比较全面的随机森林算法总结

原创 石头 机器学习算法那些事 2018-11-29

前言

上节介绍了集成学习方法包括bagging法和boosting法，随机森林是基于bagg ing框架的决策树模型，本文详细的总结了随机森林算法，尽可能的让大家对随机森林有一个全面的认识。

目录

随机森林的算法流程

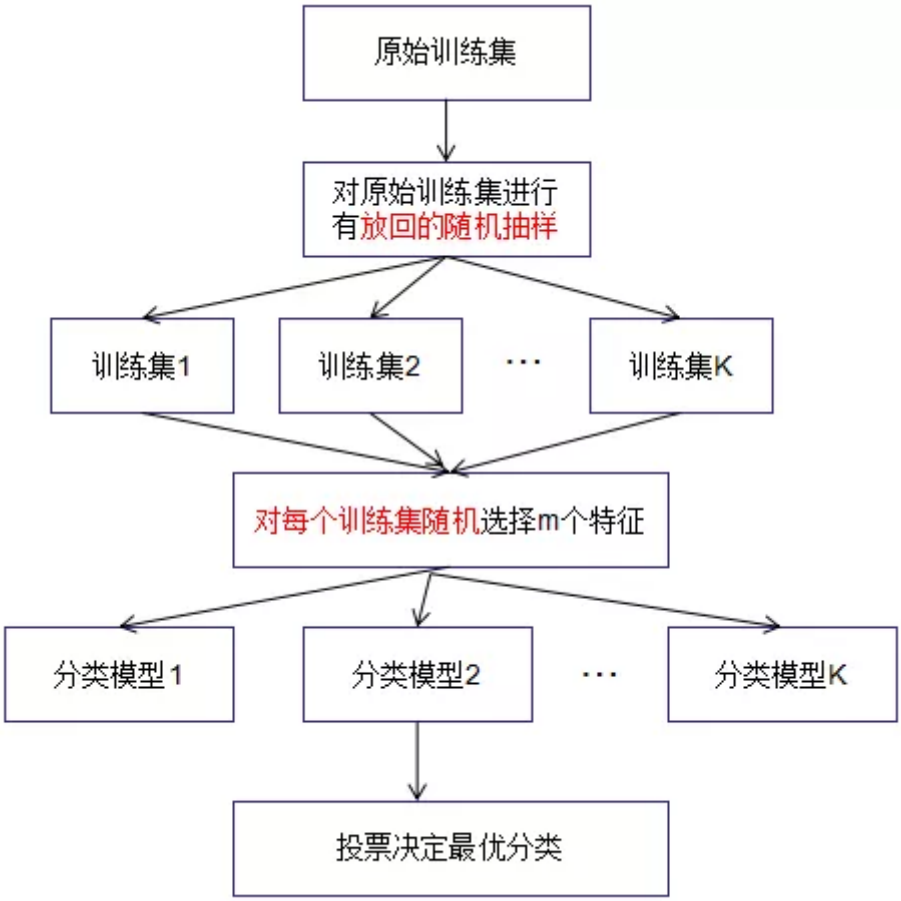
随机森林是基于bagging框架下的决策树模型，随机森林包含了很多树，每棵树给出分类结果，每棵树的生成规则如下：

（1）如果训练集大小为N，对于每棵树而言，随机且有放回地从训练中抽取N个训练样本，作为该树的训练集，重复K次，生成K组训练样本集。

（2）如果每个特征的样本维度为M，指定一个常数m<<M，随机地从M个特征中选取m个特征。

（3） 利用m个特征对每棵树尽最大程度的生长，并且没有剪枝过程。

随机森林的分类算法流程如下图：



随机森林的应用场景

> 吴恩达老师在《机器学习》公开课讲过，如何优化当前的机器学习模型，首先你要知道当前的模型是处于高方差状态还是高偏差状态，高方差需要增加训练数据或降低模型的复杂度，高偏差则需要优化当前模型，如增加迭代次数或提高模型的复杂度等。

随机森林是基于bagging思想的模型框架，我们从bagging角度去探讨随机森林的偏差与方差问题，给出应用场景。

随机森林的偏差与方差讨论：

随机森林对每一组重采样的数据集训练一个最优模型，共K个模型。
令Xi为随机可放回抽样的子数据集的N维变量，i = 1,...,K 。
根据可放回抽样中子数据集的相似性以及使用的是相同的模型，因此各模型有近似相等的 bias和variance，且模型的分布也近似相同但不独立（因为子数据集间有重复的变量）。因此：

$$E[\frac{\sum X_i}{K}] = E[X_i]$$

由上式得：bagging法模型的bias和每个子模型接近，因此，bagging法并不能显著降低bias。
a) 极限法分析bagging法模型的方差问题：

若模型完全独立，则：

$$Var(\frac{\sum X_i}{K}) = \frac{Var(X_i)}{K} \quad ,(1)$$ （上一篇文章等式右边的分母为K^2，此处存在错误，抱歉）

若模型完全一样，则：

$$Var(\frac{\sum X_i}{K}) = Var(X_i) \quad ,(2)$$

因为bagging的子数据集既不是相互独立的，也不是完全一样的，子数据集间存在一定的相似性，因此，bagging法模型的方差介于（1）（2）式两者之间。

b) 公式法分析bagging法模型的方差问题：

假设子数据集变量的方差为 σ^2，两两变量之间的相关性为 ρ

所以，bagging法的方差：

$$Var(\frac{\sum X_i}{K}) = \frac{1}{K^2} Var(\sum X_i)$$

$$\Rightarrow \frac{1}{K^2}(KVar(X_1) + 2\sum_{i=1}^{K}\sum_{j=1}^{K}\mathrm{cov}(X_i, X_j)_{i\neq j} \quad (3)$$

$$\because \quad \rho = \frac{\mathrm{cov}(X_i, X_j)}{\sqrt{X_i}\sqrt{X_j}}$$

$\therefore$ （3）式得：

$$\Rightarrow \frac{1}{K^2}(K \cdot Var(X_1) + 2\sum_{i=1}^{K}\sum_{j=1}^{K}\mathrm{cov}(X_i, X_j)_{i\neq j})$$

$$\Rightarrow \frac{1}{K^2}(K \cdot Var(X_1) + 2\sum_{i=1}^{K}\sum_{j=1}^{K}\rho \cdot \sigma^2)$$

$$\therefore \quad Var(\frac{\sum X_i}{K}) = \rho \cdot \sigma^2 + (1-\rho)\sigma^2 / K \quad (4)$$

由（4）式可得，bagging法的方差减小了。

结论：bagging法的模型偏差与子模型的偏差接近，方差较子模型的方差减小。所以，随机森林的主要作用是降低模型的复杂度，解决模型的过拟合问题。

随机森林的相关性理解

随机森林的相关性包括子数据集间的相关性和子数据集间特征的相关性。相关性在这里可以理解成相似度，若子数据集间重复的样本或子数据集间重复的特征越多，则相关性越大。

随机森林分类效果（错误率）与相关性的关系：

（1）森林中任意两棵树的相关性越大，错误率越大；

（2）减小子数据间的特征选择个数，树的相关性和分类能力也会相应的降低；增大特征个数，树的相关性和分类能力会相应的提高。

结论：（1）是随机有放回抽取的，相关性大小具有随机性，因此，特征个数是优化随机森林模型的一个重要参数。

随机森林蕴含的思想

我们再回顾随机森林学习模型的步骤：

（1）对原始数据集进行可放回随机抽样成K组子数据集；

（2）从样本的N个特征随机抽样m个特征；

（3）对每个子数据集构建最优学习模型

（4）对于新的输入数据，根据K个最优学习模型，得到最终结果。

思想：（2）的随机抽样的结果是子数据集间有不同的子特征，我们把不同的特征代表不同的领域，（3）表示在不同领域学习到最顶尖的程度，（4）表示对于某一个输入数据，用不同领域最顶尖的观点去看待输入数据，得到比较全面的结果。

随机森林的模型估计方法

对于包含m个样本的原始数据集，对该原始数据集进行可放回抽样m次，每次被采集到的概率是1/m，不被采集到的概率是(1-1/m)。m次采样不被抽到的概率是 $(1-\frac{1}{m})^m$ 。

当 $m \to \infty$ 时，$(1-\frac{1}{m})^m \to \frac{1}{e} \approx 0.368$ 。因此在bagging的每轮抽样中，训练集大约有36.8%的数据没有被采样，这份数据称之为袋外数据（Out Of Bag，简称OOB）。

Breiman在随机森林的论文中证明了袋外数据(OOB)误差估计是一种可以取代测试集的误差估计方法，即袋外数据误差是测试数据集误差的无偏估计，因此可以用OOB数据用来检测模型的泛化能力。

Table 7. Error and OB Estimates

Data Set	Test Error	OB Error	PE*(tree)	Cor.
Boston Housing	10.2	11.6	26.3	.45
Ozone	16.3	17.6	32.5	.55
Servo x10-2	24.6	27.9	56.4	.56
Abalone	4.6	4.6	8.3	.56
Robot Arm x10-2	4.2	3.7	9.1	.41
Friedman #1	5.7	6.3	15.3	.41
Friedman #2x10+3	19.6	20.4	40.7	.51
Friedman #3x10-3	21.6	22.9	48.3	.49

备注：如上图，测试数据集的误差和OOB误差很接近。

论文下载地址

https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

总结：模型的估计方法有两种，（1）袋外数据误差估计模型，（2）交叉验证率估计模型。

总结

目前，集成式学习方法的框架比较火，应用非常广。本文详细总结了随机森林算法的各个理论要点，我对学习随机森林的看法是：随机森林原理简单，但是知识点很杂，需要有耐心去深入理解它。

参考：

https://blog.csdn.net/wishchin/article/details/52515516

https://www.cnblogs.com/maybe2030/p/4585705.html

推荐阅读文章

【干货】集成学习原理总结

决策树算法总结

-END-

文章已于2018-11-29修改