近似算法

周志健

1 解决的问题

求解最优二叉树结构，以损失减少值作为评价分数:

$$\mathcal{L}_{split}=\frac{1}{2}\left[\frac{(\sum_{i\in I_L}g_i)^2}{\sum_{i\in I_L}h_i+\lambda}+\frac{(\sum_{i\in I_R}g_i)^2}{\sum_{i\in I_R}h_i+\lambda}-\frac{(\sum_{i\in I}g_i)^2}{\sum_{i\in I}h_i+\lambda}\right]-\gamma \tag{1}$$

其中，I_L和I_R是划分后左右节点的实例集，g_i和h_i分别为目标损失函数在实例$\mathbf{x}_i$处的一阶和二阶导数，λ和γ是目标损失函数中正则化项的参数。

最终求解问题:

$$x=\arg\max\mathcal{L}_{split} \tag{2}$$

2 近似求解算法

近似求解算法为了提高算法计算效率而提出，差别于基本贪心算法，它考虑排序数据的百分比信息，基本思想是从最小值开始，每隔ε比例的数据量就选择一个候选点，然后从所有候选点中选出最优值。其中候选点集D的大小受到ε值的影响，值越小，D越大。

选择候选点集D有两种方式:

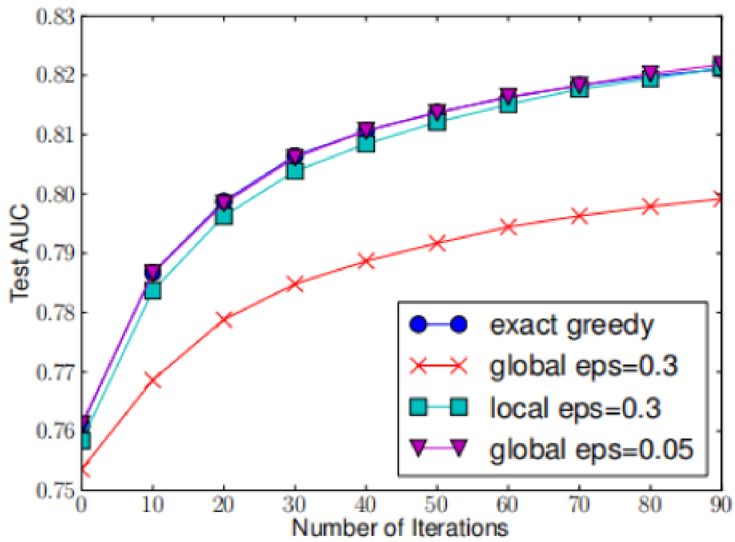
1. **全局选择:** 在构造树之前就基于总体样例提出所有特征对应的候选点，并在所有层都使用相同的候选点集D。
2. **局部选择:** 每次树向下划分后，更新左子树和右子树对应的候选点集D_L、D_R。

相较而言，全局选择比局部选择需要更少的找候选点集步骤，但由于没有细化候选点集，所以要达到与局部选择方法相同的精度则需要更小的ε值，也就是要计算更多候选点，所以局部选择方法更适合较深的树。

Algorithm 2: Approximate Algorithm for Split Finding

for $k = 1$ ***to*** m **do**
  Propose $S_k = \{s_{k1}, s_{k2}, \cdots s_{kl}\}$ by percentiles on feature k.
  Proposal can be done per tree (global), or per split(local).
end
for $k = 1$ ***to*** m **do**
  $G_{kv} \leftarrow= \sum_{j \in \{j | s_{k,v} \geq \mathbf{x}_{jk} > s_{k,v-1}\}} g_j$
  $H_{kv} \leftarrow= \sum_{j \in \{j | s_{k,v} \geq \mathbf{x}_{jk} > s_{k,v-1}\}} h_j$
end
Follow same step as in previous section to find max score only among proposed splits.

3 全局选择与局部选择的比较



（Higgs boson dataset）

观察图像，作者给出两个主要结论：

1. 相同ε值下，局部选择方法效果更好。
2. 在ε值足够小的情况下，近似算法可以获得与精确贪心算法相同的精度。