1. **集成学习**

我们可以通过方差和偏差这两个参数来判断一个模型的好坏。偏差指模型的预测值与真实值的差距，方差指模型的预测值波动的范围。若一个模型的偏差很大，代表这个模型没有很好地捕捉到数据特征，没有很好地拟合数据，即欠拟合；若一个模型的方差很大，代表这个模型在不同训练集中表现差异较大，可能在某训练集中捕捉到了一些数据的特殊特征，即过拟合。当方差和偏差都比较小时，我们可以认为这个模型是比较强大的。

在机器学习中的有监督学习的算法中，我们期望得到一个稳定的，准确的，能适应所有情况的可靠模型，然而实际上出于种种原因，我们很难得到这样的模型。但是相比于得到一个强大准确的模型，获得若干个相比较而言并不是那么完美的，有偏好的（即只在某些情况下表现较佳）的模型要现实得多。而集成学习就是将这些弱模型通过某种方式组合起来，得到一个较为完美的强模型，它的思路是即便某些模型得出错误的预测，其它的模型也能纠正它们的错误。比较典型的算法有引导聚集发和增强学习法等。

1. 引导聚集算法（Bagging）

“Bagging”一词是对“bootstrap aggregating”的简写，其中“bootstrap”又称自展法，是一种用小样本估计整体值的非参数方法。具体操作是从初始数据中有放回地抽取一定量的样本，通过对这些样本的计算得到统计量的置信区间。Bagging使用bootstrap方法从整体数据集中有放回地取出若干个数据集，针对每一个数据集训练模型，最后参考这些模型的预测结果得出答案。更具体地，对于分类问题，可以采用所有模型投票的方式得出答案；对于回归问题，对所有模型的结果求平均值。

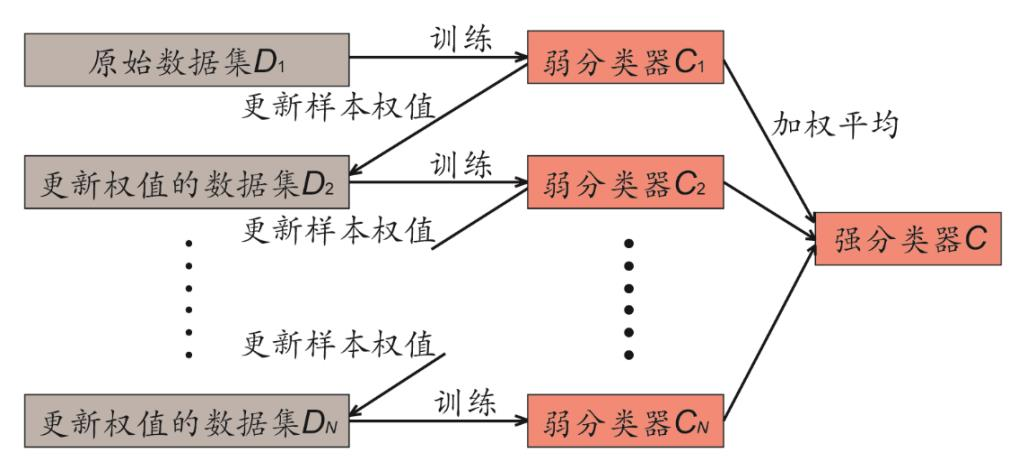
Bagging最大的优势是可以通过求平均的方式缩小基分类器的方差，从而减小过拟合。它的缺点是选取样本集的方法会带来一定量的偏差，影响最终结果。

上文中提及的随机森林算法就是用到了Bagging的思想。被应用在许多股市预测的模型当中。

1. 增强学习算法（Boosting）

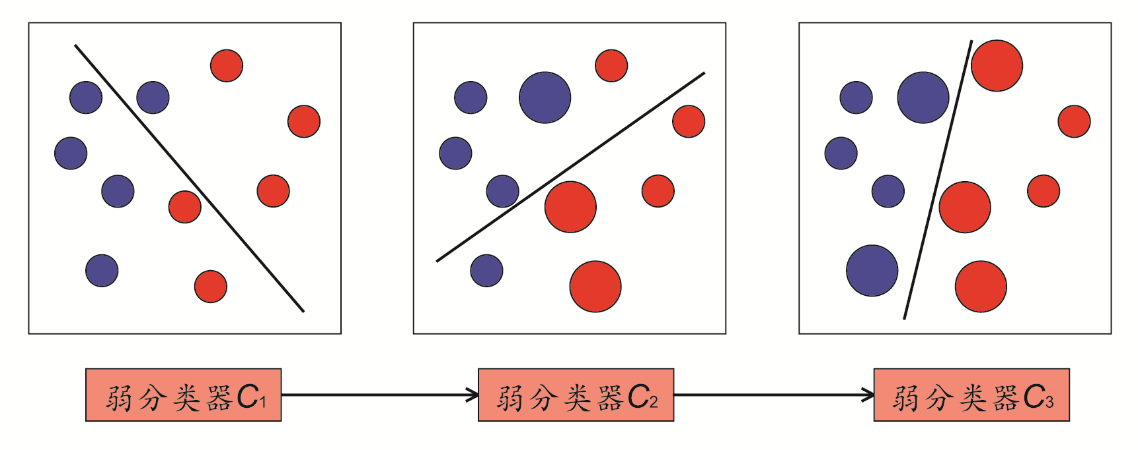
Boosting方法采用重赋权的方式迭代地训练基分类器。具体的操作是从训练集使用初始权重训练出第一个模型，再如图5.2.2中所显示，根据当前模型的错误数据更改训练集中样例的权值，训练第二个模型，以此类推，直到生成了n个弱学习器，再根据预测误差率将它们加权平均组合成一个强学习器，如图5.2.1中所示。相比于Bagging算法，Boosting算法能有效地降低模型的偏差。

**图5.2.1 boosting算法的实现流程**



资料来源：华泰证券交易所

**图5.2.2 训练集权重调整过程**



资料来源：华泰证券交易所

根据如何调整训练集的权重，以及如何组合基分类器，Boosting方法有很多不同的方式去实现，比较著名的有AdaBoost，GradienBoosting等算法。其中Adaboost是其中的成功代表，被称为数据挖掘十大算法之一[1]。

AdaBoost二元分类法的具体操作可分为以下几步：

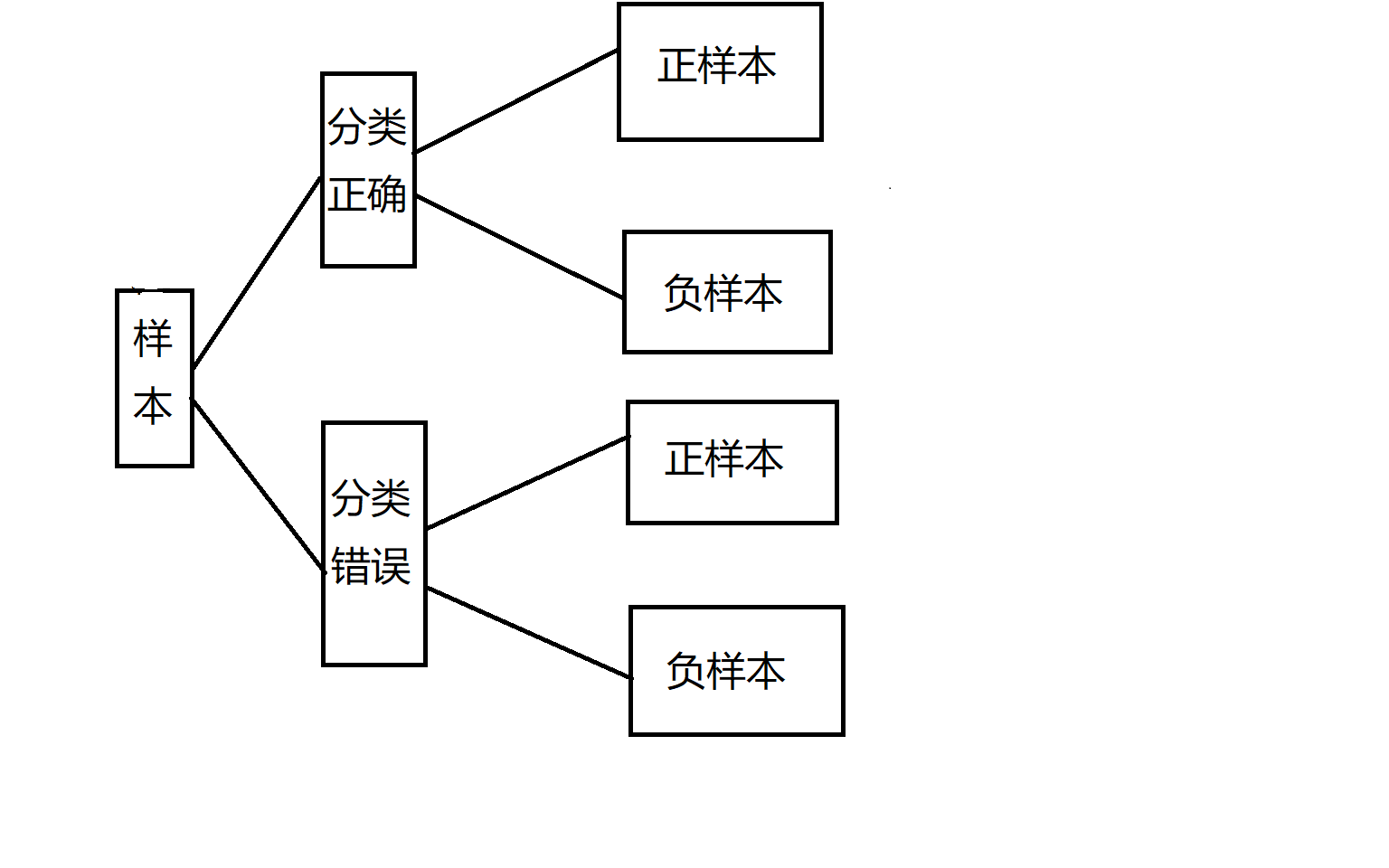
1. 我们令总数据集为T = {(x1 , y1) , (x2 , y2) ... (xn , yn)}，标记输出为{-1 , 1}，迭代次数为K
2. 初始化样本集的权值为W(1) = {w11 , w12 ... w1n}，其中w1i = 1/n，i = 1 , 2 ...n。
3. 执行多次迭代操作，令k = 1 , 2 ... K表示当前迭代数。

使用权重为D(k)的样本集对分类器进行训练，得到基本分类器Gk，并计算Gk(x)在该样本集上的分类错误率e(k) = P(Gk(xi) != yi) = 。再根据这个分类错误率得到这个分类器在最终分类器中的重要程度，其中可以看出，小于二分之一时为正，且随的减小而增大，符合“越强的基分类器在最终分类器中作用越大”的思路。最后，我们还要根据当前分类器的错误情况更新训练集的权值分布W(k + 1)。。其中是归一化因子，使变为对应的概率化分布。通过这一函数，使得原本被分类错误的样本在下一次训练中具有更高的权重，相对的，正确的样本权重减小，使得AdaBoost方法能够“集中注意”在比较难以分类的样本上。

1. 最后按加权平均组合所有的基分类器

综上，我们可以看出AdaBoost在二元分类的问题上颇有建树，所以它常常可以和其他算法如决策树以及SVM等一起使用。而在龚利琴在2018年的研究中，认为简单地区分正确和错误在金融领域是不够的。在金融界中，人们对损失往往比对相同数额的收入更为敏感，所以对于分类正确和分类错误的样本继续进行一个正负样本的区分，如图5.2.3所示。进而使用付中良提出的多分类敏感问题代价敏感的AdaBoost算法[2]，实现Alpha策略选股[3]。持此之外，华泰证券[4]也对Boosting系列算法在金融预测方面的能力进行了评估，认为在同等预测水平下，Boosting的复杂程度要远低于Bagging，表现较为优异。

**图5.2.3 分类方法**



1. **总结与展望**

本文简要阐述了包括决策树，支持向量机，神经网络，集成学习等一系列经典人工智能算法的原理，并适当地介绍了一些基于这些算法进行改进的特殊模型，并解释了他们在股市预测领域的应用。在完成本片文献的过程中，本组组员们分工合作，每人对应其中一个研究方向进行较为有深度的学习,查找了大量文献,可谓是收获颇丰。除了对算法本身有了更深刻的理解之外，我们对于金融股市的预测也有了进一步的了解：金融大数据的特点是瞬息万变，不确定性高，而机器学习恰恰能够适应这样的数据，相比于使用传统的回归模型，机器学习不但能提高预测精度，还可以将更多文本数据，新闻事件等信息加入数据集中，大大拓展了数据类型。

然而我们发现现在人工智能在股市预测领域的应用尚不成熟，处于探索阶段。存在着对公式的过度依赖问题，不能够很灵活地利用模型，导致得到的预测结果泛化能力不足的问题。未来需要进一步提高模型预测的精度，挖掘更多不同类型的数据，另外，对于机器学习当中参数设定和预测结果的经济学理论解释也是未来的研究难题。

**参考文献**

1. Zhou Z H, Yang Y, Wu X D, Kumar V, The Top Ten Algo-rithms in Date Mining New York, USA: CRC Press, 2009, 127149
2. 付中良. 多分类问题代价敏感AdaBoost算法[J]. 自动化学报, 37(8):973-983, 2011
3. 龚利琴. 基于AdaBoost算法的Alpha组合研究[D]. 郑州大学,2018.
4. 林晓明. 金工:人工智能选股之Boosting模型[R]. 华泰证券研究报告,2017
5. 张世杰.大数据下的机器学习在股市预测中的应用[J].贵阳学院学报(社会科学版),2021,16(04):43-48.DOI:10.16856/j.cnki.52-1141/c.2021.04.007.