# 集成学习

## 集成学习的分类

集成学习分为两种：投票学习(bagging)和再学习(boosting)

### Bagging

投票选举(bagging)是基于数据随机重抽样分类器构造的方法。虽然使用相同的模型，但是数据和特征的选择都加入了随机化，这样训练得到多个不同的分类器，最终根据分类器的投票结果决定最终输出，所有分类器的权重是相同的。

### Boosting

再学习(boosting)是基于所有分类器的加权求和的方法。通过调整已有分类器错分的那些数据来获得新的分类器。Boosting的结果通过分类器加权求和得到，所以权重并不相等，每个权重代表的是其对应分类器在上一轮迭代中的成功度。

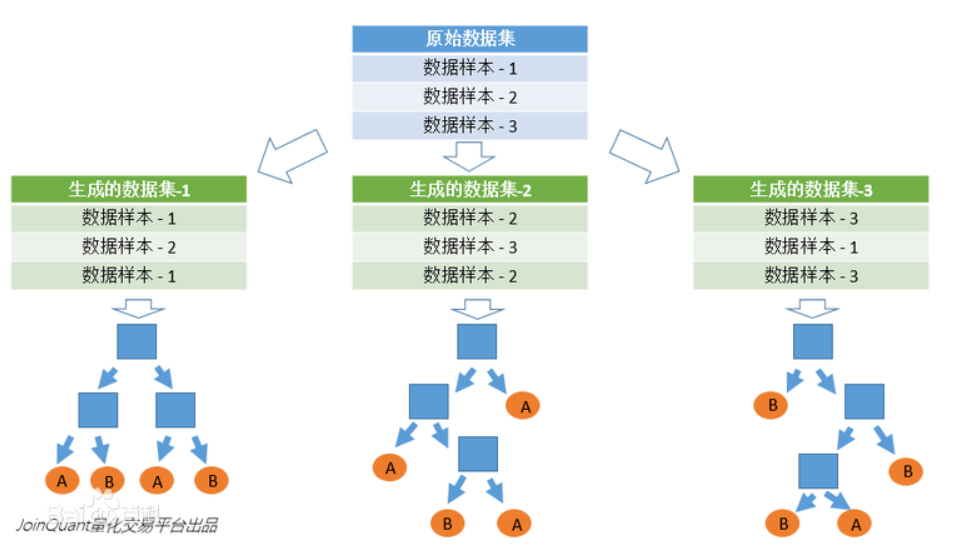
## Bagging样例：随机森林

### 整体概述

前提：完整训练集、特征集合、分类器(决策树)

主要流程：

1. 从训练集中抽取若干子集，分别用于训练不同的分类器
2. 分类器使用子集进行训练，选取特征划分子树时也加上随机化
3. 最终训练得到若干个分类器，他们的投票结果最为随机森林的最终结果

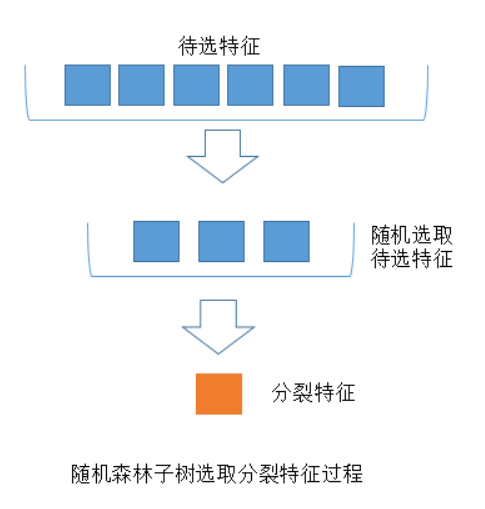


### 数据随机化

策略：有放回的随机采样，不同子集的元素个数相同，同一子集的元素可以重复

### 特征随机化

在训练分类器的过程中进行特征随机化。



流程：

1. 从待选特征中随机抽取特征子集
2. 利用决策树的方法(信息增益)从特征子集中选取最优特征
3. 根据最优特征划分子树
4. 从待选特征中删除最优特征
5. 对子树递归1~5

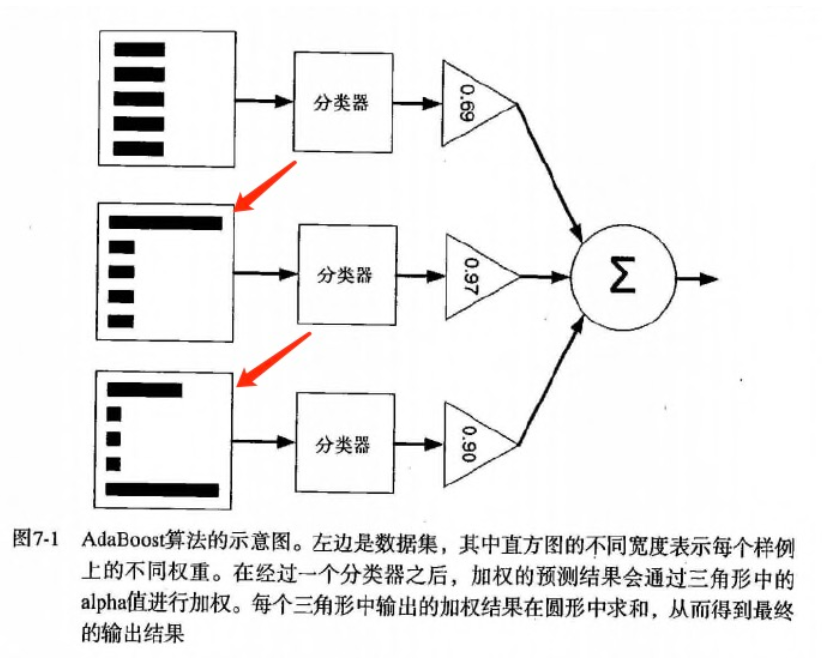
## Boosting样例：AdaBoost

### 参数说明

第i个分类器，数据权重，误差率，分类器的权重

### 主要流程

1. 初始数据权重(比如权重全相同)，训练弱分类器
2. 求的误差率，计算弱分类器在强分类器中的权重
3. 更新数据权重
4. 重复1~3



### 计算公式

弱分类器权重

某个样本j被正确分类：

某个样本j被错误分类：

### 数据权重的应用

分类器可以是决策树或者其他模型，数据权重的作用在于计算误差率。

比如对于每个数据，预测正确记为0，预测错误记为1，得到一个errArr矩阵，这个矩阵再乘以数据权重矩阵D，然后求和sum(D)就可以作为一种误差率。