# 1 算法流程

**异构深度森林算法：**

**输入：**

训练集 D，样本个数 N，$x_i$, $y_i$, i=1,...,N，$x_i$ 的维度为 n，$y_i$ 维度为 m。异构随机森林的总层数为 L，其中一层记为 j,j=1,...L。基学习器随机森林模型记为 RF，逻辑回归模型记为 LR，极端随机森林模型记为 ERF，xgboost 模型记为 XGB，lightgbm 模型记为 LGB，catboost 模型记为 CAB, $k$ 为每一层基学习器的数量。

**步骤：**

**Step 1:** j=1

**Step 2:** 初始化基学习器

$$RF_j, LR_j, ERF_j, XGB_j, LGB_j, CAB_j$$

根据训练集的特征 $x_i$ 和标签 $y_i$ 进行训练，得到模型对于特征的预测概率

$$p_{RF_{ij}}, p_{LR_{ij}}, p_{ERF_{ij}}, p_{XGB_{ij}}, p_{LGB_{ij}}, p_{CAB_{ij}}$$

其中所有的 $p$ 维度为 $m$，与 $y_i$ 相同

**Step 3:** 将特征与预测概率进行融合，令

$$x_i' = [x_i, p_{RF_{ij}}, p_{LR_{ij}}, p_{ERF_{ij}}, p_{XGB_{ij}}, p_{LGB_{ij}}, p_{CAB_{ij}}]$$

此时 $x_i'$ 的维度为 $(n + k * m)$

**Step 4:** $j = j + 1$，从 $j = 2$ 开始，

$$j = j + 1$$

训练基学习器的特征替换为 $x_i'$, 重复 step2-step3，直至 $j = L$

**Step 5:** 根据最后一次的模型预测概率

$$p_{RF_{iL}}, p_{LR_{iL}}, p_{ERF_{iL}}, p_{XGB_{iL}}, p_{LGB_{iL}}, p_{CAB_{iL}}$$

计算

$$p_{final} = \frac{1}{k}[p_{RF_{iL}} + p_{LR_{iL}} + p_{ERF_{iL}} + p_{XGB_{iL}} + p_{LGB_{iL}} + p_{CAB_{iL}}]$$

**Step 6:** 根据 $p_{final}$ 计算 $y' = \arg\max(p_{final})$ 为最终的模型预测结果

**残差深度森林算法：**

**输入：**

训练集 D，样本个数 N，$x_i$, $y_i$, i=1,...,N，$x_i$ 的维度为 n，$y_i$ 维度为 m。异构随机森林的总层数为 L，其中一层记为 j,j=1,...L。基学习器随机森林模型记为 RF，极端随机森林模型记为 ERF。$k$ 为每一层基学习器的数量。

**步骤：**

**Step 1:** j=1

**Step 2:** 初始化基学习器 $RF_j, ERF_j$，根据训练集的特征 $x_i$ 和标签 $y_i$ 进行训练，得到模型对于特征的预测概率 $p_{RF_{ij}}, p_{ERF_{ij}}$，其中所有的 $p$ 维度为 $m$，与 $y_i$ 相同。

**Step 3:** 将特征与预测概率进行融合，令

$$x_i' = [x_i, p_{RF_{ij}}, p_{ERF_{ij}}]$$

此时 $x_i'$ 的维度为 $(n + k * m * j)$

**Step 4:** 令 $x = x_i', j = j + 1$，重复 step2-step3，直至 j = L

**Step 5:** 根据最后一次的模型预测概率 $p_{RF_{iL}}, p_{ERF_{iL}}$ 计算

$$p_{final} = \frac{1}{k}[p_{RF_{iL}} + p_{ERF_{iL}}]$$

**Step 6:** 根据 $p_{final}$ 计算 $y' = \arg\max(p_{final})$ 为最终的模型预测结果

# 2 参数说明及模型对比

表 1: 异构深度森林参数设置

| 参数 | 含义 | 取值 |
|---|---|---|
| n_bins | 非缺失值的分箱数 | 255 |
| bin_type | 分箱类型 | percentile |
| max_layers | 级联图层的最大层数 | 20 |
| n_estimators | 每个级联图层中的估算器数 | 2 |
| n_trees | 每个估算器中的树数 | 100 |
| min_samples_split | 拆分内部节点所需的最小样本数 | 2 |
| min_samples_leaf | 叶节点上所需的最小样本数 | 1 |
| use_predictor | 是否构建连接到深度森林的预测变量 | False |
| predictor | 连接到深度森林的预测变量的类型 | forest |
| n_tolerant_rounds | 进行早停的层数 | 2 |
| delta | 早停的阈值 | 1e-5 |
| random_state | 随机种子标识 | 42 |
| base_model | 基学习器类型 | RandomForestClassifier<br>ExtraTreesClassifier<br>LogisticRegression<br>LGBMClassifier<br>CatBoostClassifier<br>XGBClassifier |

表 2: 残差深度森林参数设置

| 参数 | 含义 | 取值 |
|---|---|---|
| n_bins | 非缺失值的分箱数 | 255 |
| bin_type | 分箱类型 | percentile |
| max_layers | 级联图层的最大层数 | 20 |
| n_estimators | 每个级联图层中的估算器数 | 2 |
| n_trees | 每个估算器中的树数 | 100 |
| min_samples_split | 拆分内部节点所需的最小样本数 | 2 |
| min_samples_leaf | 叶节点上所需的最小样本数 | 1 |
| use_predictor | 是否构建连接到深度森林的预测变量 | False |
| predictor | 连接到深度森林的预测变量的类型 | forest |
| n_tolerant_rounds | 进行早停的层数 | 2 |
| delta | 早停的阈值 | 1e-5 |
| random_state | 随机种子标识 | 42 |
| base_model_name | 基学习器名称 | deep_forest |
| base_model_max_layers | 基学习器 (深度森林) 的最大层数 | 1 |
| layers | 残差深度森林的层数 | 3 |

表 3: 模型对比

| 模型 | Accuracy | Precision | Recall | F1 | AUC | KS |
|---|---|---|---|---|---|---|
| DeepForest | 93.75 | **99.78** | 87.60 | 93.29 | 96.76 | 87.48 |
| HGDeepForest | **93.80** | 99.69 | 87.77 | **93.35** | 96.98 | **87.52** |
| HGDeepForest_LR | 93.73 | 99.00 | 88.26 | 93.32 | **97.07** | 87.49 |
| ResDeepForest | 88.51 | 86.92 | **90.45** | 88.65 | 95.55 | 86.89 |
| HGResDeepForest | 93.60 | 98.67 | 88.29 | 93.19 | 96.78 | 87.35 |
| Stacking_LR | 93.69 | 98.70 | 88.45 | 93.30 | 96.93 | 87.38 |