集成学习

张翼鹏

November 28, 2019

目录

1 基本介绍

1.1 什么是集成学习

在机器学习的有监督学习算法中，我们的目标是学习出一个稳定且在各个方面表现都较好的模型，但实际情况往往不这么理想，有时我们只能得到多个有偏好的模型（弱监督模型，在某些方面表现得较好）。集成学习就是要组合多个弱监督模型以期得到一个更好更全面的强监督模型。集成学习潜在的思想是即便某一个弱分类器得到了错误的预测，其他的弱分类器也可以将错误纠正回来。

集成方法是将几种机器学习技术组合成一个预测模型的元算法，以达到减小方差（bagging）、减小偏差（boosting）或改进预测（stacking）的效果。

集成学习在各个规模的数据集上都有很好的策略。数据集较大时，可以划分成多个小数据集，学习多个模型进行组合；数据集较小时，可以利用Bootstrap方法进行抽样，得到多个数据集，分别训练多个模型再进行组合。

1.2 集成学习的分类

一般来说集成学习可以根据其效果分为三大类：

- 用于减少方差的Bagging
- 用于减少偏差的Boosting
- 用于提升预测结果的Stacking

也可以按数据传输形式分为如下两类：

- 串行集成方法，这种方法串行地生成基础模型（如AdaBoost）。串行集成的基本动机是利用基础模型之间的依赖，通过给错分样本一个较大的权重来提升性能。
- 并行集成方法，这种方法并行地生成基础模型（如Random Forest）。并行集成的基本动机是利用基础模型的独立性，因为通过平均能够较大地降低误差。

2 Bagging（套袋）

思想：Bagging即套袋法，在训练集中进行子抽样作为每个基模型需要的子训练集，对所有基模型的预测结果进行综合从而产生最终的预测结果。主要利用bootstrap自助法进行有放回抽样，目的是为了得到统计量的分布以及置信区间。

步骤：1、从原始样本集中抽取训练集。每轮从原始样本集中使用Bootstraping的方法抽取n个训练样本（在训练集中，有些样本可能被多次抽取到，而有些样本可能一次都没有被抽中）。共进行k轮抽取，得到k个相互独立的训练集。

2、每次使用一个训练集得到一个模型，k个训练集共得到k个模型（可以取全部的特征进行训练，也可以随机选取部分特征训练，例如随机森林就是每次随机选取部分特征；这里并没有具体的分类算法或回归方法，可以根据具体问题采用不同的方法，如决策树、感知器等，但k次训练使用的是同一模型）。

3、对分类问题：将上步得到的k个模型采用投票的方式得到分类结果；对回归问题，计算上述模型的均值作为最后的结果。

评价：1、Bagging通过降低基分类器的方差，改善了泛化误差。

2、其性能依赖于基分类器的稳定性。如果基分类器不稳定，Bagging有助于降低训练数据的随机波动导致的误差；如果稳定，则集成分类器的误差主要由基分类器的偏倚引起。

3、由于每个样本被选中的概率相同，因此Bagging并不侧重于训练数据集中的任何特定实例。

3 Boosting（提升）

思想：Boosting的主要思想是将若干弱分类器组装成一个强分类器。训练过程中利用数据的加权，对于错分数据给予较大的权重。

步骤：1、从初始训练集训练出一个基学习器。

2、根据基学习器的表现对训练样本分布进行调整（提高前一轮分错样例的权值，减小前一轮分对样例的权值），使得先前基学习器做错的训练样本在后续受到更多关注。

3、基于调整后的样本分布来训练下一个基学习器。

4、重复进行上述步骤，直至基学习器数目达到事先指定的值k，最终将这k个基学习器进行加权结合。

4 Stacking（堆叠）

思想：Stacking方法是通过一个元分类器或者元回归器来整合多个分类模型或回归模型的集成学习技术。首先我们先训练多个不同的模型（各个模型使用的算法不同），然后把之前训练的各个模型的输出作为输入来训练一个模型，以得到一个

最终的输出。