# 第十七节 基于筹码分布的股指期货交易策略

**前言**

在中国内地股市，筹码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)这个词一般指业务规模大、市值高和对经济周期不敏感的优质企业的股票(蓝筹股blue-chip)。但在中国，筹码的意义进一步扩展，筹码带上了更多博弈的意味：每一张股票即是一个筹码，每一次股票的交易即是钱和筹码的置换。这个诠释方式源于中国，而且只流行于中国。将每张股票看作一颗筹码，放置在持有者的购入价位上，当交易发生时，筹码（股票）从卖出者的成本价上取出，放到买入者的购入价位上。因此，筹码的位置代表了持有者的持仓成本，筹码的流动也就体现了行情的发展。

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，得出其筹码分布，也就是各个价位上有多少筹码的数据，然后根据这个筹码分布来预测以后的走势。筹码分析主要是研究卖盘而不是买盘，因为筹码的位置与卖方的利益有关，能暗示卖方的心理。筹码流动的动力是盈利和亏损的筹码和价格的相互作用,这是心理学的内容。一方面，筹码分布的概念完全脱离于企业的基本面，独立地定义于金融市场内发生的交易之中；另一方面，它深深根植在许多投资者的方法论中。如果有足够多的投资者相信筹码分析理论并且以之为依据行动，那么它在行为金融学层面上就会和市场的走势有紧密的联系。

**一、股指的筹码分布算法**

如同应用于个股，筹码分布同样可以应用于指数。以下我们不区分地使用“点数”（对应指数）和“价格”（对应个股）。给定某个股票指数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：

|  |
| --- |
| *图2.17.1：链式计算筹码分布图示* |
|  |
| *资料来源：WIND数据库，诺德基金FOF管理部* |

简而言之，在挂牌日，所有筹码位于基准点数上。挂牌日之后的任意一天，筹码从昨日各个价位以换手率等比例流出，汇入当日的收盘价，完成一次筹码的流动。每个交易日筹码分布都会更新一次，通过这样的链式计算可以得到任意一天的筹码分布。这里有几个比较重要的假设：第一，想要获得精确的筹码分布（即市场中每一张股票的成本），就必须掌握每一笔交易的成交价和成交量，这显然是不可能的，因此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，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，所以我们以收盘价替代。

下面以沪深300指数（000300.SH）为例提供几个筹码分布的例子。图2.17.2显示了2017年8月18日的筹码分布（为了提高可视度，将每10点指数内的筹码加和显示）。该日正处于一个较长的、稳定的上升背景，筹码分布是一个典型的类似正态分布的形态。超额峰度达到了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（当前价高于成本价的筹码比例）达到了85%。

|  |
| --- |
| *图2.17.2：沪深300指数（000300.SH）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2.17.3是沪深300指数（000300.SH）在2016年6月22日的筹码形态，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2.17.4是2018年4月17日的筹码形态，该日位于一个长时间的震荡下挫的走势中，可以发现筹码同样较为离散，超额峰度仅有0.34。

|  |
| --- |
| *图2.17.3：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2.17.4：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* |

目前市场上一个追随者很多的观点是尖的筹码峰（高峰度）是强劲的上升信号。进一步推广这个观点，我们猜测，今天的筹码形态可以预测将来的股指涨跌。这个猜测的核心论点是：在一个行情阶段内的筹码形态有一些共同性，可以被划分为一类。剩下的问题是，世界上没有哪两片叶子是一模一样的，也没有哪两个筹码分布图是一模一样的，哪些指标可以帮助我们分类千奇百怪的筹码分布呢？下一节将会给出答案。

1. **基于筹码分布的股指预测模型**

本文中提到的变量被整理在下面的变量表中，以便读者参考：

|  |
| --- |
| *表2.17.1：变量的含义* |
| |  |  | | --- | --- | | 变量 | 含义 | | N | 模型的预测时长，即预测目标为第N天后的涨跌 | | M | 计算当前价格相对位置e，平均持仓成本位置f使用的回溯时间窗口长度 | | c | 控制去除筹码极端值的比例 | | e | 当前价格的相对位置，计算公式见下文 | | f | 平均持仓成本的相对位置，计算公式见下文 | | g | 超额峰度，详细计算公式见下文 | | m | 控制根据预测开仓需要的最小信心 | | k | 某日的指数点数（价格） | | P | 在某价位上的筹码概率（筹码比例） | |
| *资料来源：诺德基金FOF管理部* |

样本数据方面，我们从WIND数据库获取了沪深300指数（000300.SH）、中证500指数(000905.SH)和上证50指数（000016.SH）自挂牌以来的日频率数据，然后利用上述算法计算日频率的筹码分布。若无另外说明，以下我们都以沪深300指数（000300.SH）为研究对象，其他指数将会在结尾讨论。

哪一个类型的模型适合描述筹码和涨跌的关系呢？线性模型很快被排除了，考虑到筹码分布其实是形态分析，市场参与者进行形态分析的过程相当主观，不太可能遵从一个简单的线性模型。尝试过决策树(Decision Tree)后，它的表现脱颖而出。最后我们采用基于决策树之上的集成模型**随机森林**(Random Forest)，因为它可以解决决策树预测稳定性不足的缺点。

理解了决策树的原理就可以很快理解随机森林的原理。直观地讲，把一个样本的若干个特征的取值从一颗决策树的根部输入，在每个节点按一个条件（比如超额峰度>0）划入两棵子树中的一个，反复进行直到抵达叶节点，每个叶节点会对应一个预测标签（比如涨或跌）。图2.17.5给出了一棵决策树的一部分，最左侧的结点就是一个叶节点（它没有子树）。

顾名思义，随机森林就是包含多棵决策树的集成模型。对比一棵决策数，随机森林随机建立多棵决策树，综合它们的投票（vote）给出预测。

一个交易策略仅仅需要一个二分的信号就可以运作，不需要数值预测，所以随机森林的预测目标是**N个交易日后的涨跌**，也就是说这是一个分类模型。

在训练随机森林之前，我们要对原始筹码分布序列进行预处理。设拥有最大筹码比例的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其筹码概率之和仍为1。*c*的默认值为0.005，在后面会对此参数进行优化分析。*c*的作用是去除极端价格（如发行价）上的微量筹码，防止它们扰乱随机森林的训练，也凸显了主力持仓。

|  |
| --- |
| *图2.17.5：训练完毕的决策树的一部分* |
|  |
| *资料来源：诺德基金FOF管理部* |

从五个初始指标出发，尝试不同的组合后，挑选出三个最有效的筹码特征作为模型输入：

给定一个筹码分布：设价格为，对应的概率为。在第j日，收盘价为,回溯之前紧邻的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。

1. **当前价格的相对位置**

当前价格的相对位置

1. **平均持仓成本的相对位置**

平均（期望）持仓成本 。平均持仓成本的相对位置

1. **超额峰度**

超额峰度

参数M的默认值为60，用今天前的60个交易日的行情作基准价格区间（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。

超额峰度在一定程度上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，市场可能会比较缺乏动能，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。因此，峰度是一个重要的特征，有助于决策树的树杈将不同筹码形态下蕴含的不同市场趋势区分开来。

预测目标是N个交易日后的涨跌标签，N的值代表了预测视野。实验证明，筹码特征的预测能力随着预测视野的变长，先增强后减弱，在中期最好。我们选定预测N=60个交易日后的涨跌。

滚动计算沪深300指数60日周期的涨跌标签(label)。方法是，假设第T日的点数为1000，第T+N日的点数为1030，则第T日的分类标签为涨(1)；反之，则为跌（0）。

|  |
| --- |
| *图2.17.6：训练集和测试集划分* |
| C:\Users\admin\Desktop\演示文稿2.png |
| *资料来源：诺德基金FOF管理部* |

以图2.17.6的方法训练随机森林，测试结果展示在表2.17.2中。2016年6月13日以后的672个交易日，模型正确的预测了四分之三的样本60个交易日后的相对价格涨跌。图2.17.7将预测信号叠加在了指数上，可以看到，模型擅长发掘长时间的、稳定的上升趋势，在震荡和下行区间的表现相对差些。

|  |
| --- |
| *表2.17.2：预测结果* |
| |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | **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 | 75.30% | 507 | 77.58% | 165 | 74.56% | |
| *资料来源：诺德基金FOF管理部* |
| *图2.17.7：多空信号* |
|  |
| *资料来源：WIND数据库，诺德基金FOF管理部* |

如何解释这样的预测结果呢？如图2.17.8-2.17.11所示，**三个特征呈现出集群的特点**，**临近的点很有可能分享同样的涨跌标签**。因此，决策树可以找到很好的空间划分分开涨点和跌点。

抛开决策树，肉眼来看能发现两个现象：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。

|  |
| --- |
| *图2.17.8：筹码分布特征的三维展示（实际）（视角1）* |
|  |
| *资料来源：WIND数据库，诺德基金FOF管理部* |

|  |
| --- |
| *图2.17.9：筹码分布特征的三维展示（实际）（视角2）* |
|  |
| *资料来源：WIND数据库，诺德基金FOF管理部* |
| *图2.17.10：筹码分布特征的三维展示（预测）（视角1）* |
|  |
| *资料来源：WIND数据库，诺德基金FOF管理部* |
| *图2.17.11：筹码分布特征的三维展示（预测）（视角2）* |
|  |
| *资料来源：WIND数据库，诺德基金FOF管理部* |

**三、策略构建**

以图2.17.6中测试集时间段回测交易策略，具体如下：

* 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；
* 每日收盘时，根据模型的当日预测结果，以收盘价做多/做空，60个交易日后以收盘价平仓；
* 在回测期结束前清算所有未结头寸，以最终现金总额计算回报；
* 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：信号过滤阈值m（）。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，同时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（其和为1），可以理解成投资者的信心。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，这时我们掌握的对未来走势的信息最少。我们选定一个阈值0.25m，滤掉最不确定的，也就是满足的预测。对于这些被过滤掉的预测我们就不开仓，减少可能的损失（同时也减少了可能的收益）。

|  |
| --- |
| *表2.17.3：概率阈值分析* |
| |  |  |  |  | | --- | --- | --- | --- | | **信号过滤阈值m** | **头寸数量** | **最大回撤（%）** | **超额收益率（%）** | | 0.3 | 258 | 5.58 | 0.18 | | 0.4 | 382 | 6.06 | 5.97 | | 0.5 | 382 | 6.06 | 5.97 | | 0.6 | 382 | 6.06 | 5.97 | | 0.7 | 452 | 6.08 | 7.92 | | 0.8 | 452 | 6.08 | 7.92 | | 0.9 | 521 | 6.51 | 8.72 | | 1.0 | 636 | 5.55 | 11.05 | |
| *资料来源：诺德基金FOF管理部* |

|  |
| --- |
| *图2.17.12：概率阈值对投资风格的影响* |
|  |
| *资料来源：诺德基金FOF管理部* |

信号过滤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2.17.3和图2.17.12中看出，随着阈值的不断提高，发起的头寸（多+空）更多，策略更加激进，会进行信心不足的投机，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跳水，但不建议将阈值设置为1（即不过滤任何信号）,因为多标签和空标签预测概率都是50%的情况下，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。我们之后默认设置。

详细的回测结果如图2.17.13和图2.17.14所示，沪深300指数线上方的红色符号刻画了看多开仓，下方的绿色符号刻画了看空开仓，另一条红色曲线表示累计收益走势。指标见表2.17.4，筹码分布发出的信号准确，年化收益率14.30%，超额年化收益率58.55%，胜率82%，赔率2.18，累计最大回撤仅有6.08%，收益回撤比达到1.30，夏普比率达到1.29。

为了避免严重的回撤，每次交易设置了恒定金额(初始资金的五十分之一)。相比全仓做空/做多,本投资方案相对稳健，有更多风险需求的投资者可以定义更激进的投资方案。

|  |
| --- |
| *图2.17.13：实证结果* |
|  |
| *资料来源：WIND数据库，诺德基金FOF管理部* |
| *图2.17.14：累计净值和每日收益* |
|  |
| *资料来源：诺德基金FOF管理部* |
| *表2.17.4：择时策略指标表现一览* |
| |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | **累计净值** | **年化收益率** | **超额年化**  **收益率** | **胜率** | **赔率** | **累计最大**  **回撤** | **收益**  **回撤比** | **夏普比率** | | 1.4929 | 14.30% | 7.92% | 81.64% | 2.18 | 6.08% | 1.30 | 1.29 | |
| *资料来源：诺德基金FOF管理部* |

**四、参数敏感性测试**

本策略涉及到三个主要参数：筹码分布预处理参数c，回溯时间窗口，和预测视野。在上述的回测分析中，统一使用了,和的参数设置。为了寻找最优的参数设置，并且观察择时策略对于参数的敏感性,进行了参数的网格搜索（Grid Search）。接下来，统计了不同的M，c组合的收益回撤比。图2.17.11中x轴和y轴分别代表了c和M的设置，z轴代表收益回撤比。从下图来看，当,参数M在20-80之间时，收益回撤比都高于1，在时达到最优收益回撤比1.62。

本文中默认参数设置未经过优化，但处于参数高原上，在默认参数附近的收益回撤比敏感性不高。实际应用中，参数还是需要灵活选择，最优参数可能是对于测试集过优化的产物。为了避免过优化，业界中较常用的参数优化方法是交叉验证（Cross-Validation）。原理是将训练集十等分，对于每一个参数组合，用九份样本训练，剩下的一份评估，轮转这份用于评估的样本，最后取平均分代表这个参数组合的质量。选出最优参数后再用测试集评估拟合度。

|  |
| --- |
| *图2.17.11：参数的网格搜索* |
|  |
| *资料来源：诺德基金FOF管理部* |

**五、总结**

上述基于筹码分布的预测模型和交易策略，因为涉及到了做空投机，主要应用于股指期货交易。除了沪深300指数(000300.SH)，中国内地股市的另外两支有期货交易的指数——中证500指数（000905.SH）和上证50指数（000016.SH）也进行了回测。使用和上文一样的配置，中证500策略的指标表现见图2.17.5，基本符合预期，甚至获得了更高的超额年化收益率。

|  |
| --- |
| *表2.17.5：中证500指数（000905.SH）择时策略指标表现一览* |
| |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | **累计净值** | **年化收益率** | **超额年化**  **收益率** | **胜率** | **赔率** | **累计最大**  **回撤** | **收益**  **回撤比** | **夏普比率** | | 1.3824 | 11.41% | 16.82% | 76.04% | 2.11 | 7.74% | 2.17 | 1.05 | |
| *资料来源：诺德基金FOF管理部* |

然而策略应用于上证50指数（000016.SH）的表现差强人意，策略指标表现见表2.17.6，超额年化收益率小于0，赔率仅有0.61。主要原因是预测模型会系统性地误判近5年大的下跌区间，导致损失较大。

|  |
| --- |
| *表2.17.6：上证50指数（000016.SH）择时策略指标表现一览* |
| |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | **累计净值** | **年化收益率** | **超额年化**  **收益率** | **胜率** | **赔率** | **累计最大**  **回撤** | **收益**  **回撤比** | **夏普比率** | | 1.0339 | 1.12% | -9.19% | 64.96% | 0.61 | -18.88% | -0.49 | -0.26 | |
| *资料来源：诺德基金FOF管理部* |

综上所述，对于沪深300(000300.SH)期指和中证500（000905.SH）期指，筹码分布形态结合当前价格和平均持仓成本的相对位置是有效的量化择时数据，作为择时信号源在历史区间内具有优异的表现。筹码形态相对大盘具有领先性质，可以较为准确地预测市场的涨跌；对于上证50（000016.SH）期指，可能受其成分股样本容量偏小的影响，策略无法提供高于简单持有的回报。

本策略基于机器学习领域的经典模型——随机森林。学习模型的特点是可以随着时间的推移不断更新训练集来学习新的信息。那么当模型的预测力衰减的时候，我们除了放弃该策略外还有其他的选择：

* 加入最近的新样本，重新训练随机森林使得模型反映最近的市场变化；
* 重新搜索最优参数，如c、M、N。

筹码分布策略的优势在于数据需求小，模型简洁，计算资源消耗少，只需要历史行情，不依赖任何金融市场以外的数据即可训练。缺点是适用性可能局限于一部分指数。

本报告初步得到了一些实用性的研究结论，但也存在如下不足，同时也提出几个未来可以深入研究的方向：

* 首先，筹码分布是形态分析，本报告只用到一个量化筹码形态的指标——峰度，除此之外还可能存在更有效的形态指标。未来可以继续研究更加复杂的形态特征，比如筹码的单峰、双峰、多峰形态等等。
* 其次，仅仅借助过往行情，能窥探到的未来市场走势是十分有限的，在重大的、意外的信息冲击下，策略持有的头寸可能面临较大损失。可以结合收入/库存周期等宏观数据和网络热度等互联网数据，生成一个复合策略。

本报告的源代码和原始数据可在此处下载：<https://github.com/yangrq1018/nuode-algotrade>