**【量化课堂】随机森林入门**

**导语：**把”机器学习“应用到量化投资领域，不同于以往的量化策略。机器学习算法是一类从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。其中，随机森林算法是一种基于统计学习理论的组合分类器。它可以将用户自选的各个因子，以机器训练的方式，自动分析其影响力度，从而给用户投资建议。本文属于入门级别，若有兴趣，后续文章将深入分析，敬请关注。  
  
*本文由JoinQuant量化课堂推出 。难度标签为进阶上，理解深度标签：level-0*  
  
**作者：**yongpeng.r  
**编辑：**宏观经济算命师

**一、相关概念**

**分类器：**分类器就是给定一个样本的数据，判定这个样本属于哪个类别的算法。例如在股票涨跌预测中，我们认为前一天的交易量和收盘价对于第二天的涨跌是有影响的，那么分类器就是通过样本的交易量和收盘价预测第二天的涨跌情况的算法。

**分裂：**在决策树的训练过程中，需要一次次的将训练数据集分裂成两个子数据集，这个过程就叫做分裂。

**特征：**在分类问题中，输入到分类器中的数据叫做特征。以上面的股票涨跌预测问题为例，特征就是前一天的交易量和收盘价。

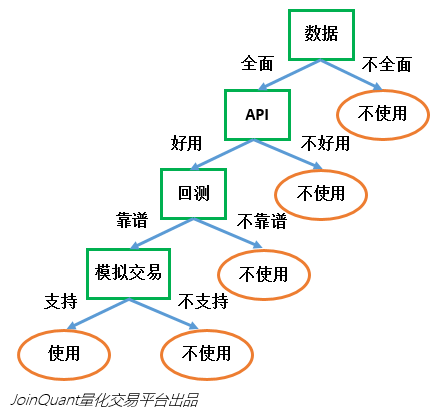
**待选特征：**在决策树的构建过程中，需要按照一定的次序从全部的特征中选取特征。待选特征就是在目前的步骤之前还没有被选择的特征的集合。例如，全部的特征是 ABCDE，第一步的时候，待选特征就是ABCDE，第一步选择了C，那么第二步的时候，待选特征就是ABDE。

**分裂特征：**接待选特征的定义，每一次选取的特征就是分裂特征，例如，在上面的例子中，第一步的分裂特征就是C。因为选出的这些特征将数据集分成了一个个不相交的部分，所以叫它们分裂特征。

**二、决策树的构建过程**

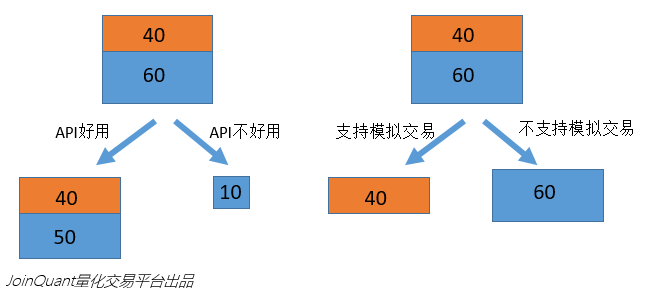
要说随机森林，必须先讲决策树。决策树是一种基本的分类器，一般是将特征分为两类（决策树也可以用来回归，不过本文中暂且不表）。构建好的决策树呈树形结构，可以认为是if-then规则的集合，主要优点是模型具有可读性，分类速度快。

我们用选择量化工具的过程形象的展示一下决策树的构建。假设现在要选择一个优秀的量化工具来帮助我们更好的炒股，怎么选呢？  
第一步：看看工具提供的数据是不是非常全面，数据不全面就不用。  
第二步：看看工具提供的API是不是好用，API不好用就不用。  
第三步：看看工具的回测过程是不是靠谱，不靠谱的回测出来的策略也不敢用啊。  
第四步：看看工具支不支持模拟交易，光回测只是能让你判断策略在历史上有用没有，正式运行前起码需要一个模拟盘吧。

这样，通过将“数据是否全面”，“API是否易用”，“回测是否靠谱”，“是否支持模拟交易”将市场上的量化工具贴上两个标签，“使用”和“不使用”。  
上面就是一个决策树的构建，逻辑可以用下图表示：  


在上图中，绿颜色框中的“数据”“API”“回测”“模拟交易”就是这个决策树中的特征。如果特征的顺序不同，同样的数据集构建出的决策树也可能不同。特征的顺序分别是“数据”“API”“回测”“模拟交易”。如果我们选取特征的顺序分别是“数据”“模拟交易”“API”“回测”，那么构建的决策树就完全不同了。

可以看到，决策树的主要工作，就是选取特征对数据集进行划分，最后把数据贴上两类不同的标签。如何选取最好的特征呢？还用上面选择量化工具的例子：假设现在市场上有100个量化工具作为训练数据集，这些量化工具已经被贴上了“可用”和“不可用”的标签。

我们首先尝试通过“API是否易用”将数据集分为两类；发现有90个量化工具的API是好用的，10个量化工具的API是不好用的。而这90个量化工具中，被贴上“可以使用”标签的占了40个，“不可以使用”标签的占了50个，那么，通过“API是否易用”对于数据的分类效果并不是特别好。因为，给你一个新的量化工具，即使它的API是易用的，你还是不能很好贴上“使用”的标签。  


再假设，同样的100个量化工具，通过“是否支持模拟交易”可以将数据集分为两类，其中一类有40个量化工具数据，这40个量化工具都支持模拟交易，都最终被贴上了“使用”的标签，另一类有60个量化工具，都不支持模拟交易，也都最终被贴上了“不使用”的标签。如果一个新的量化工具支持模拟交易，你就能判断这个量化工具是可以使用。我们认为，通过“是否支持模拟交易”对于数据的分类效果就很好。

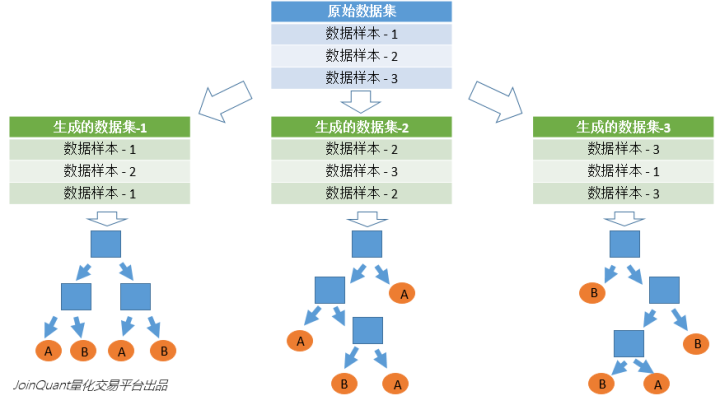
在现实应用中，数据集往往不能达到上述“是否支持模拟交易”的分类效果。所以我们用不同的准则衡量特征的贡献程度。主流准则的列举3个：ID3算法（J. Ross Quinlan于1986年提出），采用信息增益最大的特征；C4.5算法（J. Ross Quinlan于1993年提出）采用信息增益比选择特征；CART算法（Breiman等人于1984年提出）利用基尼指数最小化准则进行特征选择。

（如果想进行更深一步的学习，可以参考《统计学习方法》或者相关博文进行更一步的学习。未来的量化课堂也会涉及这方面的内容。）

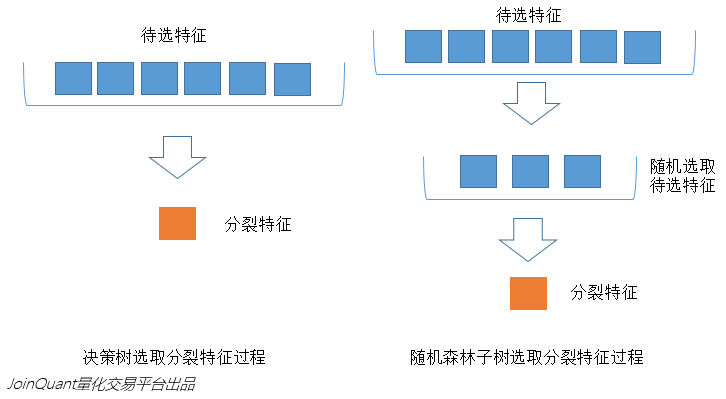
**三、随机森林的构建过程**

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。但是俗话说得好，一个诸葛亮，玩不过三个臭皮匠。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，以及待选特征的随机选取。

**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。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。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。  


**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。

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（别忘了前文提到的ID3算法，C4.5算法，CART算法等等），完成分裂。右边是一个随机森林中的子树的特征选取过程。  


**四、Random Forest的具体使用-sklearn**

以上介绍了随机森林的工作原理，那么在python环境下，我们可以利用python环境下的sklearn包来帮助我们完成任务。举个小例子：  
特征是通过收盘价数据计算的SMA，WMA，MOM指标，训练样本的特征是从2007-1-4到2016-6-2中截止前一天的SMA，WMA，MOM指标，训练样本的标类别是2007-1-4日到2016-6-2中每一天的涨跌情况，涨了就是True，跌了就是False，测试样本是2016-6-3日的三个指标以及涨跌情况。我们可以判定之后判断结果是正确还是错误，如果通过Random Forest判断的结果和当天的涨跌情况相符，则输出True，如果判断结果和当天的涨跌情况不符，则输出False。（和SVM那一篇中例子的作用是一样滴，只是为了展示如何使用，不对预测的准确性做担保啊）

具体代码见下面的研究代码啦  
  
*本文由JoinQuant量化课堂推出，版权归JoinQuant所有，商业转载请联系我们获得授权，非商业转载请注明出处。*