

一种基于随机贪心算法的横向联邦梯度提升树优化方法

技术领域

本发明涉及联邦学习技术领域，具体为一种基于随机贪心算法的横向联邦梯度提升树优化方法。

背景技术

联邦学习是一个机器学习框架，能有效帮助多个机构在满足用户隐私保护、数据安全和政府法规的要求下，进行数据使用和机器学习建模，让参与方在未共享数据的基础上联合建模，能从技术上打破数据孤岛，实现 AI 协作，在此框架下通过设计虚拟模型解决不同数据拥有方在不交换数据的情况下进行协作的问题，虚拟模型是各方将数据聚合在一起的最优模型，各自区域依据模型为本地目标服务，联邦学习要求建模结果应当无限接近传统模式，即将多个数据拥有方的数据汇聚到一处进行建模的结果，在联邦机制下，各参与者的身份和地位相同，可建立共享数据策略，贪心算法是一种对某些求最优解问题的更简单、更迅速的设计技术，贪心算法的特点是一步一步地进行，常以当前情况为基础根据某个优化测度作最优选择，而不考虑各种可能的整体情况，省去了为找最优解要穷尽所有可能而必须耗费的大量时间，贪心算法采用自顶向下，以迭代的方法做出相继的贪心选择，每做一次贪心选择，就将所求问题简化为一个规模更小的子问题，通过每一步贪心选择，可得到问题的一个最优解，虽然每一步上都要保证能获得局部最优解，但由此产生的全局解有时不一定是最优的，所以贪心算法不要回溯。

然而，现有横向联邦梯度提升树算法需要各个参与方和协调方在频繁传递直方图信息，对协调方网络带宽要求很高，训练效率容易受网络稳定性的影响，并且由于传递的直方图信息中包含用户信息，存在泄漏用户隐私的风险，在引入多方安全计算、同态加密、秘密共享等隐私保护方案后，可以减少用户隐私泄漏的可能性，但会更加本地计算负担，降低训练效率。

发明内容

本发明的目的在于提供一种基于随机贪心算法的横向联邦梯度提升树优化方法，以解决上述背景技术中提出现有横向联邦梯度提升树算法需要各个参与方和协调方在频繁传递直方图信息，对协调方网络带宽要求很高，训练效率容易受网络稳定性的影响，并且由于传递

的直方图信息中包含用户信息，存在泄漏用户隐私的风险，在引入多方安全计算、同态加密、秘密共享等隐私保护方案后，可以减少用户隐私泄漏的可能性，但会更加本地计算负担，降低训练效率的问题。

为实现上述目的，本发明提供如下技术方案：一种基于随机贪心算法的横向联邦梯度提升树优化方法，包括其步骤如下：

步骤一：协调方设置梯度提升树模型相关参数，包括但不限于决策树最大数量 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 。

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 $n=1$ 。

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 。

步骤八：协调方将最终确定的分割信息，包括但不限于确定分割特征 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，继续步骤六；反之，继续下一步。

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。优选的，所述步骤六中的最优分割点算法：

一、确定分割目标函数：包括但不限于以下目标函数，

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

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

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

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

5 基尼系数：

$$\begin{aligned} \text{Gini}(D) &= \sum_{k=1}^{|K|} \sum_{k' \neq k} p_k p_{k'} \\ \text{Gini_index}(D, a) &= \sum_{v=1}^V \frac{|D^v|}{|D|} \text{Gini}(D^v) \end{aligned}$$

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。

10 二、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；分割值列表可以根据以下方法确定：

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

15 似相等。

遍历分割值候选列表，寻找使目标函数最优的分割点。

优选的，所述步骤七中的 Epsilon 贪心算法：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；

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；

- 5 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 ；

$$v = \sum_{i=1}^P \frac{N_i g_i}{\sum_{j=1}^P N_j g_j} v_i$$

将分割值分发到各参与方，进行节点分割。

优选的，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。

- 10 优选的，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。

优选的，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。

优选的，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。

优选的，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。

- 15 优选的，所述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。

优选的，所述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。

- 与现有技术相比，本发明的有益效果是：该基于随机贪心算法的横向联邦梯度提升树优化方法，通过协调方设置梯度提升树模型相关参数，包括但不限于决策树最大数量 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 ，令树计数器 $t=1$ ，对每个参与方 p_i ，令树层数计数器 $l=1$ ，令当前层节点计数器 $n=1$ ，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，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 ，协调方将最终确定的分割信息，包括但不限于确定分割特征 f 和分割值 v ，下发给各个参与方，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，
- 20 令 $n=n+1$ ，如果 n 小于或等于当前层最大节点数，继续步骤六，反之，继续下一步，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，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，
- 25

结束，利支持的横向联邦学习中包括参与方和协调方，参与方拥有本地数据，协调方不拥有任何数据，进行参与方信息聚合的中心，参与方分别计算直方图，将直方图发送给协调方，协调方汇总全部直方图信息后，根据贪心算法寻找最优分割点，然后分享给各个参与方，配合内部的算法进行工作。

5

附图说明

为了更清楚地说明本发明实施例中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其它的附图。

10

图 1 为本发明基于随机贪心算法的横向联邦梯度提升树优化方法架构示意图；

图 2 为本发明基于随机贪心算法的横向联邦梯度提升树优化方法步骤示意图；

图 3 为本发明基于随机贪心算法的横向联邦梯度提升树优化方法判断示意图。

具体实施方式

15

下面将结合本发明实施例中的附图，对本发明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发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，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，都属于本发明保护的范围。

请参阅图 1-3，本发明提供一种技术方案：一种基于随机贪心算法的横向联邦梯度提升树优化方法，包括其步骤如下：

20

步骤一：协调方设置梯度提升树模型相关参数，包括但不限于决策树最大数量 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 。

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 $y_0 = y, \hat{y}_0 = base$ 。

25

步骤四：令树层数计数器 $l=1$ 。

步骤五：令当前层节点计数器 $n=1$ 。

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。

30

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

步骤八：协调方将最终确定的分割信息，包括但不限于确定分割特征 f 和分割值 v，下发给各个参与方。

5 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，继续步骤六；反之，继续下一步。

10 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；进一步的，步骤六中的最优分割点算法：

一、确定分割目标函数：包括但不限于以下目标函数，

15 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

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

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：

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

其中

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

20 基尼系数：

$$\begin{aligned} \text{Gini}(D) &= \sum_{k=1}^{|K|} \sum_{k' \neq k} p_k p_{k'} \\ \text{Gini_index}(D, a) &= \sum_{v=1}^V \frac{|D^v|}{|D|} \text{Gini}(D^v) \end{aligned}$$

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。

- 5 二、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；分割值列表可以根据以下方法确定：

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。
- 10

遍历分割值候选列表，寻找使目标函数最优的分割点；

进一步的，步骤七中的 Epsilon 贪心算法：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；

- 15 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；

$$v = \sum_{i=1}^P \frac{N_i g_i}{\sum_{j=1}^P N_j g_j} v_i$$

- 20 将分割值分发到各参与方，进行节点分割；

进一步的，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，更好的进行比对工作；

进一步的，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，更好的进行工作；

- 25 进一步的，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，能够更好的帮助预测；

进一步的，分割点是决策树中非叶节点进行数据分割的切分位置，更好的进行分割；

进一步的，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，更直观的进行表示；

进一步的，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，

5 更好的进行数据录入；

进一步的，录入设备包括处理器，被处理器执行时实现步骤一到十二中的任一项算法。

工作原理：步骤一：协调方设置梯度提升树模型相关参数，包括但不限于决策树最大数量 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 ，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
10 $y_0 = y, \hat{y}_0 = base$ ，步骤四：令树层数计数器 $l=1$ ，步骤五：令当前层节点计数器 $n=1$ ，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，一、确定分割目标函数：包括但不限于以下目标函数，

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
15 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

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

信息增益率：

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

其中

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

20 基尼系数：

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

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

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,

- 5 二、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；分割值列表可以根据以下方法确定：

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
10 近似相等，

遍历分割值候选列表，寻找使目标函数最优的分割点，步骤七：协调方统计全部参与方的切割点信息，根据 epsilon-贪心算法，确定分割特征 f 和分割值 v ，步骤七中的 Epsilon 贪心算法：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 ，
15 本地目标函数增益 g_i ；其中 i 代表各参与方；

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 $x \leq \text{epsilon}$ ，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；

- 20 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 ；

$$v = \sum_{i=1}^P \frac{N_i g_i}{\sum_{j=1}^P N_j g_j} v_i$$

- 将分割值分发到各参与方，进行节点分割，步骤八：协调方将最终确定的分割信息，包括但不限于确定分割特征 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，继续步骤六；反之，继续下一步，步骤十一：
25 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，通过协调方设置梯度提升树模型相关参数，包括但不限于决策树最大数量、树最大深度、初始预测值等，并下发到各个参与方，协调方将最终确定的分割信息，包括但不限于确定分割特征和分割值，下发给各个参与方，各个参与方根据分割特征和分割值分割当前节点数据集，利支持的横向联邦学习中包括参与方和协调方，参与方拥有本地数据，协调方不拥有任何数据，进行参与方信息聚合的中心，参与方分别计算直方图，将直方图发送给协调方，协调方汇总全部直方图信息后，根据贪心算法寻找最优分割点，然后分享给各个参与方，配合内部的算法进行工作。

尽管已经示出和描述了本发明的实施例，对于本领域的普通技术人员而言，可以理解在不脱离本发明的原理和精神的情况下可以对这些实施例进行多种变化、修改、替换和变形，

10 本发明的范围由所附权利要求及其等同物限定。