# 【实践】随机森林算法参数解释及调优

原创 石头 机器学习算法那些事 2018-11-30

## 前言

上篇文章梳理了随机森林的各理论要点，本文首先详细解释了随机森林类的参数含义，并基于该类讲解了参数择优过程。

随机森林类库包含了RandomForestClassifer类，回归类是RandomForestRegressor类。RF的变种ExtraTress也有ExtraTressClassifier类和ExtraTressRegressor类。由于这四个类的参数基本相同，只要完全理解其中一个类，其他三个类很快就能上手。本文只介绍RandomForestClassifer类。

随机森林是基于bagging框架的决策树模型，因此随机森林的参数择优包括两部分：（1）RF框架的参数择优；（2）RF决策树的参数择优。因此，理解RF框架参数和决策树参数的含义是模型参数择优的前提。

## 目录



请参考Scikit-learn官网RandomForestClassifier类的参数来阅读前两节：

### RF框架参数含义

n_estimators：对原始数据集进行有放回抽样生成的子数据集个数，即决策树的个数。若n_estimators太小容易欠拟合，太大不能显著的提升模型，所以n_estimators选择适中的数值，版本0.20的默认值是10，版本0.22的默认值是100 。

bootstrp：是否对样本集进行有放回抽样来构建树，True表示是，默认值True

oob_score：是否采用袋外样本来评估模型的好坏，True代表是，默认值False，上篇文章提到袋外样本误差是测试数据集误差的无偏估计，所以推荐设置True。

RF框架的参数很少，框架参数择优一般是调节n_estimators值，即决策树个数。

### RF决策树参数含义

max_features：构建决策树最优模型时考虑的最大特征数。默认是“auto“，表示最大特征数是N的平方根；“log2”表示最大特征数是$\log_2 N$；"sqrt"表示最大特征数是$\sqrt{N}$。如果是整数，代表考虑的最大特征数；如果是浮点数，表示对（N*max_features）取整。其中N表示样本的特征数。

max_depth：决策树最大深度。若等于None，表示决策树在构建最优模型的时候不会限制子树的深度。如果模型样本量多，特征也多的情况下，推荐限制最大深度；若样本量少或者特征少，则不限制最大深度。

min_samples_leaf：叶子节点含有的最少样本。若叶子节点样本数小于min_samples_leaf，则对该叶子节点和兄弟叶子节点进行剪枝，只留下该叶子节点的父节点。整数型表示个数，浮点型表示取大于等于（样本数*min_samples_leaf）的最小整数。min_samples_leaf默认值是1。

min_samples_split：节点可分的最小样本数，默认值是2。整数型和浮点型的含义与min_samples_leaf类似。

max_leaf_nodes：最大叶子节点数。int设置节点数，None表示对叶子节点数没有限制。

min_impurity_decrease：节点划分的最小不纯度。假设不纯度用信息增益表示，若某节点划分时的信息增益大于等于min_impurity_decrease，那么该节点还可以再划分；反之，则不能划分。

criterion：表示节点的划分标准。不纯度标准参考Gini指数，信息增益标准参考"entrop"熵。

min_samples_leaf：叶子节点最小的样本权重和。叶子节点如果小于这个值，则会和兄弟节点一起被剪枝，只保留该叶子节点的父节点。默认是0，则不考虑样本权重问题。一般来说，如果有较多样本的缺失值或偏差很大，则尝试设置该参数值。

## RF参数择优实例

**RF参数择优思想**：RF模型可以理解成决策树模型嵌入到bagging框架，因此，我们首先对外层的bagging框架进行参数择优，然后再对内层的决策树模型进行参数择优。在优化某一参数时，需要把其他参数设置为常数。

**（1）训练数据集下载：**

```
X, y = make_classification(n_samples=1000,n_features=50,
                           n_clusters_per_class=1,n_informative=15,
                           random_state=RANDOM_STATE)
```

make_classification构建样本数1000和特征数50的二分类数据 。

**所有参数都采用默认值，查看分类情况：**

```
rf0 = RandomForestClassifier(oob_score=True,random_state=10)
rf0.fit(X,y)
print rf0.oob_score_
print "accuracy: %f" % rf0.oob_score_
```

准确率为：

```
accuracy: 0.823000
```

**（2）对外层的bagging框架进行参数择优，即对n_estimators参数择优，其他参数仍然是默认值。**

n_esimators参数择优的范围是：1~101，步长为10。十折交叉验证率选择最优n_estimators 。

```
param_test1 = {"n_estimators":range(1,101,10)}
gsearch1 = GridSearchCV(estimator=RandomForestClassifier(),param_grid=param_test1,
                        scoring="roc_auc",cv=10,)
gsearch1.fit(X,y)
print gsearch1.grid_scores_
print gsearch1.best_params_
print "best accuracy:%f" % gsearch1.best_score_
```

输出结果：

```
{'n_estimators': 81}
best accuracy:0.986152
```

因此，最佳的子决策树个数是81，准确率98.61%，相比默认参数的82.3%有较大的提高。

**（3）优化决策树参数的最大特征数max_features，其他参数设置为常数，且n_estimators为81。**

max_features参数择优的范围：1~11，步长为1，十折交叉验证率选择最优max_features 。

```
param_test2 = {"max_features":range(1,11,1)}
gsearch1 = GridSearchCV(estimator=RandomForestClassifier(n_estimators=81,random_state=10),
                        param_grid=param_test2,scoring="roc_auc",cv=10)
gsearch1.fit(X,y)
print gsearch1.grid_scores_
print gsearch1.best_params_
print "best accuracy:%f" % gsearch1.best_score_
```

**结果：**

```
{'max_features': 6}
best accuracy:0.986399
```

因此，选择最佳的最大特征数为6，准确率为98.63%，相比默认的最大特征数，准确率有一个非常小的提高。决策树的其他最优参数也是按照类似的步骤去搜寻，这里就不一一介绍了。

**（4）用最优参数重新训练数据，计算泛化误差：**

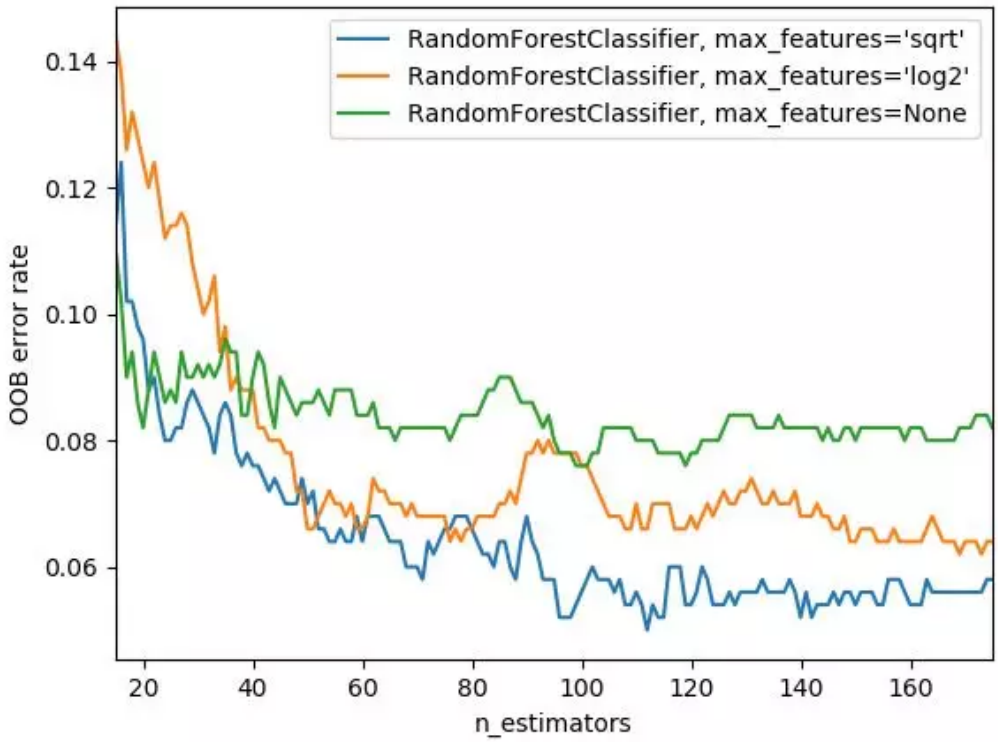
```
rf0 = RandomForestClassifier(n_estimators=81,max_features=6,
                             oob_score=True,random_state=10)
rf0.fit(X,y)
print rf0.oob_score_
print "accuracy: %f" % rf0.oob_score_
```

泛化误差：

```
accuracy: 0.928000
```

**总结**

随机森林模型优化主要是考虑如何选择子数据集个数（n_estimators）和最大特征个数（max_features），参数优化顺序可参考下图：




首先增大n_estimators，提高模型的拟合能力，当模型的拟合能力没有明显提升的时候，则再增大n_estimaors，提高每个子模型的拟合能力，则相应的提高了模型的拟合能力。上节的参数调优是比较常用的一种参数调优方法，可应用到其他模型的参数优化过程。

**参考**

https://scikit-learn.org

https://www.cnblogs.com/pinard/p/6160412.html?utm_source=itdadao&utm_medium=referral

**-END-**

长按二维码关注

机器学习算法那些事

微信：beautifulife244

砥砺前行 不忘初心