Stacking 方法介绍

在机器学习过程中，通常会遇到这样一类问题，模型在训练集上表现效果较好，但在测试集上表现效果较差，一般这种情况是由于模型过拟合所导致的。为了在保持较高精度的情况下减小过拟合的影响，集成学习算法的思想被提出来。集成学习算法一般通过融合精度较高并且个体之间差异度较大的学习器产生，并且在大多数情况下可以显著提高学习器的泛化能力，stacking算法是众多集成学习算法中的一种。一般而言，stacking算法能够取得比个体学习器更高的精度，且没有过拟合现象的产生。

Stacking算法的思想：

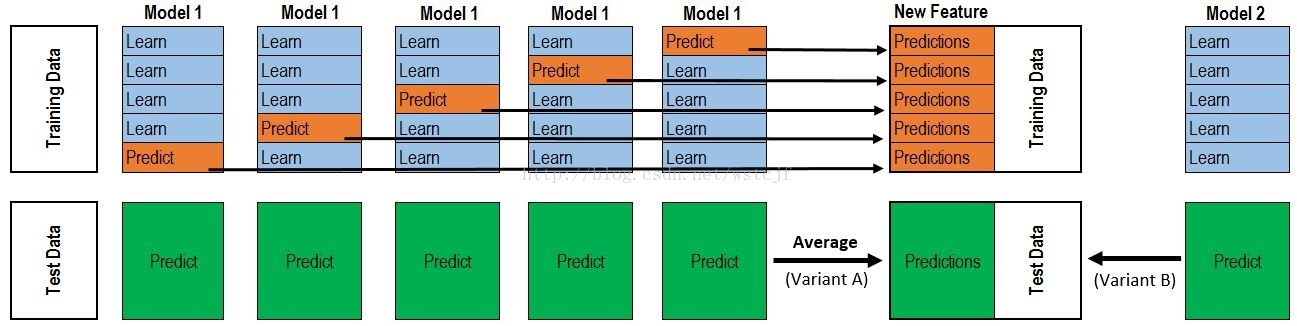


图2 stacking算法

图2为stacking算法的经典描述，本文将以该图为例介绍stacking算法。

Stacking算法包括两层：第一层首先使用单个学习器对数据进行预测，第二层把第一层的预测结果作为数据，再使用学习器进行预测。

第一层：假如Training Data有10000个数据，Test Data有3000个数据。用Model1对Training Data进行5折交叉验证，将Training Data 分为5部分，每部分有2000个数据，先利用前面1,2,3,4部分的8000个数据对model1进行训练，记得到的模型为model11，然后利用1,2,4,5部分的8000个数据得到model12，以此类推，可以得到model13，model14，model15。接着利用model11对第五部分数据进行预测，得到第五部分的2000个预测值，利用model12对第四部分数据进行预测，得到第四部分的2000个预测值，以此类推，可以得到整个Training Data的预测值，记为A1。同样可以利用model11至model15，分别对Test Data进行预测，并将预测值取平均，即可当得到Test Data的一组预测值，一个3000行一列的数据，记为B1。

同理可利用model2产生A2，B2，利用model3产生A3，B3。将A1，A2，A3列合并，产生新的Training Data，将B1，B2，B3列合并，产生新的Test Data。

第二层：使用新的Training Data训练模型，使用新的Test Data验证模型。