

## 【量化专题】：机器学习模型理论—— Boosting 和 Bagging

### 专题报告

#### 摘要

- 集成是合并多个机器学习模型来构建更强大模型的方法，通常可获得比单一学习器更优越的泛化性能。在机器学习文献中有许多模型属于这类。
- 集成学习方法大致可分为两类，一是个体学习器之间存在强依赖关系、必须串行生成的序列化方法，代表为 Boosting；二是个体学习器之间不存在强依赖关系、可同时生成的并行化方法，代表为 Bagging 和随机森林。本文在这两类集成学习方法中选择常用的 AdaBoosting 和随机森林两种算法进行介绍。

作者姓名：姜慧丽

邮箱：jianghuili@csc.com.cn  
电话：023-81157278  
期货从业资格号：F3081375  
期货投资咨询从业证书号：Z0018496

发布日期：2024 年 3 月 28 日

风险提示：本报告仅对模型作客观呈现，不具备任何投资建议。历史业绩不代表未来业绩，回测业绩不代表实盘业绩，期市有风险，入市需谨慎。

目录

<b>【量化专题】：机器学习模型理论—AdaBoosting 和随机森林 .....</b>	<b>1</b>
摘要 .....	1
一、集成学习概述 .....	1
二、Boosting 与 AdaBoosting.....	1
三、Bagging 与随机森林 .....	2
四、总结 .....	4

## 一、集成学习概述

集成学习，即分类器集成，是指通过构建并结合多个学习器来完成学习任务。一般结构是：先产生一组“个体学习器”，再用某种策略将它们结合起来。结合策略主要有平均法、投票法和学习法等。

## 二、Boosting 与 AdaBoost

### 2.1 Boosting 算法简介

在序列化的方法中，组合起来的不同弱模型之间不再相互独立地拟合。其思想是迭代地拟合模型，每一个模型都在尝试增强整体的效果。具体而言，先从初始训练集训练出一个基学习器，再根据基学习器的表现对训练样本分布进行调整，使得先前学习器处理的较差的训练样本在后续受到更多的关注。然后基于调整后的样本分布来训练下一个学习器。如此重复进行，直至学习器数目达到预先指定的值，使模型在给定步骤上的训练依赖于之前拟合的模型。

提升法(Boosting)是最著名的序列化方法的一种。Boosting 着眼于以一种适应性很强的方式顺序拟合多个弱学习器：序列中每个模型在拟合的过程中，会更加重视之前的模型处理的很糟糕的观测数据，每个模型都把注意力集中在目前最难拟合的观测数据上。这样一来，在这个过程的最后，我们就获得了一个强学习器。

下边将介绍一个重要的 Boosting 算法：自适应提升(adaboost)算法。

### 2.1 AdaBoost 算法简介

在自适应 adaboost 中，我们将集成模型定义为  $L$  个弱学习器的加权和：

$$S_L(\cdot) = \sum_{l=1}^L c_l \times w_l(\cdot)$$

其中  $c_l$  为系数， $w_l$  为弱学习器。寻找最佳集成模型是一个困难的优化问题，不可能一次性找到给出最佳整体加法模型的所有系数和弱学习器，因此使用一种更易于处理的迭代优化过程。

即将弱学习器逐个添加到当前的集成模型中，在每次迭代中寻找可能的最佳组合（系数、弱学习器），

循环地将  $s_l$  定义如下：

$$s_l(\cdot) = s_{l-1} + c_l \times w_l(\cdot)$$

其中， $c_l$  和  $w_l$  被挑选出来，使得  $s_l$  是最适合训练数据的模型，因此这是对  $s_{l-1}$  的最佳可能改进。

在考虑二分类问题时，可以将 adaboost 算法重新写入以下过程：首先，更新数据集中观测数据的权重，训练一个新的弱学习器，该学习器重点关注当前集成模型误分类的观测数据。其次，它会根据一个表示该弱模型性能的更新系数，将弱学习器添加到加权和中：弱学习器的性能越好，对强学习器的贡献就越大。

因此，假设面对一个二分类问题：数据集中有  $N$  个观测数据，我们想在给定一组弱模型的情况下使用 adaboost 算法。在算法的起始阶段，所有的观测数据都拥有相同的权重  $1/N$ 。然后，将下面的步骤重复  $L$  次（作用于序列中的  $L$  个学习器）：

- 1、用当前观测数据的权重拟合可能的最佳弱模型；
- 2、计算更新系数的值，更新系数是弱学习器的某种标量化评估指标，它表示相对集成模型来说，该弱学习器的分量如何；
- 3、通过添加新的弱学习器与其更新系数的乘积来更新强学习器计算新观测数据的权重，该权重表示我们想在下一轮迭代中关注哪些观测数据。

重复这些步骤，顺序地构建出  $L$  个模型，并将它们聚合成一个简单的线性组合，然后由表示每个学习器性能的系数加权。

### 三、 Bagging 与随机森林

#### 3.1 Bagging 算法简介

在并行化的方法中，单独拟合不同的学习器，因此可以同时训练它们。最著名的方法是自助聚合 (Bagging)，它的目标是生成比单个模型更好的集成模型。

自助法：这种统计技术先随机抽取出作为替代的  $A$  个观测值，然后根据规模为  $N$  的初始数据集生成大小为  $A$  的样本，称为自助样本。

在某些假设条件下，这些样本具有非常好的统计特性：在一级近似中，它们可被视为直接从真实的底层数据分布中抽取出来的，并且彼此之间相互独立。因此，它们被认为是真实数据分布的代表性和独

立样本。为了使这种近似成立，必须验证两个方面的假设：

初始数据集的大小  $N$  应该足够大，以捕获底层分布的大部分复杂性。这样，从数据集中抽样就是从真实分布中抽样的良好近似。

与自助样本的大小  $A$  相比，数据集的规模  $N$  应该足够大，这样样本之间就不会有太大的相关性。

举例而言，自助样本通常用于评估统计估计量的方差或置信区间。根据定义，统计估计量是某些观测值的函数。因此，随机变量的方差是根据这些观测值计算得到的。为了评估这种估计量的方差，需要对从感兴趣分布中抽取出来的几个独立样本进行估计。在大多数情况下，相较于实际可用的数据量来说，考虑真正独立的样本所需要的数据量可能太大。然而，可以使用自助法生成一些自助样本，它们可被视为独立同分布的样本。这些自助样本可以通过估计每个样本的值，近似得到估计量的方差。

Bagging 的工作机制为：

1、从原始样本集中抽取出  $K$  个训练集。每轮从原始样本集中使用自助法，抽取  $n$  个训练样本，随机森林中，还会随机抽取一定数量的特征。

2、 $K$  个训练集分别训练，共得到  $k$  个模型体现了并行化的方法。

## 3.2 随机森林

随机森林是具有代表性的 Bagging 集成学习算法，它的所有基评估器都是决策树，分类树组成的森林就称为“随机森林分类器”，回归树所集成的森林就称为“随机森林回归器”。

### 3.2.1 随机森林算法步骤

随机森林采用决策树作为分类器，在有放回的随机采样基础上，进一步增加了特征的随机选择。其算法步骤如下：

- 1、从样本数据集中有放回的随机选取  $n$  个样本；
- 2、从所有特征中随机选择  $k$  个特征，基于有放回的随机选取的样本以及随机选取的特征建立决策树；
- 3、重复第一步、第二步，生成  $m$  棵决策树，且每棵决策树都最大可能地进行生长而不进行剪枝；
- 4、通过对所有的决策树加总进行预测，其中分类时采用多数投票策略、回归时采用平均策略。

### 3.2.2 袋外错误率

随机森林的一个重要优点是不需要对它进行交叉验证或者用一个独立的测试来获得误差的无偏估计。它可以在内部进行评估，即生成决策树的过程中就可以建立一个无偏估计。

在构建决策树时，模型通过有放回的随机抽取样本。对第  $i$  棵树而言，约  $1/3$  的训练样本没有参与该决策树的生成，它们称为第  $i$  棵树的 oob 样本，可以进一步进行 oob 估计，计算步骤如下：

- 1、对每个样本，计算它作为 oob 样本的树对它的分类情况；
- 2、然后以多数投票作为该样本的分类结果；
- 3、然后用误分个数占样本总数的比率作为随机森林的 oob 误分率。

### 3.2.3 特征选择

随机森林进行特征重要性评估思想是计算每个特征在随机森林不同决策树上做了多大的贡献，然后取其平均值，并比较不同特征之间的贡献大小。贡献度的衡量指标包括 Gini 系数、袋外错误率。

## 四、总结

本文主要介绍了集成学习以及具体的算法原理。通俗地说，集成学习就是利用群众的智慧去学习同样的数据集，不断地迭代以达到比单个模型更好的效果，因此集成学习一般都有很高的准确性。但需要注意的是上述集成学习的方法还是有各自的局限性，比如会存在过度拟合，分类器数目的设定，对离群点敏感等。

## 联系我们

全国统一客服电话：400-8877-780

网址：[www.cfc108.com](http://www.cfc108.com)

获取更多研报报告、专业客户经理一对一服务、  
了解公司更多信息，扫描右方二维码即可获得！



## 重要声明

本报告内容仅供符合《证券期货投资者适当性管理办法》规定可参与期货交易的投资者参考。在任何情形下都不构成对接收本报告内容投资者的任何投资建议，投资者应充分了解各类投资风险并谨慎考虑本报告发布内容是否符合自身特定状况，自主做出投资决策并自行承担投资风险。中信建投期货不因任何订阅或接收本报告的行为而将订阅人视为中信建投的客户，投资者依据本报告内容作出的任何决策与中信建投期货或作者无关。

本报告发布内容如属于系列解读，则投资者可能会因缺乏对完整内容的了解而对其中假设依据、研究依据、结论等内容产生误解，提请投资者参阅我司已发布的完整系列报告，仔细阅读其所附各项声明、数据来源及风险。

中信建投期货对本报告所载资料的准确性、可靠性、时效性及完整性不作任何明示或暗示的保证，本报告意见仅代表报告发布之时的判断，相关研究观点可能依据我司后续发布的报告在不发布通知的情形下作出更改。

本报告发布内容为中信建投期货所有。未经我司书面许可，任何机构和个人不得以任何形式对本报告进行翻版、复制和刊发，如需引用、转发等，需注明出处为“中信建投期货”，且不得对本报告进行任何增删或修改。亦不得从未经我司书面授权的任何机构、个人或其运营的媒体平台接收、翻版、复制或引用本报告发布的全部或部分内容。版权所有，违者必究。

**全国统一客服电话：400-8877-780**

**网址：[www.cfc108.com](http://www.cfc108.com)**