# 第十七节 基于筹码分布的股指期货交易策略

1. **筹码分布理论简介**

筹码，在中国内地股市，是投资者耳熟能详的一个技术概念。在西方股票市场中，筹码(chip)一般被用于形容业务规模大、市值高和对经济周期不敏感的企业的优质股票(蓝筹股)。而在中国，筹码带上了更多的博弈的意味，筹码的意义被更进一步地扩展，每一张股票即是一颗筹码。股票的交易，因此等价于钱和筹码的置换。将每张股票看作一颗筹码，放置在持有者的购入价位上，这样形成的价格与筹码密度的关系，就是**筹码分布**。当交易发生时，筹码（股票）从卖出者的购入价位上取出，放置在买入者的购入价位上。筹码的位置代表了持有者的持仓成本，筹码的移动体现了行情的发展。

筹码分析理论，就是根据筹码流动的特点，对大盘或者个股的历史成交情况进行分析，得出其**筹码结构**（筹码分布），然后根据这个筹码结构来预测以后的走势。筹码分析主要是研究卖盘，而不是买盘。筹码流动的动力，是心理学的内容，是盈利和亏损的筹码和价格的相互作用。

筹码分布是一个较少见的发源于中国投资者，而且被局限在中国股票市场的概念。 一方面，筹码分布的概念完全脱离企业的基本面，独立地定义在金融市场内发生的交易之上；另一方面，它深深根植在许多投资者的方法论中，因此可能的后果是，即使筹码理论是空中楼阁，只要足够多的投资者相信而且以其为依据行动，那么它在行为金融的层面上，就会和市场的走势有着紧密的联系。

1. **股指的筹码分布算法**

如同某只个股，筹码分布同样可以应用于指数上，以下我们不区分地使用“点数”和“价格”。给定某个股票指数的自挂牌以来的价格和换手率，我们可以使用如下算法得到任意时间点的筹码分布：

假设在某日，在任意价位上的筹码所占比例为，第日的换手率为，收盘价格为。则第日的筹码分布为

简单地说，我们假设筹码在各个价位等比例地流出，然后汇入当日的收盘价。这里有几个比较强的假设。第一，想要获得无限精确的筹码分布（即市场中每一张股票的成本），就必须获得每一笔交易的成交价和成交量，这显然是不可能的，因为我们使用当日的收盘价来近似所有交易的成交价。第二，流通盘的移动未必是均匀的，而为了简化问题我们假定了各个价位的筹码以同样的换手率流出。第三，一般筹码算法会使用成交金额/成交量来估算平均成交价格，但是指数并没有可用的成交量数据，所以使用收盘价来。

图2.17.1是沪深300指数在2017年8月18日的筹码分布，为了提高可视性，将每10点指数内地所有筹码加和显示。该日处于一个较长的、稳定的、回撤很少上升阶段。筹码分布是一个典型的类似正态分布的形态，超额峰度(excess kurtosis)达到了8，意味着有大量的筹码非常密集地集中在峰顶价位。获利盘高达85%。

|  |
| --- |
| *图2.17.1：沪深300指数2017年8月18日的筹码分布* |
|  |
| *资料来源：WIND数据库，诺德基金FOF管理部* |

图2.17.2是沪深300指数筹码在2016年6月22日的形态，该日处于一个短期而且伴随着回撤的小幅度攀升中。筹码分散，超额峰度仅有0.81，获利盘仅有22%。图2.17.3是沪深300指数筹码在2018年4月17日，位于一个长时间的震荡下挫的走势中的筹码分布，可以发现筹码同样较为离散，更接近标准正态分布，超额峰度仅有0.34。

通过选取了三个具有代表性的走势内的筹码分布，我们可以发现筹码的形态和股指的涨跌趋势是具有一定的联系。最为显著的是，越尖的筹码峰（越高的峰度）越是强劲牛市的信号。基于筹码分布的股指预测模型和交易策略，正是建立在这样的假说上：在某个行情阶段内，每日的筹码形态具有一些共同性，因而可以被划分为一类。当我们找到这些可以有效的分类筹码形态的指标之后，就可以建立一个预测模型来预测未来的指数走势。

|  |
| --- |
| *图2.17.2：沪深300指数2016年6月22日的筹码分布* |
|  |
| *资料来源：WIND数据库，诺德基金FOF管理部* |

|  |
| --- |
| *图2.17.3：沪深300指数2018年4月17日的筹码分布* |
|  |
| *资料来源：WIND数据库，诺德基金FOF管理部* |

1. 基于筹码分布的股指预测模型

样本数据方面，我们从万德获取了沪深300，中证500和上证50三个指数自挂牌以来的日频率数据，利用上文的算法计算出每日的筹码分布。基于日频率的数据，模型的预测时长是未来的N个交易日。若无说明，以下我们都以沪深300指数为研究对象，中证500和上证50的应用表现将会在最后讨论。

首先我们猜测，这些指标可能会反应当下的持仓成本结构（即筹码结构），从而影响未来的走势：1.当前价格的相对位置2.平均持仓成本的相对位置3.峰度4.当前价格附近的筹码比例5.盈利盘比例。下一步，如何选择一个合适的模型类型呢？线性模型被尝试后很快被派出了，因为表现不佳，并且考虑到筹码分布其实是形态分析，形态分析的复杂度很高且相当主观，符合一个简单的线性模型的可能性比较小。最初我们想要预测未来一段时间的回报率，但这显然难度过大，而且对仅仅建立一个交易策略来说，有些多余，所以预测目标改为N个交易日后的涨跌。

最后，二叉决策树(Decision tree)的拟合度较好。决策树是一个非线性、简洁但是强大的机器学习模型。直观地说，决策树的原理是给定一个样本中的若干个特征的取值，将一个样本从树的根部输入，在每个节点按条件（如超额峰度>0）划分入两个子树中的一个，直到抵达叶节点。每个叶节点对应预测的标签（涨/跌），示例见图2.17.4。**随机森林**(Random Forecast)是包含多个决策树的集成学习模型（ensemble），优点是可以增强模型的稳定性（应对不同的测试集的表现更加一致），使得实盘应用的风险减低。我们最终选用sklearn的随机森林RandomForecastClassifier，按sklearn的默认设置配置参数。更多有关决策树和随机森林，可以参考sklearn的[官方文档](https://scikit-learn.org/stable/user_guide.html)。

对前面提到过的筹码分布算法得到的筹码分布序列进行预处理。设最高概率的价位的筹码概率（即落在此价位上的筹码比例）为，去除任何概率小于的价位上的筹码，对剩余的筹码进行放大使得概率之和仍为1，c的默认值为0.005，在后面会对此参数进行优化分析。因子的作用是去除极端价格（如发行价）上的微量筹码造成的不稳定性，同时凸显主力持仓。

|  |
| --- |
| *图2.17.4：训练完毕的决策树的一部分* |
|  |
| *资料来源：诺德基金FOF管理部* |

从五个初始特种出发，尝试不同的组合后，精选出三个最有效的特征作为模型输入（自变量）：

对于一个筹码分布：价格，对应的概率.

1. **当前价格的相对位置**

在第j日，收盘价为,回溯紧邻的之前的M个交易日。在长度为M的时间窗口内，设最高收盘价为,最低收盘价为,当前价格的相对位置

1. **平均持仓成本的相对位置**

在第j日，平均（期望）持仓成本 . 平均持仓成本的相对位置

1. **超额峰度**.

参数M的默认值为60，也就是说用前60个交易日的行情作为一个基准区间-从最低价到最高价-来衡量当前价格和当前平均持仓成本是处于一个偏高还是偏低的位置。超额峰度一定程度可以探测主力大量囤货的市场异动，以及量化整个持仓成本结构的分散度。当筹码散落在不同价位的时候，市场可能会更加缺乏动能，因为小的价格波动不能造成一大批持有者的盈亏突然转换，未来比较可能处于一个震荡的态势中。因此，峰度是一个重要的特征，十分有利于决策树的树杈将不同筹码形态下代表的不同市场潜能区分开来，再利用相对价格特征辅助判断涨跌。

我们的预测目标是N个交易日后的涨跌标签，N的数值代表了我们的预测视野。实验证明，筹码特征的预测能力随着预测视野边长而增强。沪深300指数而言，短期模型的预测能力不佳，中期N=60时预测能力最优，也就是说筹码分布信号有三个月左右的延迟。因此我们决定预测60交易日的中期指数变化。滚动计算沪深300指数N日周期的涨跌标签(label)，假设第一日的价格为1000，第N日的价格为1030，则第一日分类为涨（）；第二日的价格为1010，第N+1日的价格为1005，则第二日分类为跌(），依此类推。以2016年6月13日为界，之前共2657个有效样本作为训练集，之后共672个有效样本作为测试集。测试结果展示在表2.17.1中。

可以看到，模型的拟合度还是相当不错的。2016年6月13日以后的672个交易日，模型正确的预测了75%的交易日60个交易日后的相对价格涨跌。图2.17.5将预测信号叠加在了指数上，可以看到，模型很擅长发掘长时间的、稳定的上升趋势，在震荡和下行区间的表现相对较差。

|  |  |
| --- | --- |
| *表2.17.1：测试结果* | |
| |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | **测试样本** | **上涨次数** | **下跌次数** | **预测准确率** | **看多次数** | **看多胜率** | **看空次数** | **看空胜率** | | 672 | 425 | 257 | 0.75 | 507 | 0.78 | 165 | 0.75 | |  |
| *资料来源：诺德基金FOF管理部* | |

|  |
| --- |
| *图2.17.5：指数与信号* |
|  |
| *资料来源：WIND数据库，诺德基金FOF管理部* |

图2.17.6显示出，筹码分布的三个特征的确呈现出集群的特点，临近的点极有可能分享同样的标签，这是随机森林模型的力量源泉，每棵决策树都会寻找最优的空间划分。有两个明显的现象：第一，上层的高峰度空间几乎全部被上涨点占据，印证了尖峰筹码是很强的牛市信号。第二，底层的低峰度空间，处于高位的平均持仓成本是可能的上涨信号。当前价格的相对位置和另外两个特征的相互作用比较微妙，不在此展开。

|  |
| --- |
| *图2.17.6：筹码分布特征的三维展示* |
|  |
|  |
| *资料来源：WIND数据库，诺德基金FOF管理部* |

1. 策略构建

选择2016年6月13日-2019年6月13日作为回测区间，2016年6月13日之前的数据用作训练。

构建基于上述预测模型的交易策略，具体地，策略如下设置

* 将初始资金等分为50份，以每份的金额为恒定的买入/卖出金额，如资金不足则当日不开仓
* 假设可以做空，杠杆为1
* 每日收盘，根据模型的当日预测结果，以收盘价做多/做空，60天后以收盘价平仓
* 在回测期结束前清算所有开仓未结头寸，以最终现金总额计算回报
* 交易费率根据中金所2019年4月的最新公告，非平今仓的单边交易费率为股指的万分之零点二三，我们使用万分之零点二三作为每次交易的手续费率
* 引入参数：概率预测阈值（）。对于每一个待预测的样本，随机森林模型除了提供一个二分法的分类之外，也可以提供一个平均多个树的概率预测，分配给两个标签涨/跌各一个概率预测（和为1），可以理解成信心。设涨（标签1）的预测概率为u，则跌（标签0）的概率为(1-u)，当u=0.5时，实际是一个中性预测，我们掌握的对未来走势的信息最少。所以我们过滤掉最不确定的，也就是的预测,这时我们不开仓。

|  |  |
| --- | --- |
| *表2.17.2：概率阈值对策略表现的影响* | |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **概率阈值** | **夏普比率** | **头寸数量** | **最大回撤（%）** | **累计净值** | **超额收益率（%）** | | 0.3 | 0.5249 | 258 | 5.5790 | 1.2100 | 0.1832 | | 0.4 | 1.1361 | 382 | 6.0560 | 1.4177 | 5.9681 | | 0.5 | 1.1361 | 382 | 6.0560 | 1.4177 | 5.9681 | | 0.6 | 1.1361 | 382 | 6.0560 | 1.4177 | 5.9681 | | 0.7 | 1.2946 | 452 | 6.0782 | 1.4929 | 7.9227 | | 0.8 | 1.2946 | 452 | 6.0782 | 1.4929 | 7.9227 | | 0.9 | 1.3096 | 521 | 6.5081 | 1.5244 | 8.7217 | | 1.0 | 1.4826 | 636 | 5.5455 | 1.6188 | 11.0527 | |  |
| *资料来源：诺德基金FOF管理部* | |
| *图2.17.7：概率阈值对策略表现的影响* | |
|  | |
| *资料来源：诺德基金FOF管理部* | |

概率阈值与其说是一个参数，不如说是一个交给策略使用者定义的风格系数。从表2.17.2和图2.17.7中可以看到，随着阈值的不断提高，发起的头寸（多+空）更多，策略更加激进，信心不足的投机也会进去，结果是，超额年化收益率更多，最大回撤更大。虽然时最大回撤突然减少，但不建议将阈值设置为1（无过滤）,因为50%多50%空的输出预测是跟标签设置的顺序有关的，如果标签的顺序倒置，投机的盈亏会翻转。建议，之后默认使用此设置。

下面是策略的实证结果。

1. 参数敏感性测试
2. 总结
3. 研究不足与未来方向

选择60天的持有周期意味着投资者暴露在短期内的震荡中，可能会因为短期内的波动（而不是长期内的趋势）蒙受资产净值上的损失。例如模型