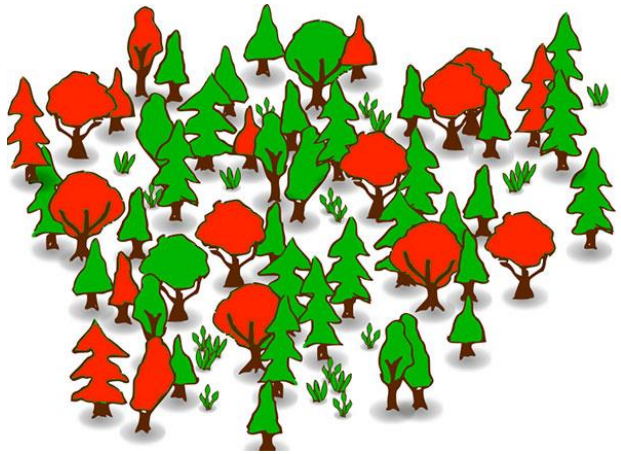
随机森林（Random forest）

- 随机森林指的是利用多棵树对样本进行训练并预测的一种分类器。该分类器最早由Leo Breiman和Adele Cutler提出。
- 随机森林顾名思义，是用随机的方式建立一个森林，森林里面有很多的决策树组成，随机森林的每一棵决策树之间是没有关联的。在得到森林之后，当有一个新的输 入样本进入的时候，就让森林中的每一棵决策树分别进行一下判断，看看这个样本应该属于哪一类（对于分类算法），然后看看哪一类被选择最多，就预测这个样本 为那一类。



随机森林（Random forest）

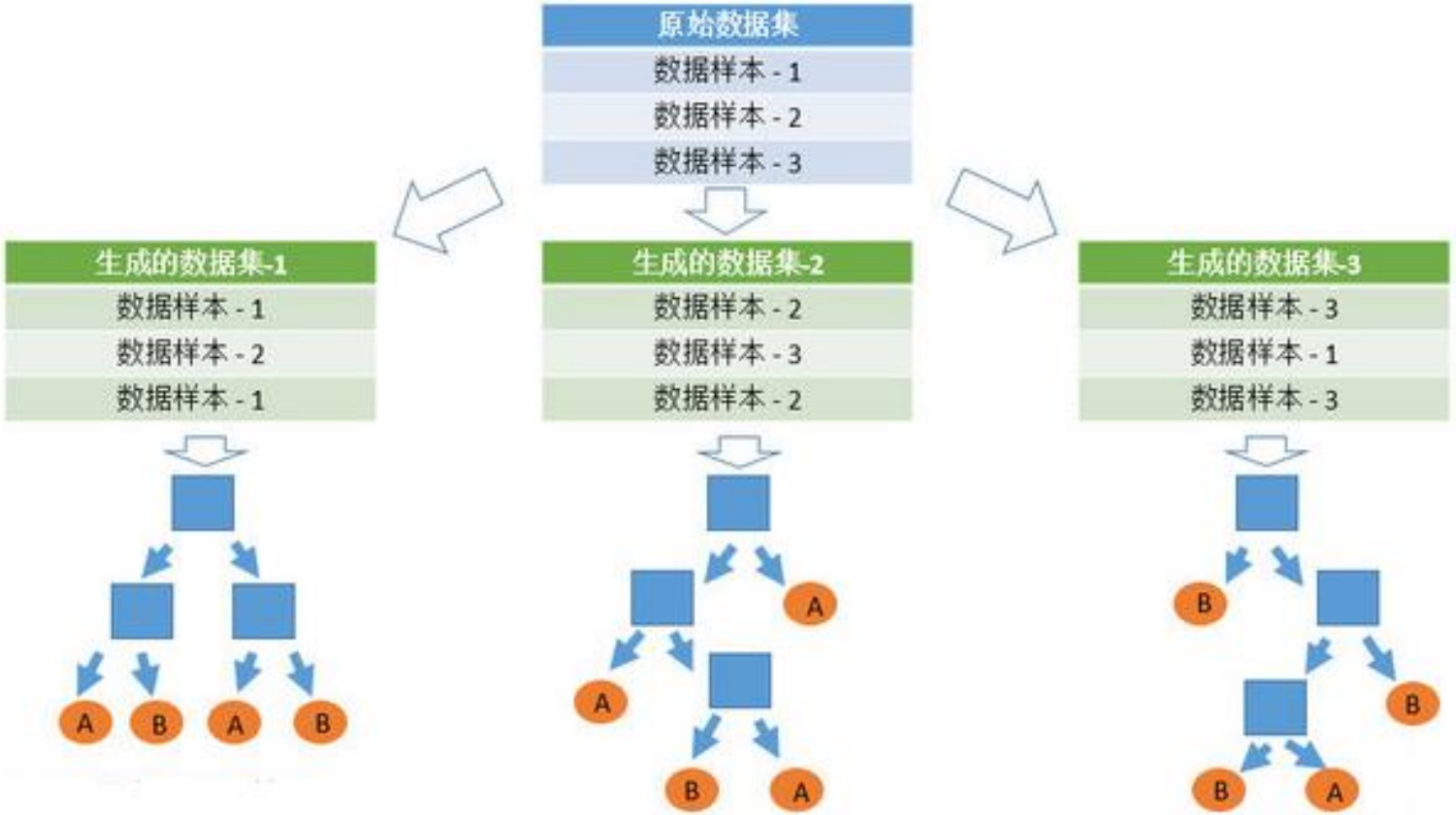
- 决策树相当于一个大师，通过自己在数据集中学到的知识对于新的数据进行分类。但大师是可遇不可求的。
- 俗话说得好，三个臭皮匠顶个诸葛亮。随机森林就是希望构建多个臭皮匠，希望最终的分类效果能够超过单个大师的一种算法。
- 那随机森林具体如何构建呢？有两个方面：
  - 数据的随机性选取
  - 待选特征的随机选取。

随机森林（Random forest）

- 数据的随机选取：
- 首先，从原始的数据集中采取有放回的抽样，构造子数据集，子数据集的数据量是和原始数据集相同的。不同子数据集的元素可以重复，同一个子数据集中的元素也可以重复。

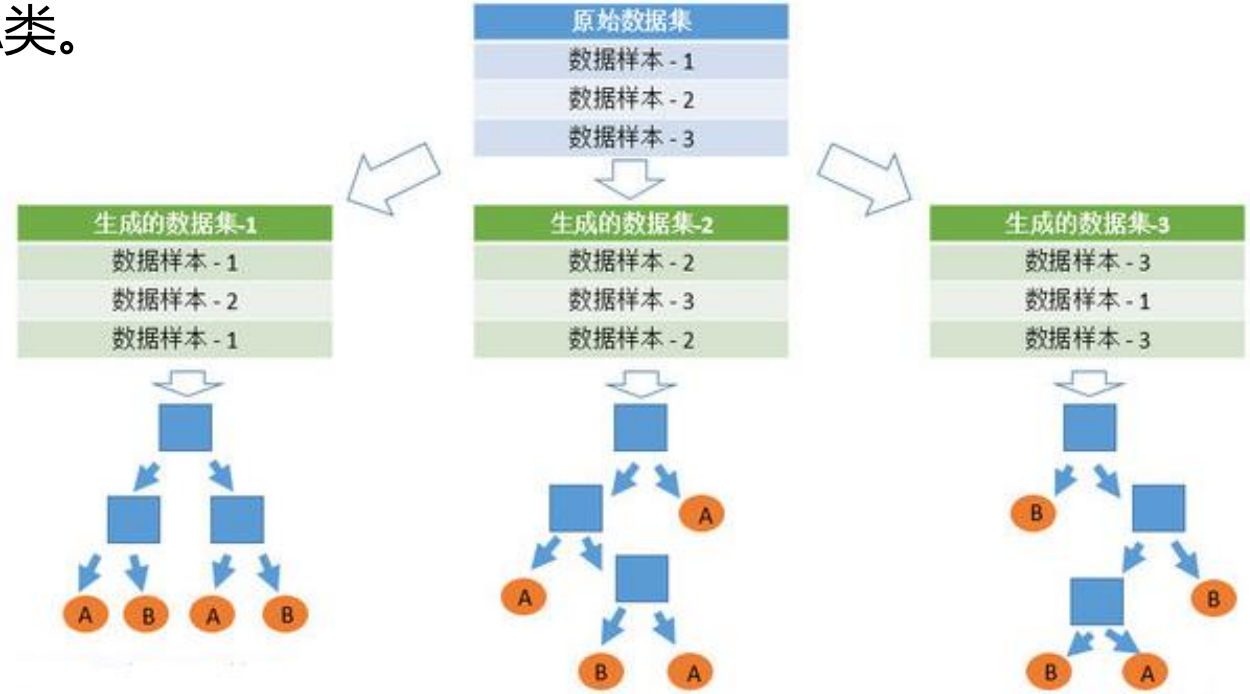
随机森林（Random forest）

- 第二，利用子数据集来构建子决策树，将这个数据放到每个子决策树中，每个子决策树输出一个结果。



随机森林（Random forest）

- 最后，如果有了新的数据需要通过随机森林得到分类结果，就可以通过对子决策树的判断结果的投票，得到随机森林的输出结果了。如下图，假设随机森林中有3棵子决策树，2棵子树的分类结果是A类，1棵子树的分类结果是B类，那么随机森林的分类结果就是A类。

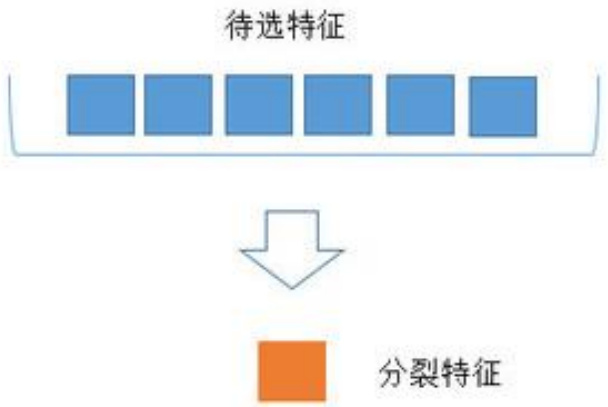


随机森林（Random forest）

- 待选特征的随机选取：
- 与数据集的随机选取类似，随机森林中的子树的每一个分裂过程并未用到所有的待选特征，而是从所有的待选特征中随机选取一定的特征，之后再在随机选取的特征中选取最优的特征。
- 这样能够使得随机森林中的决策树都能够彼此不同，提升系统的多样性，从而提升分类性能。

随机森林（Random forest）

- 下图中，蓝色的方块代表所有可以被选择的特征，也就是目前的待选特征。黄色的方块是分裂特征。左边是一棵决策树的特征选取过程，通过在待选特征中选取最优的分裂特征（如C4.5算法等），完成分裂。右边是一个随机森林中的子树的特征选取过程。



决策树选取分裂特征过程

随机森林（Random forest）

- 随机森林有很多的优点：
  - 在数据集上表现良好
  - 在当前的很多数据集上，相对其他算法有着很大的优势
  - 能够处理很高维度（feature很多）的数据，并且不用做特征选择
  - 在训练完后，它能够给出哪些feature比较重要
  - 在创建随机森林的时候，对generlization error使用的是无偏估计
  - 训练速度快
  - 在训练过程中，能够检测到feature间的互相影响
  - 容易做成并行化方法
  - 实现比较简单
  - 不会产生过拟合问题

随机森林（Random forest）

- 美国金融银行业的大数据算法：随机森林模型+综合模型

- 模型组合（比如说有Boosting，Bagging等）与决策树相关的算法比较多，这些算法最终的结果是生成N(可能会有几百棵以上）棵树，这样可以大大的减少单决策树带来的毛病，有点类似于三个臭皮匠顶个诸葛亮的做法，虽然这几百棵决策树中的每一棵都很简单（相对于C4.5这种单决策树来说），但是他们组合起来确是很强大。

随机森林预测股市

- 如果我有昨天或者历史该股票的很多信息，哪些信息能决定今日的股价变动呢?
- 特征的选取
  - 技术指标
    - 这种因子每天都跟股价与成交量息息相关,对第二天的股价应该有很强的预测。常用的一些技术指标类的因子是：KDJ，BollDown（布林线）,MassIndex（梅斯线）,macd等等。
  - 基本面因子
    - 基本面因子主要反映了公司的基本面信息，反映了公司财务的状况，适合价值投资者。但是当基本面发生改变时，股价抖动也会非常大，主要选取如下因子：ETOP,PE

随机森林预测股市

- 特征的选取
  - 分析师因子
    - 分析师因子应该是分析师的预测，也许很有指导意义，否则作为反向指标也行啊，先加入2个试试：DAREC,DAREV
  - 大盘相关的数据，指数与融资融券
    - 一些大盘相关的数据，包括个股资金流入流出应该都会对股价造成很大影响。包括两融余额，个股资金净流入，上证指数的涨跌幅

随机森林预测股市

- 创建并且训练一个随机森林模型（包含1000个决策树），根据上一个交易日的因子预测涨跌，返回预测涨幅最大的前20支股票。
- 回测起始时间： 2013-01-01
- 回测结束时间：2016-06-01
- 基准：沪深300指数
- 股票池：沪深300股票
- 每20日调仓

随机森林预测股市

- 创建并且训练一个随机森林模型（包含1000个决策树），根据上一个交易日的因子预测涨跌，返回预测涨幅最大的前20支股票。

年化收益率	基准年化收益率	阿尔法	贝塔	夏普比率	收益波动率	信息比率	最大回撤	换手率
30.9%	7.0%	24.0%	0.94	0.87	31.1%	1.33	51.1%	34.82