第10章：集成学习

蒋良孝 |  中国地质大学（武汉）

机器学习与数据挖掘团队

ljiang@cug.edu.cn

http://grzy.cug.edu.cn/jlx/

本章内容

一、集成学习基础知识

二、集成学习常用方法

三、集成学习结合策略

一、集成学习基础知识

- 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-classifier system)、基于委员会的学习(committee-based learning) 等。
- 集成学习先产生一组“个体学习器”(individual learner)，再用某种策略将它们结合起来。
- 集成学习分同质集成和异质集成。同质集成中的个体学习器由相同的学习算法生成，个体学习器称为基学习器；异质集成中的个体学习器由不同的学习算法生成，个体学习器称为组件学习器。

一、集成学习基础知识

- 集成学习要显著优于单一个体学习器必须满足两个必要条件：1）个体学习器之间应该是相互独立的；2）个体学习器应当好于随机猜测学习器。
- 满足第2个条件往往比较容易，因为在现实任务中，出于种种考虑，比如希望使用较少的个体学习器，或者是希望重用关于常见学习器的一些经验等，人们往往会使用比较强的个体学习器。
- 满足第1个条件往往比较困难，个体学习器是为解决同一个问题训练出来的，显然不可能互相独立！事实上，个体学习器的“准确性”和“多样性”本身就存在冲突。一般的，准确性很高之后，要增加多样性就需要牺牲准确性。

因此，如何产生“好而不同”的个体学习器是集成学习研究的核心！

二、集成学习常用方法

- **那么，如何在保持个体学习器足够“好”的前提下增强多样性呢？**一般的思路是在学习过程中引入随机性，常用方法主要包括：

✓ 训练样本扰动

✓ 输入属性扰动

✓ 输出标记扰动

✓ 算法参数扰动

✓ 混合扰动

二、集成学习常用方法

1）训练样本扰动

- 训练样本扰动通常是用抽样的方法从原始训练样本集中产生出不同的样本子集，然后再利用不同的样本子集训练出不同的个体学习器。比如，在装袋Bagging中使用自助采样，在提升Boosting中使用序列采样。
- 此类方法简单高效，使用也最广，但只对不稳定基学习器有效，比如决策树、神经网络等；对稳定基学习器效果不明显，比如线性学习器、支持向量机、朴素贝叶斯、k-最近邻学习器等。

二、集成学习常用方法

2）输入属性扰动

- 输入属性扰动通常是从初始属性集中抽取出若干个属性子集，然后利用不同的属性子集训练出不同的个体学习器。比如，随机子空间[Ho, 1998]和随机森林[Breiman, 2000]。
- 此类方法对包含大量冗余属性的数据集有效，但若数据集只包含少量属性，或者冗余属性很少，则不宜使用。

二、集成学习常用方法

3）输出标记扰动

- 输出标记扰动通常是对训练样本的类标记稍作变动，比如，可将原来的多分类问题随机转化多个二分类问题来训练基学习器，纠错输出码[Dietterich and Bakiri, 1995]就是这类方法的典型代表。
- 此类方法对类数足够多的数据集有效，但若数据集包含的类数较少，则不宜使用。

二、集成学习常用方法

4）算法参数扰动

- 算法参数扰动通常是通过随机设置不同的参数来训练差别较大的个体学习器。比如，神经网络的隐层神经元数、初始连接权值等
- 此类方法对参数较多的算法有效，对参数较少的算法，可通过将其学习过程中某些环节用其他类似方式代替?从而达到扰动的目的。

二、集成学习常用方法

5）混合扰动

- 混合扰动是指在在同一个集成算法中同时使用上述多种扰动方法。比如，随机森林就同时使用了训练样本扰动和输入属性扰动。

三、集成学习结合策略

- 到此为止，我们都是在讨论：**怎么通过集成学习的常用方法生成“好而不同”的个体学习器？**
- 剩下来要解决的问题：**怎么结合生成的个体学习器，具体的结合策略有哪些？**
- 对分类任务来说，最常见的结合策略就是投票法(voting)，具体包括：

绝对多数投票法(majority voting)

相对多数投票法(plurality voting)

加权投票法(weighted voting)

三、集成学习结合策略

- 绝对多数投票法：即若某标记得票过半数，则分类为该标记，否则拒绝分类。
- 相对多数投票法：分类为得票最多的标记，若同时有多个标记获最高票，则从中随机选取一个。
- 加权投票法：给每个个体学习器预测的类标记赋一个权值，分类为权值最大的标记。这里的权值通常为该个体学习器的分类置信度（类成员概率）。