步骤十二：令 $t=t+1$ ，如果 t 大于或等于决策树最大数量 T ，继续步骤 3；反之，结束。

2. 根据权利要求 1 所述的一种基于随机贪心算法的横向联邦梯度提升树优化方法，其特征在于：所述步骤三中的最优分割点算法：

25 确定分割目标函数：包括目标函数，

信息增益：信息增益是度量样本集合纯度最常用的一种指标。假设节点样本集合 D 中共有 K 类样本，其中第 k 类样本所占的比例为 p_k ，则 D 的信息熵定义为

$$Ent(D) = - \sum_{k=1}^{|K|} p_k \log_2 p_k$$

假设节点根据属性 a 切分为 V 个可能的取值，则信息增益定义为

$$\text{Gain}(D, a) = \text{Ent}(D) - \sum_{v=1}^V \frac{|D^v|}{|D|} \text{Ent}(D^v)。$$

信息增益率：

$$\text{Gain_ratio}(D, a) = \frac{\text{Gain}(D, a)}{\text{IV}(a)}$$

其中

$$\text{IV}(a) = - \sum_{v=1}^V \frac{|D^v|}{|D|} \log_2 \frac{|D^v|}{|D|}$$

基尼系数：

$$\text{Gini}(D) = \sum_{k=1}^{|K|} \sum_{k' \neq k} p_k p_{k'}$$

$$\text{Gini_index}(D, a) = \sum_{v=1}^V \frac{|D^v|}{|D|} \text{Gini}(D^v)$$

5 结构系数：

$$\text{Gain} = \frac{1}{2} \left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + \lambda} \right) - \gamma$$

其中 G_L 为根据分割点分割数据集后划分到左节点的数据集的一阶梯度和， H_L 为左节点的数据集的二阶梯度和， G_R 及 H_R 为相应右节点的梯度信息和， γ 为树模型复杂度惩罚项， λ 为二阶正则项；

确定分割值候选列表：根据当前节点数据分布，确定分割值列表；分割值包括分割特征和分割特征值；分割值列表根据以下方法确定：

数据集中所有特征的所有取值；

针对数据集中每个特征的取值范围，确定离散分割点；

分割点的选择可以根据数据的分布，均匀分布在取值范围内；其中均匀体现在分割点间的数据量近似相等或者二阶梯度和近似相等；

15 遍历分割值候选列表，寻找使目标函数最优的分割点。

3. 根据权利要求1所述的一种基于随机贪心算法的横向联邦梯度提升树优化方法，其特征在于：所述步骤七中的Epsilon贪心算法包含：

针对节点n各参与方把节点分割点信息发送给协调方，包括分割特征 f_i ，分割值 v_i ，节点样本数量 N_i ，本地目标函数增益 g_i ；其中i代表各参与方；

20 协调方根据各参与方分割信息，基于最大数原则，确定最优分割特征 $f_{\max}$ 设X为均匀分

布在[0,1]之间的随机数，对 X 随机取样得 x ；如果 $x \leq \epsilon$ ，则在各参与方分割特征中随机选择一个作为全局分割特征；反之，选择 $f_{\max}$ 为全局分割特征；

各参与方根据全局分割特征重新计算分割信息，并发送给协调方；

协调方根据一下公式确定全局分割值：如果参与方总数为 P ；

$$5 \quad v = \sum_{i=1}^P \frac{N_i g_i}{\sum_{j=1}^P N_j g_j} v_i \text{ 将分割值分发到各参与方，进行节点分割。}$$

4. 根据权利要求 1 所述的一种基于随机贪心算法的横向联邦梯度提升树优化方法，其特征在于：所述横向联邦学习，是联邦学习的一种分布式结构，其中各个分布式节点的数据特征相同，样本空间不同。

5. 根据权利要求 1 所述的一种基于随机贪心算法的横向联邦梯度提升树优化方法，其
10 特征在于：所述梯度提升树算法，是一种基于梯度提升和决策树的集成模型。

6. 根据权利要求 1 所述的一种基于随机贪心算法的横向联邦梯度提升树优化方法，其特征在于：所述决策树是梯度提升树模型的基础模型，基于树结构，在节点通过给定特征判断样本的预测方向。

7. 根据权利要求 1 所述的一种基于随机贪心算法的横向联邦梯度提升树优化方法，其
15 特征在于：所述分割点是决策树中非叶节点进行数据分割的切分位置。

8. 根据权利要求 1 所述的一种基于随机贪心算法的横向联邦梯度提升树优化方法，其特征在于：所述直方图是表示节点数据中一阶梯度和二阶梯度的统计信息。

9. 根据权利要求 1 所述的一种基于随机贪心算法的横向联邦梯度提升树优化方法，其特征在于：录入设备可以是计算机、手机等数据终端或者是移动终端的一种或多种。

20 10. 根据权利要求 1 所述的一种基于随机贪心算法的横向联邦梯度提升树优化方法，其特征在于：录入设备包括处理器，被所述处理器执行时实现步骤一到十二中的任一项所述算法。