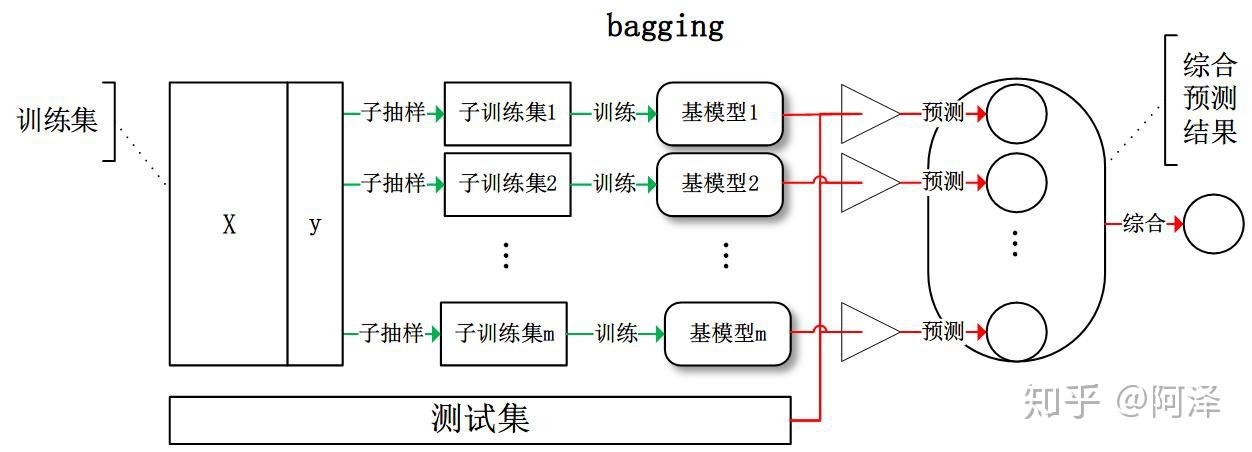
有关模型集成方法的讨论与总结

吴锐 李林 李震

**Bagging**

bagging在训练过程中，各基分类器之间无强依赖，可以并行训练，如随机森林。为了让基分类器之间相互独立，将训练集分为若干子集（子集之间有交集）。Bagging方法是一个集体学习的过程，每个个体进行单独学习，学习的内容可以相同，也可以不同，也可以部分重叠。但由于个体之间存在差异性，最终做出的判断不会完全一致。在最终做出决策时，每个个体单独做出判断，在通过投票的方式做出最后的集体决策。

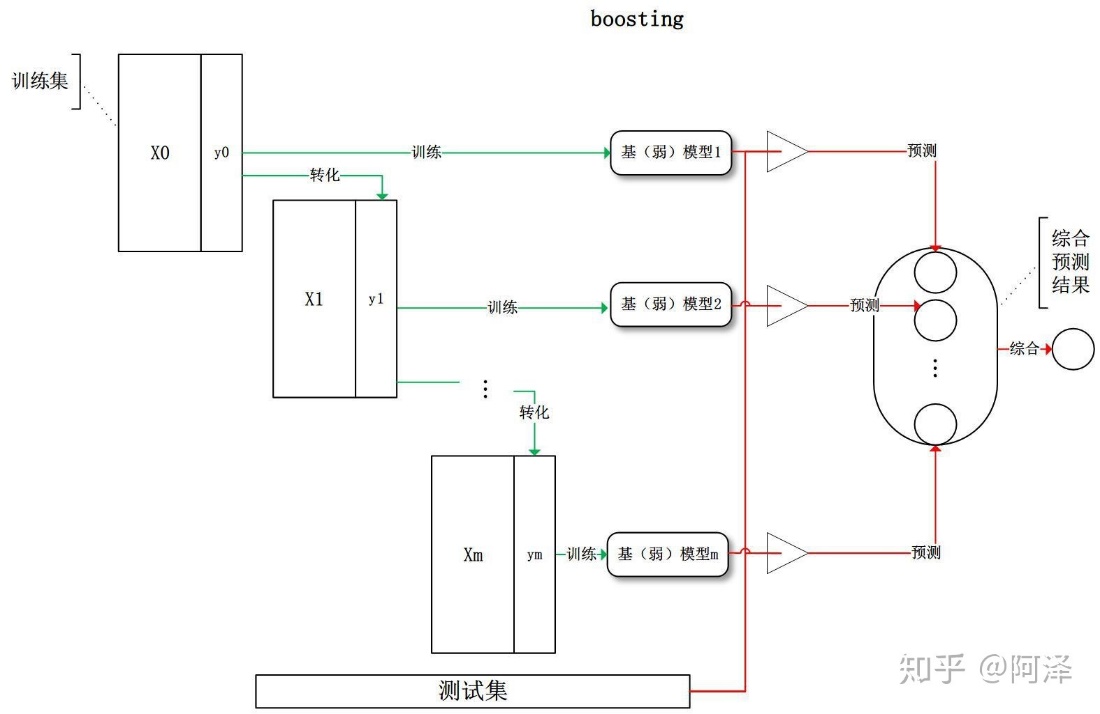


并行集成学习，基学习器对数据集采样（0.632），每个基学习器对采样数据进行训练，综合所有基学习器的预测结果（平均或投票）得到最终预测结果。

随机森林是Bagging的一个代表，简单高效，除了数据采样带来的样本多样性，还有不同决策树随机选择的属性多样性。

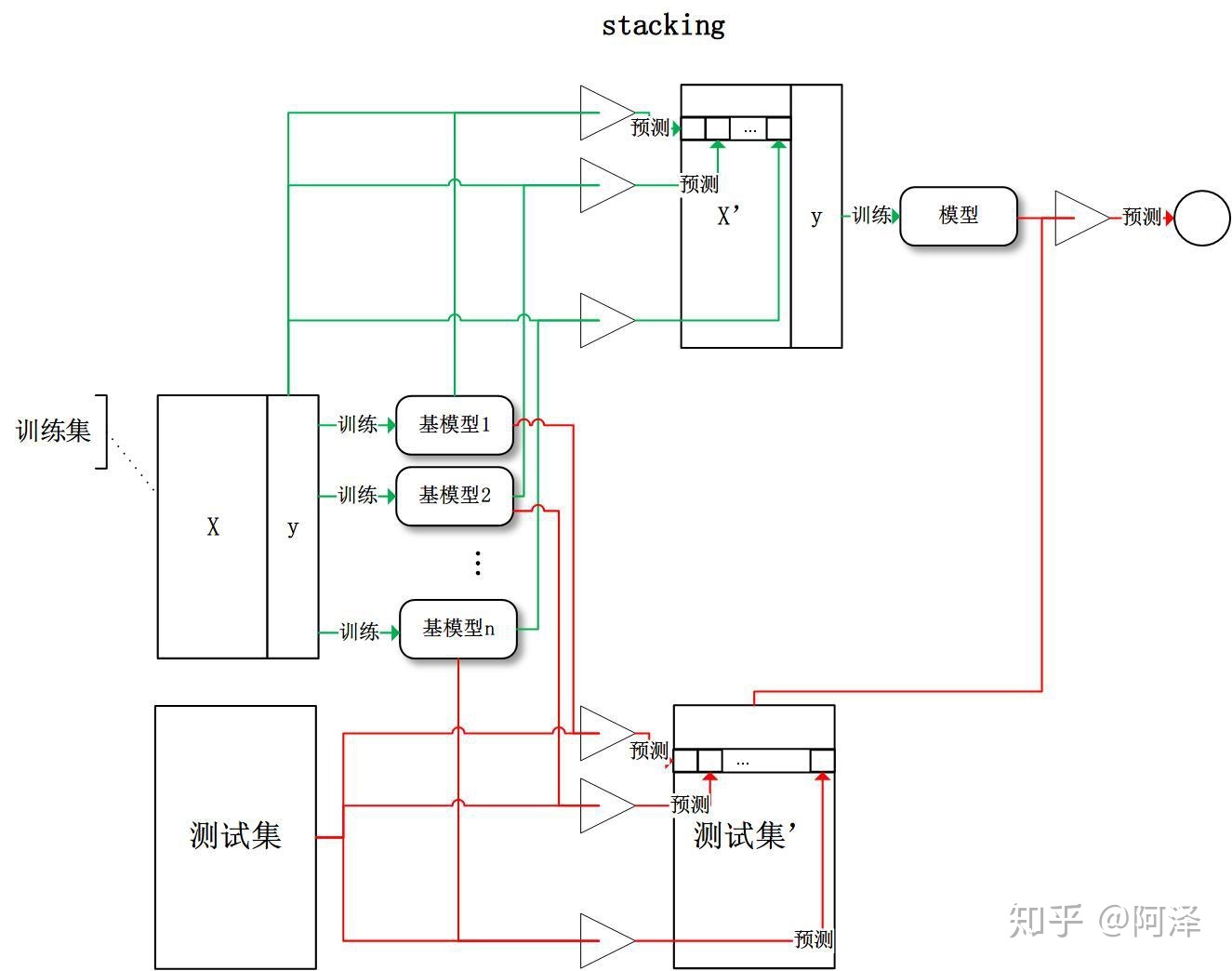
**Boosting**

基本思路：将基分类器层层叠加，每一层在训练的时候对前一层基分类器分错的样本给予更高的权重。测试时，根据各层分类器的结果加权得到最终结果。迭代式学习。



提升集成学习，将弱学习器提升为强学习器，根据初始数据训练出基学习器，然后根据基学习器的表现对数据分布调整，更加关注错例样本，学习下一个基学习器；迭代T次，最终把T个基学习器加权求和。AdaBoost即Boosting的代表，就是对每次迭代得到的基学习器加权求和的过程，后面迭代的学习器应该对前面迭代的学习器的错误进行纠正。Boosting主要关注降低偏差。

**Stacking**

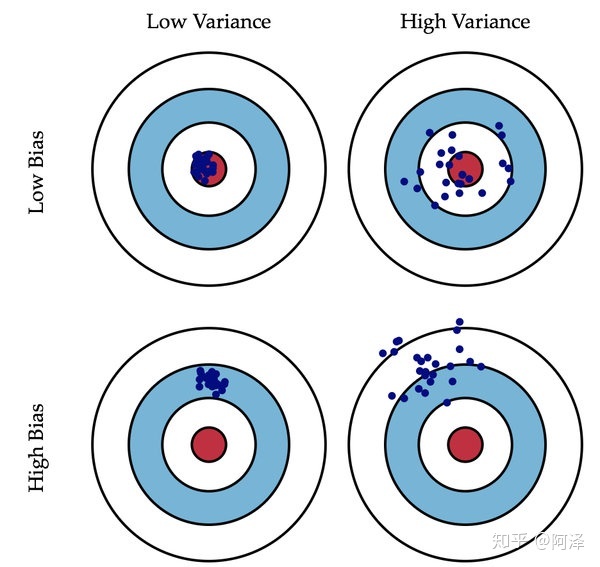


初始学习器的预测值作为训练数据，原始训练数据的标签保持不变，然后在此基础上进行训练。初始训练集和次级训练集应该不相交（比如用交叉验证中的验证集构建次级训练集），防止过拟合；本质上，stacking就是对输出结果进行拟合的过程。

**集成学习的偏差与方差**

**偏差**（学习器预测值与真实值的偏离程度）

**方差**（学习器之间预测值的离散程度）



弱学习器偏差高，可以通过类似于Boosting的集成方式，不断的对错例调整，使得预测值接近真实值。强学习器可能对自己看到的数据过拟合而出现不同强学习器预测值方差高，可以通过Bagging投票或平均的方式进行集成。

偏差：通常是我们对模型做了错误的假设所导致的，体现在训练误差上。（适合用boosting方式来解决）

方差：通常是由于模型的复杂度相对于训练样本数过高导致的。体现在测试误差相对于训练误差的增量上。（适合用bagging方式来解决）

集成学习步骤：

1. 找到误差相互独立的基分类器。一般选取决策树作为基分类器，因为树型模型简单且容易产生随机性。
2. 训练基分类器
3. 合并基分类器的结果

常用的基分类器是什么？

1. 最常用的基分类器是决策树，主要有以下3个方面的原因。

（1）决策树可以较为方便地将样本的权重整合到训练过程中，而不需要使用 过采样的方法来调整样本权重。

（2）决策树的表达能力和泛化能力，可以通过调节树的层数来做折中。

（3）数据样本的扰动对于决策树的影响较大，因此不同子样本集合生成的决策树基分类器随机性较大，这样的“不稳定学习器”更适合作为基分类器。此外， 在决策树节点分裂的时候，随机地选择一个特征子集，从中找出最优分裂属性， 很好地引入了随机性。

1. 神经网络模型也适合作为基分类器，主要由于神经网络模型 也比较“不稳定”，而且还可以通过调整神经元数量、连接方式、网络层数、初始 权值等方式引入随机性。