第八章　集成学习

Castor Ye

1　个体与集成

集成学习（ensemble learning）通过构建并结合多个学习器来完成学习任务，有时也被称为多分类器系统（multi-classfier system）、基于委员会的学习（committee-based learning）等。

图 1 显示出集成学习的一般结构：先产生一组“个体学习器”（in 地 vi learner），再用某种策略将它们结合起来。

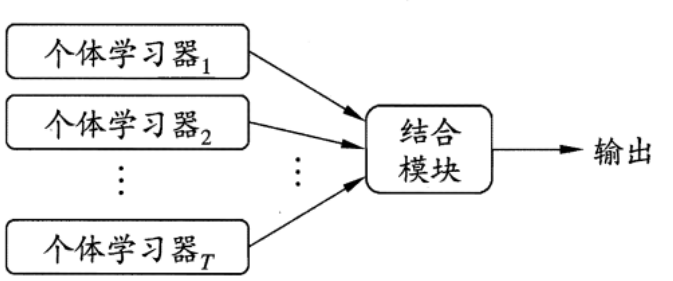

图 1: 集成学习示意图

在集成模型中，若个体学习器都属于同一类别，例如都是决策树或神经网络，则称该集成为“同质”的（homogenneous）。同质集成中的个体学习器亦称“基学习器”（base learner），相应的学习算法称为“基学习算法”（base learning algorithm）。若个体学习器为不同类型的，则称该集成为“异质”的（heterogenous）。异质集成中的个体学习器由不同的学习算法生成，此时不再有基学习器，而称“组件学习器”（component learner）或直接称“个体学习器”。

上面我们已经提到要让集成起来的泛化性能比单个学习器好，但也存在短板效应，所以我们引入两个重要概念：准确性和多样性（diversity）。准确性指的是个体学习器不能太差，要有一定的准确度；多样性则是个体学习器之间的输出要具有差异性。

	测试例1	测试例2	测试例3
h_1	√	√	×
h_2	×	√	√
h_3	√	×	√
集成	√	√	√

(a) 集成提升性能

	测试例1	测试例2	测试例3
h_1	√	√	×
h_2	√	√	×
h_3	√	√	×
集成	√	√	×

(b) 集成不起作用

	测试例1	测试例2	测试例3
h_1	√	×	×
h_2	×	√	×
h_3	×	×	√
集成	×	×	×

(c) 集成起负作用

图 2: 集成个体应“好而不同”

0