CLAIMS

1. A horizontal federated Gradient Boosting Decision Tree optimization method based on a random greedy algorithm, comprising the following steps:

Step 1: a coordinator setting relevant parameters of a Gradient Boosting Decision Tree model, including a maximum number of decision trees T, a maximum depth of trees L, an initial predicted value base, etc., and sending the relevant parameters to respective participants p_i;

Step 2: letting a tree counter t=1;

Step 3: for each participant p_i, initializing a training target of a kth tree $y_k = y_{k-1} - \hat{y}_{k-1}$; wherein $y_0 = y, \hat{y}_0 = \text{base}$;

Step 4: letting a tree layer counter l=1;

Step 5: letting a node counter of a current layer n=1;

Step 6: for each participant p_i, determining a segmentation point of a local current node n according to the data of the current node and an optimal segmentation point algorithm and sending the segmentation point information to the coordinator;

Step 7: the coordinator counting the segmentation point information of all participants, and determining a segmentation feature f and a segmentation value v according to an epsilon- greedy algorithm;

Step 8, the coordinator sending the finally determined segmentation information, including the determined segmentation feature f and segmentation value v, to respective participants;

Step 9: each participant segmenting a data set of the current node according to the segmentation feature f and the segmentation value v, and distributing new segmentation data to child nodes;

Step 10: letting n=n+1, and continuing with the Step 3 if n is less than or equal to a maximum number of nodes in the current layer; otherwise, proceeding to a next step;

Step 11: resetting the node information of the current layer according to the child nodes of a node of a lth layer, so that l=l+1, and continuing with the Step 5 if l is less than or equal to the maximum tree depth L; otherwise, proceeding to a next step;

Step 12: letting t=t+1, and continuing with the Step 3 if t is greater than or equal to the maximum number of decision trees T; otherwise, ending.

2. The horizontal federated Gradient Boosting Decision Tree optimization method based on a random greedy algorithm according to claim 1, wherein the optimal segmentation point algorithm in the Step 6:

determines a segmentation objective function, including an objective function,

information gain: information gain is the most commonly used index to measure a purity of a sample set; assuming that there are K types of samples in a node sample set D, in which a proportion of a kth type of samples is p_k, an information entropy of D is defined as:

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

assuming that the node is segmented into V possible values according to an attribute a, the information gain is defined as:

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

information gain rate:

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

where

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

a Gini coefficient:

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

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

a structural coefficient:

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

where G_L is a sum of first-order gradients of the data set divided into a left node according to the segmentation point, H_L is a sum of second-order gradients of the data set of the left node, and G_R and H_R are sums of the gradient information of a corresponding right node, γ is a tree model complexity penalty term and λ is a second-order regular term;

determines a candidate list of segmentation values: determining the list of segmentation values according to the data distribution of the current node, wherein the segmentation values

comprise segmentation features and segmentation feature values; the list of segmentation value is determined according to the following method:

all values of all features in the data set;

determining a discrete segmentation point according to a value range of each feature in the data set;

wherein the selection of the segmentation points can be evenly distributed within the value range according to the distribution of data, and the amount of data uniformly reflected between the segmentation points is approximately equal or a sum of the second-order gradients is approximately equal;

traversing the candidate list of segmentation values to find the segmentation point that makes the objective function optimal.

3. The horizontal federated Gradient Boosting Decision Tree optimization method based on a random greedy algorithm according to claim 1, wherein the Epsilon greedy algorithm in the Step 7 comprises:

for the node n, each participant sending the node segmentation point information to the coordinator, including a segmentation feature f_i, a segmentation value v_i, a number of node samples N_i and a local objective function gain g_i, where i represents respective participants;

according to the segmentation information of each participant and based on a maximum number principle, the coordinator determining an optimal segmentation feature f_{max}, where X is a random number evenly distributed among [0,1] and randomly sampling X to obtain x; if x<=epsilon, randomly selecting one of the segmentation features of each participant as a global segmentation feature; otherwise, selecting f_{max} as the global segmentation feature;

each participant recalculating the segmentation information according to the global segmentation feature and sending the segmentation information to the coordinator;

the coordinator determining a global segmentation value according to the following formula: if the total number of participants is P,

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

distributing the segmentation value to each participant to perform node segmentation.

4. The horizontal federated Gradient Boosting Decision Tree optimization method based on a random greedy algorithm according to claim 1, wherein the horizontal federated learning is a distributed structure of federated learning, in which each distributed node has the same data

feature and different sample spaces.

5. The horizontal federated Gradient Boosting Decision Tree optimization method based on a random greedy algorithm according to claim 1, wherein the Gradient Boosting Decision Tree algorithm is an integrated model based on gradient boosting and decision tree.

6. The horizontal federated Gradient Boosting Decision Tree optimization method based on a random greedy algorithm according to claim 1, wherein the decision tree is a basic model of a Gradient Boosting Decision Tree model, and a prediction direction of a sample is judged at the node by given features based on a tree structure.

7. The horizontal federated Gradient Boosting Decision Tree optimization method based on a random greedy algorithm according to claim 1, wherein the segmentation point is a segmentation position of non-leaf nodes in the decision tree for data segmentation.

8. The horizontal federated Gradient Boosting Decision Tree optimization method based on a random greedy algorithm according to claim 1, wherein the histogram is statistical information representing the first-order gradient and the second-order gradient in node data.

9. The horizontal federated Gradient Boosting Decision Tree optimization method based on a random greedy algorithm according to claim 1, wherein an input device can be one or more of data terminals such as computers and mobile phones or mobile terminals.

10. The horizontal federated Gradient Boosting Decision Tree optimization method based on a random greedy algorithm according to claim 1, wherein the input device comprises a processor, and when executed by the processor, the algorithm of any one of steps 1 to 12 is implemented.