机器学习优化股票多因子模型的研究与实证分析

摘要

本文以中国 A 股市场所有股票和 Auto-Trader 中十二大类 500 多个因子作为研究对象，利用机器学习方法对多因子选股模型进行了优化，并基于风险管理模型建立了 SVM-RC 多因子选股策略。首先，利用 IC 显著性分析从大量候选因子中甄别出选股能力最强的基础因子，进一步结合主成分分析和聚类分析构建新的选股主因子，并以此建立了多因子选股模型。其次，本文利用三种较为知名的机器学习算法对所建选股模型进行优化与增强，通过比较发现支持向量机算法的提升效果最优，并以此为例提出了消除机器学习算法在选股模型上的过拟合现象的处理方法，建立了 SVM-CO 提升模型。最后，在上述模型基础之上，结合三种风险控制方法建立了一整套可实际操作的量化投资策略。通过大量的实证分析，发现本策略不仅表现出年化收益率超过 50%的高收益性，在不同市场行情下都能连续获利的稳定性，而且具有可以将最大回撤控制在 10%以内的低风险性。

关键词：量化投资、多因子模型、主成分分析、机器学习、风险管理

目录

0 引言 1
0.1 问题重述 1
0.2 本文主要工作和创新点 1
1 单因子筛选和主因子合成 3
1.1 数据的选取与预处理 3
1.1.1 数据的选取 3
1.1.2 残缺测量矩阵及其完备化 4
1.1.3 因子数据去极值 4
1.1.4 因子数据标准化 5
1.2 IC 显著性筛选候选因子 5
1.2.1 计算因子 IC 值 5
1.2.2 因子作用方向 6
1.2.3 因子显著性分析 6
1.3 单因子策略构建和绩效分析 7
1.3.1 单因子选股策略构建 7
1.3.2 单因子选股策略绩效分析 8
1.4 主因子合成 9
1.4.1 主成分分析方法 9
1.4.2 正交旋转 11
1.4.3 等权重线性模型选股策略 12
2 机器学习算法增强因子选股模型 15
2.1 利用随机森林算法提升多因子选股模型 15
2.1.1 “随机森林”基本原理[26] 15
2.1.2 随机森林基本构建步骤 15
2.1.3 随机森林算法提升多因子选股模型测试 16
2.2 利用 AdaBoost 提升算法提升多因子选股模型 17
2.2.1AdaBoost 算法基本原理 18
2.2.2 AdaBoost 二元分类算法基本步骤 18
2.2.3 AdaBoost 二元分类算法提升多因子选股模型 19
2.3 利用支持向量机提升因子表现 20
2.3.1 线性支持向量机算法原理 21
2.3.2 核支持向量机 22
2.3.2 支持向量机算法提升多因子选股模型 23
2.4 对最优机器学习提升模型的过拟合分析与处理 24
2.4.1 过拟合问题的原因 25
2.4.2 过拟合问题的应对 25
2.4.3 交叉验证方法 26
2.4.4 机器学习算法模型过拟合评价标准 26

2.4.3 以基于 SVM 算法的多因子选股模型为例的过拟合处理方案.....27
3 风险控制 32
3.1 行业、风格中性组合 32
3.1.1 行业中性对冲.....32
3.1.2 风格中性.....33
3.2 波动率控制风险 34
3.2.1 波动率模型构建.....34
3.3 模型实证检验 34
3.4 实证结果 35
3.4.1 模型样本期内回测结果.....35
3.4.2 模型长跨度时期回测检验.....36
3.4.3 分行情实证分析.....36
4 结语 39
参考文献 41

0 引言

0.1 问题重述

（一）针对 Auto-Trader 中 2016 年 1 月 1 日至 2018 年 9 月 30 日十二大类因子的日频数据，建立单因子选股策略对因子与股票收益率间的关系进行探究，通过对策略的绩效分析，筛选出年化夏普比率最高的因子。

（二）分别使用不同的机器学习算法对筛选出的因子进行提升，利用 2016 年 1 月 1 日至 2018 年 9 月 30 日我国 A 股市场的数据对比不同机器学习算法选股策略与等权重线性多因子模型选股策略的回测绩效。

（三）在机器学习算法选股策略的基础上使用风险对冲方法对选股策略进行风险控制，在确保选股策略收益的同时将选股策略的最大回撤控制在 10%以内。

0.2 本文主要工作和创新点

本文基于题目中所给的分析角度和学术领域内已有研究工作的基础上，结合主成分分析方法、机器学习算法与对冲思想，建立了一套完整的量化投资策略。

（一）通过数据分析筛选最优因子的角度。本文首先对预处理后的因子数据通过 IC 显著性分析和选股能力鉴别得到了最优的因子，其次对于筛选出的因子数据进行主成分分析方法的基础上，使用 k-means 聚类分析和正交旋转得到了两个新的且分别能代表不同经济意义的合成因子，最后在此基础上建立了等权重多因子选股模型。

（二）机器学习算法提升因子表现的角度。首先本文使用三种不同类型的机器学习算法去提升上述等权重多因子选股模型的表现，并通过对不同提升模型的回测绩效分析来进行不同算法优化效果的筛选，得到了提升效果最为优异的一种机器学习算法。其次本文基于 AUC 指标与 k 折交叉检验，针对提升效果最优的机器学习算法，提出了消除过拟合现象的处理方案，进而建立了 SVM-CO 提升模型，并对其进行了回测绩效分析。

（三）对选股策略进行风险控制的角度。在 SVM-CO 提升模型的基础上连续使用了三种风险对冲方法，在确保选股模型收益的同时又降低了模型的回撤水平，

并最终建立了 SVM-RC 多因子选股策略。

经过实证分析发现，SVM-RC 多因子选股策略不仅具有稳定的高收益，而且能够充分适应各种市场行情，同时还具有较低的投资风险。

1 单因子筛选和主因子合成

本章针对问题一，首先对 Auto-Trader 中各大类因子数据进行预处理，然后通过 IC 显著性分析和因子选股能力筛选出几个最优因子，最后通过 PCA 和聚类分析得到最终的合成因子。

1.1 数据的选取与预处理

本节主要详细介绍了本文数据来源以及对原始数据进行预处理的步骤。

1.1.1 数据的选取

本文选取中国 A 股市场所有股票以及 BP 股票量化因子库[①]作为研究对象，研究区间为 2016 年 1 月 1 日至 2018 年 9 月 30 日。根据 BP 因子库分类将因子分为十二大类：基础科目衍生类、质量类、收益风险类、情绪类、成长类、常用技术指标类、动量类、价值类、每股指标类、模式识别类、特色技术指标、行业与分析师类（表 1）。数据来源于点宽网商用数据库，具有一定的权威性，且数据质量较高。数据分析软件和编程软件为 SPSS、AutoTrader、Python 与 MATLAB。

表 1 因子分类介绍

基础科目衍生类	该类别主要由公司原始财报数据构成，是其他类别数据的基础。
质量类	该类别因子基于上市公司的财务数据计算所得出，可以更加有效的观察公司的整体状况。
收益风险类	该指标从股票收益，风险，以及风险收益比三个角度来度量股票的表现。
情绪类	该指标以成交量及换手率数据为基础，并结合 k 线数据来判断市场上的资金走向。
成长类	该指标反映了每家上市公司的成长性。

① 数字动能量化研究部根据上市公司财务报表，交易行情计算而成。

常用技术指标类	该类别包含主流的技术指标，使用前复权价格计算，反映了股票的量价信息。
动量类	该指标通过计算不同种类的价格动量，帮助投资者综合判断股价的变化趋势。
价值类	该指标主要反映市场对上市公司的估值大小。
每股指标类	该指标从每股角度展现股票的各种盈利能力，大多数因子的计算涉及财务报表数据。
模式识别类	该指标是对典型的 K 线模式进行分类，从中找出股价运动的规律并以此对买卖点进行判断。
行业与分析师类	该指标包含分析师关于个股的预测及相关评级指标，以及个股相对于其所在行业的表现。
特色技术指标	该指标是 BP 运用量价指标计算出对于常见技术指标的补充。

1.1.2 残缺测量矩阵及其完备化

记 $\{X_j\}_{j=1}^{N}$ 为所考虑的全部因子，$\{a_i\}_{i=1}^{M}$ 为股票所对应的数据，其中 N, M 分别表示因子和股票的个数，a_i 为 N 维列向量，$A=(a_1,\dots,a_M)^T$ 为测量矩阵，A 的每一行为一个股票，每一列为一个因子。由于某些指标因为停牌或资产重组等原因而无法获取，从而导致测量矩阵中的某些数据缺失，称这样的测量矩阵为残缺的。对于残缺测量矩阵无法直接使用回归分析，因而需要对其进行完备化。即利用补齐的办法，构造一个能够充分反应原始数据，且不存在数据缺失的测量矩阵。经过验证，本文所使用数据来源质量较高，测量矩阵 A 缺失值并不太多，由此只需要简单利用拉格朗日插值法对残缺矩阵 A 补全数据。

1.1.3 因子数据去极值

为避免数据中极端值对回归结果产生过多影响，本文使用“拉依达准则（3σ 准则）”对数据进行处理，将超过上下限的极端值用上下限值代替。

设因子数据为 $x_1,\dots,x_n$，其算术平均值为 $\bar{x}$，则剩余误差为

$$v_i = x_i - \bar{x} \qquad (i = 1,2,\dots,n)$$

从而得到标准偏差为

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

若某个测量值 x_i 的剩余误差 v_i 满足下式

$$|v_i| = |x_i - x| > 3\sigma$$

则认为 x_i 有比较大的误差，应予代替。

1.1.4 因子数据标准化

由于选取的数据包含不同类别的因子，而不同因子之间在取值范围与度量单位等方面存在较大差异。因此需要将因子数据进行标准化处理，将其按照一定比例缩放到 $0\sim1$ 之间。本文采用零均值标准化对数据进行处理，经过处理的数据，其均值为 0，标准差为 1。记 $x = \{x_i\}_{i=1}^{N}$ 为原始数据， $x^* = \{x_i^*\}_{i=1}^{N}$ 为标准化的数据，则转化公式如下：

$$x_i^* = \frac{x_i - \mu}{\sigma}, \qquad i = 1,\dots,N$$

其中 μ 是原始数据的均值，而 σ 为原始数据的标准差。

1.2 IC 显著性筛选候选因子

本节主要计算因子库中因子的 IC 值，并通过 IC 显著性筛选出表现较为优异的基础因子。

1.2.1 计算因子 IC 值

在金融领域，信息系数 IC（information coefficient）是一个用于衡量预测值优劣的绩效指标，可以将它作为衡量因子收益预测能力的重要参数，由此来筛选出一些预测能力较为优秀的因子。由于信息系数与相关系数相似，因此本文通过计算因子在各个股票上的因子暴露与对应股票下一期收益序列间的 Spearman 相关系数序列来确定信息系数，即

$$IC_f = \text{corr}（FactorValue_f^{t-1}，StockYeild_f^t）$$

$$\mathrm{IC} = (IC_1, \ldots, IC_M)$$

其中 $FactorValue_f^{t-1}$ 为 $t-1$ 期对应第 f 只股票的因子值，$StockYeild_f^t$ 为 t 期对应第 f 只股票收益率。由于采用按月换仓的交易策略，同时部分基于基本面数据的因子需要根据各公司月度财报数据进行更新，故本文以月为单位周期滚动计算相关系数，得到因子的 IC 值序列。

1.2.2 因子作用方向

因子作用方向由因子 IC 值决定。IC 值为负，表示对应股票的因子值越小此股票的收益率越高，因子值越大此股票收益率越低；IC 值为正，则表示对应股票的因子值越大此股票收益率越高，因子值越小股票收益率越低。所以本文使用 IC 值的绝对值去判断因子值大小与股票收益之间的关系。

1.2.3 因子显著性分析

每个因子针对每一只个股都有一个与之对应的 IC 绝对值，但是因子对单一个股的影响情况并不能有代表性的反映出因子值与股票收益率序列之间的显著性。本文利用每个因子针对全市场所有股票的 IC 绝对值序列的均值，来判断此因子值与股票市场收益率之间的关系是否显著，以充分反映出因子的优劣。由此本文将因子显著性定义为因子针对全市场股票的 IC 绝对值序列的均值，序列均值越大，表明该因子对股价的相关性越高，则其对股价的预测能力就越强。

通过对 BP 因子库中的 500 多个因子所进行的测试，并根据因子显著性最终筛选出最优的 20 个候选因子（详见表 2，），测试数据为 2016 年 1 月 1 日至 2018 年 9 月 30 日中国 A 股市场所有股票的历史交易数据及 BP 股票量化因子库数据。

表 2 候选因子 IC 显著性值统计表

因子简称	IC 显著性值	因子简称	IC 显著性值
LINEARREG_INTERCEPT	0.238293184	EMA20	0.216676481
LINEARREG	0.235324254	WMA	0.215869591
HT_TRENDLINE	0.232439959	NegMktValue	0.21556976
DEMA	0.219244186	LFLO	0.212727672

EMA12	0.218957538	EMA26	0.212184201
BBI	0.218621928	TEMA10	0.211502877
MA20	0.218477633	MktValue	0.211443091
EMA10	0.218382184	TEMA5	0.21011255
MIDPOINT	0.217909082	KAMA	0.208771848
MA10	0.216985748	T3	0.207979982

1.3 单因子策略构建和绩效分析

本节在前文筛选出的候选因子基础之上，根据因子的选股能力再次精选出其中几个最为优异的因子作为最后的筛选结果。

1.3.1 单因子选股策略构建

在使用 IC 显著性分析筛选出相对显著的候选因子后，一个可以更加直接考察该因子选股能力的方式，就是观察因子值高的股票是否能够保持盈利。由此本文建立如下单因子测试模型（模型 1），以通过策略绩效来挑选因子。

模型（1）：单因子测试模型

测试样本范围：全市场

测试样本期：2016 年 1 月 1 日至 2018 年 9 月 30 日

模型建立要点：

1. 初始资金为 1000 万元整，手续费为双边千分之三，每月月初调仓。
2. 在每个截面期的最后一个交易日（本文中的截面期为一个月），提取样本内股票因子值，并剔除因子值缺失的股票。
3. 按照股票因子值的大小将样本内股票排序，根据因子值作用方向选取因子值最好的前 N 支（本文中 N 为全市场股票中的百分之三）股票作为备选股票池。
4. 在下一个截面期的首个交易日，以当天的收盘价将持仓股票更换为备选股票池中股票（根据不同股票价格等权重配置资金）并剔除当天因停牌、涨停等因素不能交易的股票。

5. 最后对 N 支股票的历史收益率进行回测，计算其年化收益率、最大回撤、夏普比率等值。

1.3.2 单因子选股策略绩效分析

使用模型（1）对表 2 中的 20 个候选因子进行回测，得到与其对应的单因子选股策略的年化收益率、最大回撤、夏普比率（详见表 3）。

表 3 单因子选股回测结果

因子简称	年化收益	最大回撤	夏普比率
LINEARREG_INTERCEPT	10.81%	23.59%	0.54
LINEARREG	9.60%	23.99%	0.46
HT_TRENDLINE	10.58%	23.62%	0.52
DEMA	9.21%	23.96%	0.43
EMA12	8.95%	24.20%	0.41
BBI	8.85%	24.26%	0.41
MA20	9.21%	24.13%	0.43
EMA10	8.85%	24.26%	0.41
MIDPOINT	9.65%	23.90%	0.46
MA10	9.01%	24.03%	0.42
EMA20	9.29%	24.11%	0.44
WMA	9.30%	24.09%	0.44
NegMktValue	15.38%	20.96%	0.70
LFLO	15.08%	19.75%	0.73
EMA26	9.45%	24.03%	0.45
TEMA10	9.36%	23.90%	0.44
MktValue	18.82%	20.84%	0.89
TEMA5	9.34%	24.10%	0.44
KAMA	9.78%	23.85%	0.47
T3	7.06%	25.38%	0.30

由于夏普比率能够同时反映出收益与风险，因而本文以此作为度量选股能力的指标。通过观察上表，选取夏普值最高的六个因子为最终因子池：MktValue、NegMktValue、LFLO、LINEARREG_INTERCEPT、HT_TRENDLINE、KAMA。

1.4 主因子合成

为了防止不同因子数据之间存在内在关联而导致信息重叠，同时也为了对高维特征进行属性约减，以提高后文将要进行机器学习的效率，本节对 1.3 节所得的六个基本因子进行了主成分分析与聚类分析，从而构建出了两个足够代表原始高维因子数据的主因子。

1.4.1 主成分分析方法

主成分分析方法（PCA）是一种被广泛使用的数据降维算法，其主要思想是将 n 维特征映射到 k 维上，这全新的 k 维正交特征也被称为主成分，是在原有 n 维特征的基础上重新构造出来的 k 维特征。

记 $\mathbf{x}=(x_1,\dots,x_n)$ 是一个 n 维样本数据，样本均值与方差分别为

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\qquad S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

样本 $\mathbf{x}$ 和样本 $\mathbf{y}$ 的协方差为

$$\mathrm{cov}(\mathbf{x},\mathbf{y})=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

协方差为正则正相关，为负则负相关，为零则不相关。

设 $\mathbf{X}=(\mathbf{x}_1,\cdots,\mathbf{x}_m)^T$ 是 m 维随机变量，$\mathbf{x}_i$ 是一个 n 维样本数据，则 $\mathbf{X}$ 是一个 m 行 n 列的矩阵。$\mathbf{X}$ 的协方差矩阵是一个 m 阶的方阵

$$\mathrm{cov}(\mathbf{X})=\begin{bmatrix}\mathrm{cov}(\mathbf{x}_1,\mathbf{x}_1) & \mathrm{cov}(\mathbf{x}_1,\mathbf{x}_2) & \cdots & \mathrm{cov}(\mathbf{x}_1,\mathbf{x}_m)\\ \mathrm{cov}(\mathbf{x}_2,\mathbf{x}_1) & \mathrm{cov}(\mathbf{x}_2,\mathbf{x}_2) & \cdots & \mathrm{cov}(\mathbf{x}_2,\mathbf{x}_m)\\ \vdots & \vdots & \ddots & \vdots\\ \mathrm{cov}(\mathbf{x}_m,\mathbf{x}_1) & \mathrm{cov}(\mathbf{x}_m,\mathbf{x}_2) & \cdots & \mathrm{cov}(\mathbf{x}_m,\mathbf{x}_m)\end{bmatrix}$$

设 $\mathrm{cov}(\mathbf{X})$ 的特征值按从大到小排列为 $\lambda_1\geq\cdots\geq\lambda_n$，其对应的单位特征向量分别为 $\mathbf{v}_1,\dots,\mathbf{v}_n$，则第 i 个主成分即为 $\mathbf{v}_i$，对应的因子线性组合记为 F_i，则

$$F_i=(x_1,\dots,x_n)\mathbf{v}_i,\qquad i=1,\dots,n$$

前 k 个主成分的累计方差贡献率记为

$$E(k)=\frac{\sum_{j=1}^{k}\lambda_j}{\sum_{j=1}^{n}\lambda_j}$$

累积方差贡献率越高，相应的主成分能够涵盖的信息量就越大。本文确定 k 使得 $E(k)>85\%$，同时要求相应的特征值大于 1。

对经过标准化的实际数据通过 SPSS 进行主成分分析后，得到如下表 4 所示的主成分方差贡献累积和特征值。

表4：主成分方差贡献累积和特征值表

成份	初始特征值			提取平方和载入		
	合计	方差的 %	累积 %	合计	方差的 %	累积 %
1	4.056	67.598	67.598	4.056	67.598	67.598
2	1.182	19.693	87.291	1.182	19.693	87.291
3	.359	5.991	93.282			
4	.262	4.368	97.651			
5	.130	2.173	99.824			
6	.011	.176	100.000			

由表 4 可知，提取出大于 1 的前 2 个特征值，其累积方差贡献率达到了 87.291%，对应的因子载荷矩阵如下表 5

表5：因子载荷矩阵a

	成份	
	1	2
'LINEARREG_INTERCEPT'	.649	.679
'HT_TRENDLINE'	.883	.344
'NegMktValue'	.862	-.468
'MktValue'	.863	-.242
'LFLO'	.863	-.460
'KAMA'	.789	.335

将因子载荷矩阵除以主成分相对应的特征值开平方根，便得到最终 2 个主成

分因子

$$\mathbf{v}_1 = \begin{bmatrix} 0.3223 \\ 0.4382 \\ 0.4280 \\ 0.4285 \\ 0.4285 \\ 0.3918 \end{bmatrix}, \qquad \mathbf{v}_2 = \begin{bmatrix} 0.6245 \\ 0.3164 \\ -0.4305 \\ -0.2226 \\ -0.4231 \\ 0.3081 \end{bmatrix}$$

1.4.2 正交旋转

如果两个主成分对因子的影响程度相似，则这两个主成分是很难区分的，为此需要进行适当的坐标转换，将其换成新的主成分。

记 $A_{n\times k} = (\mathbf{v}_1, \dots, \mathbf{v}_k)$ 为主成分系数矩阵， A 的每一行代表 k 维空间中的一点，共计 n 个点。利用 k-mean 聚类法，将这 n 个点分为 k 类。根据上述分类将 A 改写为分块矩阵形式

$$A = \begin{bmatrix} A_1 \\ \vdots \\ A_k \end{bmatrix},$$

其中 A_i 是 n_i 行 k 列的矩阵，其所有行表示第 i 类中的所有 n_i 个点，则 $n_1 + \cdots + n_k = n$。

构造 m 阶正交旋转矩阵 S，对 A 进行旋转得到 $B = AS = (B_1, \dots, B_m)$，使得 B_i 的分量只在 A_i 所对应行的位置上显著，在其它行不显著。从而得到如下优化问题：

$$\max \sum_{i=1}^{m} \left\| A_i S_i^T \right\|^2$$
$$\text{s.t.} \quad SS^T = I$$

其中 S_i 是 S 的第 i 行。

由上节计算结果可知

$$A = (\mathbf{v}_1, \mathbf{v}_2) = \begin{bmatrix} 0.3223 & 0.6245 \\ 0.4384 & 0.3164 \\ 0.4280 & -0.4305 \\ 0.4285 & -0.2226 \\ 0.4285 & -0.4231 \\ 0.3918 & 0.3081 \end{bmatrix}$$

A 的每行对应平面中的一个点，共有六个点。利用 k-means 聚类可将这六个点分为两类：第 1、2、6 个点属于第一类，第 3、4、5 个点属于第二类。

通过求解上述优化问题，得到正交旋转矩阵：

$$S = \begin{bmatrix} 0.6534 & 0.7570 \\ 0.7570 & -0.6534 \end{bmatrix}$$

对 A 进行旋转即得到本文最终的主因子合成系数矩阵：：

$$B = AS = \begin{bmatrix} 0.6833 & -0.1641 \\ 0.5260 & 0.1252 \\ -0.0462 & 0.6053 \\ -0.1115 & 0.4698 \\ -0.0403 & 0.6008 \\ 0.4892 & 0.0952 \end{bmatrix}$$

B 的第一列对应第一个主因子，与其关系密切的是第 1、2、6 个基本因子，即为三个价值类因子 NegMktValue、MktValue、LFLO，其所包含的信息主要为公司的基本面信息，由此将其命名为 Value 因子。

B 的第二列对应第二个主因子，与其关系密切的是第 3、4、5 个基本因子，即为三个技术指标类因子 LINEARREG_INTERCEPT、HT_TRENDLINE、KAMA，其所包含的信息主要为股票行情 K 线的基本数据，于是将其命名为 Index 因子。

根据矩阵 B 得到两个主因子表达式分别为：

$$\begin{aligned} \text{Value} = {} & 0.6833 \times \text{'LINEARREG\_INTERCEPT'} + 0.5260 \times \text{'HT\_TRENDLINE'} \\ & - 0.0462 \times \text{'NegMktValue'} - 0.1115 \times \text{'MktValue'} - 0.0403 \\ & \times \text{'LFLO'} + 0.4892 \times \text{'KAMA'} \end{aligned}$$

$$\begin{aligned} \text{Index} = {} & -0.1641 \times \text{'LINEARREG\_INTERCEPT'} + 0.1252 \times \text{'HT\_TRENDLINE'} \\ & + 0.6053 \times \text{'NegMktValue'} + 0.4698 \times \text{'MktValue'} + 0.6008 \\ & \times \text{'LFLO'} + 0.0952 \times \text{'KAMA'} \end{aligned}$$

1.4.3 等权重线性模型选股策略

为了验证 Value 因子和 Index 因子在量化选股策略中的有效性，本小节使用等权重线性选股策略测试其在历史回测中的效果。

多元线性回归模型可以表示为：

$$y = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_p x_p$$

其中 $x_1, x_2, \ldots, x_p$ 是样本的 p 个特征，y 是样本的标签，而 $\mathbf{w} = (w_0, \ldots, w_p)$ 是需要拟合的系数向量。

从多因子选股的角度来看，$x_1, x_2, \ldots, x_p$ 可以视作截面期的 p 个因子暴露度，y 是下期收益， $\mathbf{w}$ 反映了不同因子对收益的影响方向和程度。等权重线性回归模型中取 $\mathbf{w} = (1, \ldots, 1)$，由此我们得到如下针对合成因子 Value、Index 的等权重线性模型选股策略（模型 2）。

模型（2）： Value、Index 等权重线性模型

测试样本范围：全市场

测试样本期：2016 年 1 月 1 日至 2018 年 9 月 30 日

模型建立要点

1. 初始资金为 1000 万元整，手续费为双边千分之三，每月月初调仓。
2. 在每个截面期的最后一个交易日（本文中的截面期为一个月），提取样本内股票因子值，并剔除因子值缺失的股票。
3. 通过股票因子值计算出合成因子 Value 和 Index 值，并将合成因子值按照权重 $\mathbf{w}$ 加总作为该股票的综合得分。
4. 按照综合得分值大小将样本内股票排序，根据得分值最大的前 N 支（本文中 N 为全市场股票中的百分之三）股票作为备选股票池。
5. 在下一个截面期的首个交易日，以当天的收盘价将持仓股票更换为备选股票池中股票（根据不同股票价格登权重配置资金）并剔除当天因停牌、涨停等因素不能交易的股票。
6. 最后对 N 支股票的历史收益率进行回测，计算其年化收益率、最大回撤。

通过测试得到等权重线性模型回测结果，如下图 2 所示：

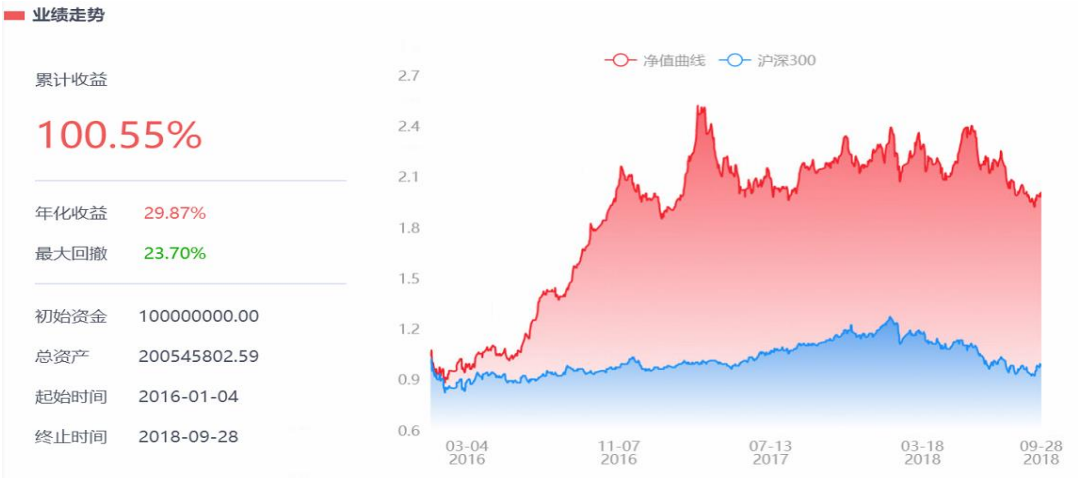


图 2 等权重线性模型净值曲线图

从图 2 中可以明确看到，本文所构建的主因子等权重模型净值曲线远胜沪深 300 收益率曲线，其收益率远远超过市场平均水准，同时策略详细绩效指标如下表 6 所示：

表 6 等权重线性模型净值回测绩效指标表

年化收益率	阿尔法比率	贝塔比率	夏普比率	信息比率	最大回撤
29.8%	0.30	0.81	1.21	1.80	23.70%

该模型的夏普比率达到了 1.21，同时其年化收益率达到了 29.8%，远远超过前面所测试的任何单因子选股模型，并且最大回撤只有 23.70%。由此可见，本文所构造合成的两个主因子在选股层面效果极其显著。

2 机器学习算法增强因子选股模型

本章针对问题二，利用三种机器学习算法对第一章得到的多因子选股模型进行提升，并通过提升模型的回测绩效分析对提升效果进行了比对。进而以提升效果最为优异的算法模型为例，提出了对于机器学习算法在量化投资策略中过拟合现象的处理方法，并以此为基础构建了一个基于机器学习算法优化后的最优多因子选股模型。

2.1 利用随机森林算法提升多因子选股模型

随机森林（Random Forest）作为一种比较新的机器学习方法，近年来在业内的关注度与受欢迎程度得到逐步提升。它是由多个弱学习器（决策树）组成，在运算量没有显著增加的前提下提高了预测精度，同时其运算结果对缺失数据和非平衡的数据也能达到相当稳健的水平。因此本文先将“随机森林”算法应用于多因子选股模型提升上并探究其表现。

2.1.1 “随机森林”基本原理[10]

随机森林使用 Bagging（并行）方式建立一个相互没有关联的决策树森林，这之中的每棵决策树都是基于多个特征进行分类决策，在树的每个结点处，根据特征的表现通过某种规则分裂出下一层的子节点，终端的子节点即为最终的分类结果。由于本文主要是对股票进行分类，所以当新的输入样本进入时，让森林中的每一颗决策树判断其类别，哪一类被选择最多就将样本预测为那一类。

2.1.2 随机森林基本构建步骤

随机森林算法主要采用 Bootstrap 随机采样思想，从训练集中有放回的采集固定个数的样本产生每个训练子集。之后算法为每一个训练子集分别建立一棵决策树，最终生成 N 棵决策树从而形成“森林”，每棵决策树自由生长不做剪枝处理，建造每棵决策树包括以下两个关键过程：

a) 节点分裂

节点分裂方式是算法的核心，在节点分裂时，将每个属性的所有划分按照某种规则进行排序，然后按照规则选择某个属性作为分裂属性，并按照其划分实现决策树的分支生长。本文使用确保信息增益最大原则（CART 算法）来形成每棵树的分支。

b） 输入变量的随机选取：

指随机森林算法在生成的过程中，为使每棵决策树之间的相关性减少，并提升分类精度从而提升整个算法的性能，其从全部 M 个特征中随机选择 m 特征（m 小于 M）为新的训练集，训练一棵决策树。对于本文所用分类预测来说，N 棵决策树投出最多票数的类别或者类别之一为最终类别。

2.1.3 随机森林算法提升多因子选股模型测试

将随机森林算法用于前文所提模型 2，并探究其表现：

模型（3）：基于随机森林算法的多因子选股模型

测试范围：全市场

测试区间：2016 年 1 月 1 日至 2018 年 9 月 30 日

模型建立步骤：

1. 提取特征值和预处理：每个自然月的最后一个交易日，计算每只股票之前报告里的合成因子（Index、Value）暴露度作为特征值，同时对第一章中矩阵 A 进行预处理。
2. 参数确定[5]：将参数设置为森林中决策树的个数 $\mathrm{n}=500$，随机选择的特征个数 $m=2$。
3. 机器学习算法预测：

a) 以 T 月月末截面期所有样本（即个股）预处理后的特征作为模型的输入，得到每个样本的 $T+1$ 月的预测值 $f(x)$（判别函数值，即样本到分类超平面的距离）。

b) 将随机森林模型的预测值视作单因子，因子值为空的股票不参与测试。

c) 在每个自然月最后一个交易日按预测值排序，选取预测上涨概率最大的 30 支股票在下个自然月首个交易日按当日收盘价等资金配置换仓。

4. 最后对 N 支股票的历史收益率进行回测，计算其年化收益率、最大回撤、

夏普比率等值。

通过测试得到基于随机森林算法的多因子选股模型回测结果，如下图 3 所示：

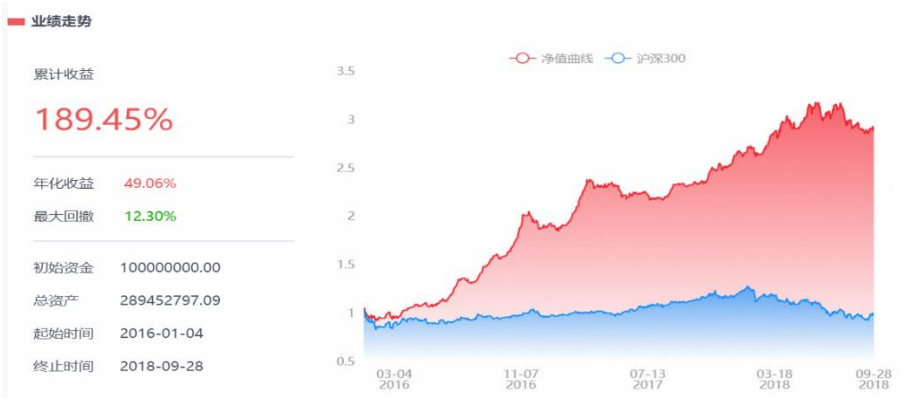


图 3 基于随机森林算法的多因子选股模型净值曲线图

从图中可以明确看到模型 3 的净值曲线远胜于模型 2 的曲线，其收益率远远超过市场平均水准，同时策略详细绩效指标如下表 7：

表 7 基于随机森林算法的多因子选股模型回测绩效表

年化收益率	阿尔法比率	贝塔比率	夏普比率	信息比率	最大回撤
49.06%	0.41	0.49	3.29	3.56	12.30%

跟初始等权重线性模型绩效相比，随机森林算法将其年化收益从 29.8%提升到 49.06%，于此同时还大幅降低了最大回撤，算法模型夏普比率达到了 3.29，说明随机森林算法对于提升多因子表现的效果非常显著。

2.2 利用 AdaBoost 提升算法提升多因子选股模型

提升算法的主要目的是将预测能力有限的单个弱学习器（例如决策树）组合成一个集成强学习器。依据弱学习器的组合方式，可将集成学习算法分为两大种类，一类为 Bagging 系列（并行方法），一类为 Boosting 系列（串行方法）。对于多棵决策树，如果以 Bagging 的方式组合起来，可以得到上文中随机森林算法；如果以 Boosting 的方式组合起来，就得到了接下来的 AdaBoost 算法。

2.2.1AdaBoost 算法基本原理

对于使用 Boosting 方式的集成算法有两个关键问题：1. 如何修正串行中每一轮训练数据的权值；2. 如何将一组弱学习器组合成强学习器。

首先 AdaBoost 的做法[①]是提高那些被前一轮弱分类器错误分类样本的权值，并降低那些被正确分类样本的权值。使得那些没有得到正确分类的数据，由于其权值的加大而将受到后一轮弱分类器的更大关注。于是，分类问题被一系列的弱分类器逐一解决。

其次 AdaBoost 采取加权多数表决的方法组合弱分类器。具体来说，加大分类误差率小的弱分类器权值，使其在表决中起较大的作用，减小分类误差率大的弱分类器的权值，使其在表决中起较小的作用。

对于选股来说主要针对股票下一期上涨或下跌的概率进行分类，主要涉及 AdaBoost 二元分类算法，接下来主要对此算法进行介绍。

2.2.2 AdaBoost 二元分类算法基本步骤

假设输入为样本集 $T=\{(x_1,y_1),(x_2，\ y_2),\dots,(x_m,y_m)\}$，输出为 $\{-1,+1\}$，弱分类器迭代次数 K，输出是强分类器 $f(x)$。

算法主要有以下三步：

（一） 初始化样本集的权重为

$$D(1)=\ (w_{11},w_{12},\dots,w_{1m}),\qquad w_{1i}=\frac{1}{m},\quad i=1,2,\dots,K$$

（二） 对于 $k=1,2,\dots,K$：

（1）使用具有权重 $D(k)$ 的样本集来训练数据，得到弱分类器 $G_k(x_i)$。

（2）计算第 K 个分类器 $G_k(\mathrm{x})$ 的分类误差率

$$e_K=P(G_k(x_i)\neq y_i)$$

（3）计算弱分类器 $G_k(x)$ 的权重系数

$$\alpha_k=\mu\ln\frac{1-e_K}{e_K}$$

μ 为学习率超参数（默认为 1）[8]，预测器的准确率越高，其权重就越

[①] AdaBoost—CSDN 博客-https://blog.csdn.net/webzjuyujun/article/details/49888719

高。如果它只是随机猜测，则其权重接近于 0。

（4）更新样本集的权重分布

对于 $i = 1,2,\dots,m$

$$w^{(i)} \leftarrow \begin{cases} w^{(i)} & \left(\hat{y}_j^{(i)} = y^{(i)}\right) \\ w^{(i)} \exp(\alpha_j) & (\hat{y}_j^{(i)} \neq y^{(i)}) \end{cases}$$

然后将所有实例的权重归一化(即除以 $\sum_{i=1}^{m} w^{(i)}$)，使用更新后的权重训练一个新的预测器，然后重复整个过程，当达到所需要数量的预测器或得到完美的预测器时，算法停止。

（三）最后，构建强分类器：

$$f(x) = \text{sign}\left(\sum_{k=1}^{N} \alpha_k G_k(x)\right)$$

得到预测类别，其中 N 是分类器的数量。

2.2.3 AdaBoost 二元分类算法提升多因子选股模型

将 AdaBoost 二元分类算法用于模型 2 中，并探究其表现：

模型（4）：基于 AdaB oost 算法的多因子选股模型
测试范围：全市场
测试区间：2016 年 1 月 1 日至 2018 年 9 月 30 日
1. 提取特征值和预处理：每个自然月的最后一个交易日，计算每只股票之前报告里的合成因子（Index、Value）暴露度作为特征值，同时取第一章中矩阵 A 对其进行预处理。
2. 参数确定[4]：将参数设置为弱学习器权重缩减系数 $v = 0.15$，最大弱学习器个数 $N = 100$。
3. 机器学习算法预测：
a) 以 T 月月末截面期所有样本（即个股）预处理后的特征作为模型的输入，得到每个样本的 $T+1$ 月的预测值 $f(x)$（判别函数值，即样本到分类超平面的距离）。
b) 将 AdaBoost 模型的预测值视作单因子，因子值为空的股票不参与测试。

c) 在每个自然月最后一个交易日按预测值排序，选取预测上涨概率最大的 30 支股票在下个自然月首个交易日按当日收盘价等资金配置换仓。

4. 最后对 N 支股票的历史收益率进行回测，计算其年化收益率、最大回撤、夏普比率等值。

通过测试得到模型 4 的回测结果，如下图 3 所示：

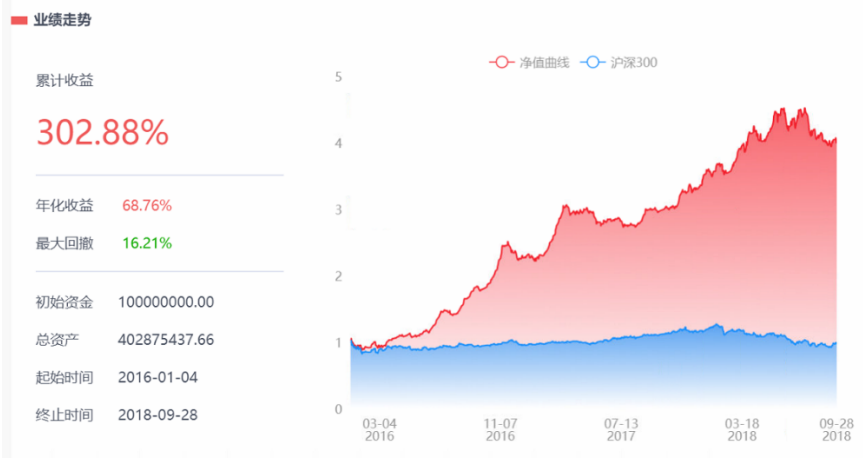


图 3 基于 AdaBoost 算法的多因子选股模型净值曲线图

从图中也可以明确看到模型 4 的净值曲线远胜模型 2 的曲线，其收益率远远超过市场平均水准，同时策略详细绩效指标如下表 8 所示：

表 8 基于 AdaBoost 算法的多因子选股模型回测绩效指标表

年化收益率	阿尔法比率	贝塔比率	夏普比率	信息比率	最大回撤
68.76%	0.54	0.63	3.64	4.50	16.21%

跟初始等权重线性模型绩效相比，AdaBoost 算法也在大幅提升年化收益的同时降低了最大回撤，跟基于随机森林算法的模型 3 相比，AdaBoost 算法收益率更高但是其最大回撤高于模型 3，综合来说模型 4 的夏普比略高于随机森林算法，其在提升因子表现的能力上略优于随机森林算法。

2.3 利用支持向量机提升因子表现

在 20 世纪 90 年代，支持向量机（Support Vector Machine，SVM）由于其极高的预测正确率，并且能够解决非线性分类问题，成为当时最流行的机器学习方法。SVM[9]在样本空间线性可分的情况下直接寻找到最优分界平面；在线性不可分的情况下，通过核函数将输入空间进行非线性变化至高维特征空间，然后在新

的空间中寻找最优分界平面。由此 SVM 可分为线性支持向量机和核支持向量机，前者针对线性分类问题，后者属于非线性分类器。

2.3.1 线性支持向量机算法原理

1. 最大间隔分类

在二维平面中，分类边界是一条一维直线；在三维空间中，分类边界是一个二维平面；在 N 维度的空间中，分类边界是一个 $N-1$ 维空间。我们将上述用于分类的一维直线、二维平面和$N-1$维空间统称为分类超平面。线性支持向量机通过最大间隔分类来确定最优的分类超平面。

我们对于给定的训练数据集 T、超平面 (w,b) 以及 n 个样本点 (x_i,y_i) 得到下面线性可分支持向量机的优化问题[2]：

$$\min \quad \frac{1}{2}\|w\|^2$$

$$\text{s.t} \quad y_i((\boldsymbol{w})^T x_i + b) - 1 \geq 0 \ , i = 1,2,\dots,n$$

2. 松弛变量

在绝大多数情况下，数据中包含的噪音使得难以用一条直线将两类样本完美区分，此时目标函数约束条件无法满足。为了应对线性不可分情况，本文引入松弛变量的概念。对每个样本点赋予一个松弛变量的值[9]：如果该点落在最大边缘超平面正确的一侧，则松弛变量 $\xi = 0$；否则，松弛变量的值等于该点到最大边缘超平面的距离。

此时我们将线性支持向量机的目标函数改写为[2]：

$$\min_{w,b} \ \left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i\right)$$

$$\text{s.t.} \quad y_i\left((\boldsymbol{w})^T x^i + b\right) - 1 \geq \xi_i \ , i = 1,2,\dots,n$$

$$\xi_i \geq 0$$

其中 ξ_i 为第 i 个样本的松弛变量，n 为样本个数。

3．惩罚系数 C

引进松弛变量后，目标函数在原有的基础之上新加入一项 $C\sum_{i=1}^{n}\xi_i$，即所有样本松弛变量之和乘以系数 C。该系数 C 称为惩罚系数，表示模型对错误分类的容忍度。当 C 取较大的数时，即使很小的松弛变量 ξ_i 也会造成很大的损失，

因此分类器对错误分类的容忍度较低，将尽可能保证分类正确，从而导致较高的训练集正确率[3]。反之，当 C 取较小的数时，分类器对错误分类的容忍度较高，允许错误分类的存在，分类器倾向于以最大间隔分类的原则进行分类。

2.3.2 核支持向量机

1.核函数概述

核支持向量机的核心思想正是将非线性分类转化为线性分类。首先通过非线性映射 ϕ 把原始数据 x 变换到 k 个高维特征空间：

$$x \mapsto \phi(x) = (\phi_1(x), \dots, \phi_k(x))$$

随后对于高维空间下的数据 $\phi(x)$，使用线性支持向量机进行分类，从而解决非线性分类问题。此时线性支持向量机对偶问题的目标函数为：

$$\max_{\alpha}\left[\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \phi(x_i)^T \phi(x_j)\right]$$

目标函数里的 $\phi(x_i)^T\phi(x_j)$ 实质上是两个向量的内积，记为 $\langle\phi(x_i), \phi(x_j)\rangle$。而任意一种映射方式 ϕ 的内积 $\langle\phi(x_i), \phi(x_j)\rangle$，可以用一个确定的核函数 $K(x_i, x_j)$ 加以刻画。

2.常用核函数[9]

理论上任何一种映射方式都对应一种确定的核函数，满足一定数学性质的函数都可以作为支持向量机的核函数，但在实际应用中，通常使用以下几种核函数：线性核、多项式核、Sigmoid 核、高斯核。本文采用三次多项式核：

$$K(x_i, x_j) = \left(\gamma\langle x_i, x_j\rangle + 1\right)^d = \left(\gamma\sum_{k=1}^{p} x_i^{(k)} x_j^{(k)} + 1\right)^d$$

3.分布系数 γ 值

γ 值为核支持向量机的重要参数，多项式核、Sigmoid 核和高斯核的核函数都包含 γ 值一项。γ 决定了原始数据映射到高维数据后，在高维特征空间中的分布。

2.3.2 支持向量机算法提升多因子选股模型

将三次多项式核 SVM 算法用于模型 2 中，并探究其表现：

模型（5）：基于支持向量机算法的多因子选股模型

测试范围：全市场

测试区间：2016 年 1 月 1 日至 2018 年 9 月 30 日

1. 提取特征值和预处理：每个自然月的最后一个交易日，计算每只股票之前报告里的合成因子（Index、Value）暴露度作为特征值，同时取第一章中矩阵 A 对其进行预处理。
2. 算法选择与参数确定[3]：通过比较使用不同的核函数的运算速度和效果，本文 SVM 核函数选取 3 阶多项式核。将参数先确定为 $C = 0.1, \gamma = 0.01$。
3. 机器学习算法预测：

a) 以 T 月月末截面期所有样本（即个股）预处理后的特征作为模型的输入，得到每个样本的 $T+1$ 月的预测值 $f(x)$（判别函数值，即样本到分类超平面的距离）。

b) 将 SVM 模型的预测值视作单因子，因子值为空的股票不参与测试。

c) 在每个自然月最后一个交易日按预测值排序，选取预测上涨概率最大的 30 支股票在下个自然月首个交易日按当日收盘价等资金配置换仓。

4. 最后对 N 支股票的历史收益率进行回测，计算其年化收益率、最大回撤、夏普比率等绩效指标值。

通过测试得到基于支持向量机算法的多因子选股模型回测结果，如下图 5 所示：

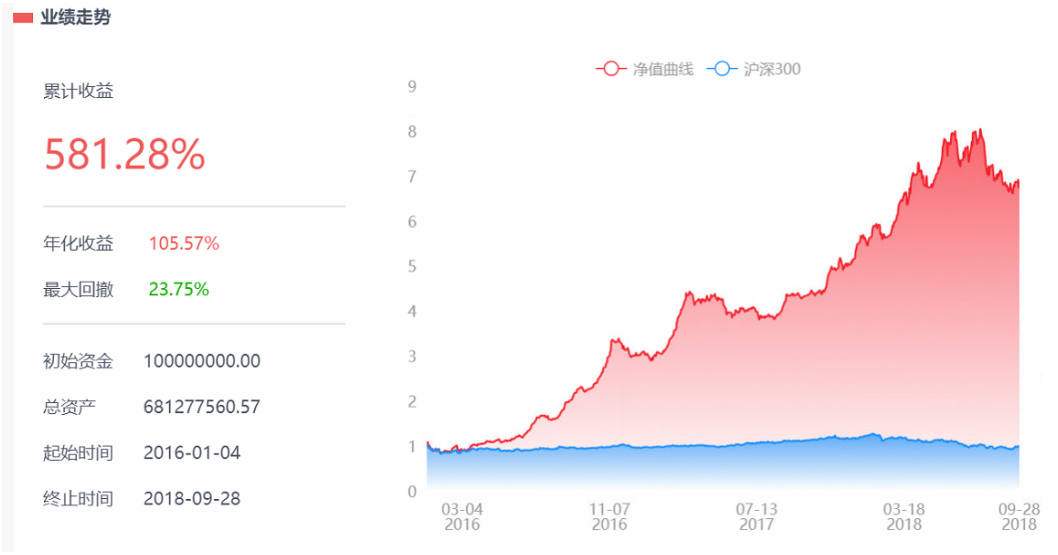


图 5 基于支持向量机算法的多因子选股模型净值曲线图

从图中可以明显看到本文所构建的基于支持向量机算法的多因子选股模型净值曲线远胜模型 2 的曲线，其收益率远远超过市场平均水平，同时策略详细绩效指标如下表 9 所示：

表 9 基于支持向量机算法的多因子选股模型回测绩效图

年化收益率	阿尔法比率	贝塔比率	夏普比率	信息比率	最大回撤
105.57%	0.76	0.88	4.07	5.42	23.75%

跟初始等权重线性模型绩效相比，支持向量机算法将年化收益提高了近 5 倍，同时回撤也并没有提高很大，与基于随机森林和 AdaBoost 算法的多因子模型相比，其收益率最高，但是回撤也较大，综合各指标来看，模型 5 的夏普比优于其他两种算法所以本文将支持向量机定为提升多因子模型的最优算法。

2.4 对最优机器学习提升模型的过拟合分析与处理

通过对三种机器学习算法的回测结果分析，本文得到了对多因子模型提升最高的 SVM 3 阶多项式核函数算法，下文将以此算法为例，给出消除机器学习算法的过拟合现象的处理方法，其他算法的处理也大致类似。

2.4.1 过拟合问题的原因

1.数据原因

绝大多数机器学习算法都需要依赖大量的数据，即使最简单的问题，很可能需要成千上万的示例。如果训练集本身是充满噪音的，或者数据集太小（会导致采样噪音），那么很可能会导致算法模型检测噪声里的模式，显然这样就无法泛化至新的实例，也就是过拟合。Michele Banko 和 Eric Brill [6]指出，截然不同的机器学习算法（包括相当简单的算法）在自然语言歧义消除这个复杂的问题上，表现几乎完全一致，导致准确率差异的根本原因就是数据量的区别，训练的数据量越大，得到的准确率越高。数据比算法更加重要这一思想由 Peter Norving 等人[7]进一步的推广。所以本文在进行每一个机器学习算法前都花费大量时间通过一系列数学方法清理训练数据，尽量规避数据层面带来的过拟合问题。

2.超参数原因

超参数（hyperparameter）[①]定义为：机器学习算法模型的外部变量，是使用者用来确定模型的参数。例如前文涉及的支持向量机模型中核函数类型、惩罚系数C、分布系数 γ，随机森林模型的树棵数、最大特征数，AdaBoost 模型的弱学习器权重缩减系数 v、最大弱学习器个数 N 等，这些都属于模型的超参数。而惯常的参数（parameter）是模型的内部变量，是模型通过学习可以确定的参数。机器学习算法模型中超参数的选择对模型是否过拟合也会产生极大影响。

本文接下来主要针对超参数引起的过拟合问题进行探讨研究。

2.4.2 过拟合问题的应对

避免过拟合的重要方法之一是进行交叉验证（cross-validation）。英国统计学家 Mervyn Stone 和美国统计学家 Seymour Geisser 是交叉验证理论的先驱。交叉验证理论并非仅针对机器学习模型，而是针对任何统计模型。

交叉验证的核心思想是先将全部样本划分成两部分，一部分用来训练模型，称为训练集；另外一部分用来验证模型，称为验证集。随后考察模型在训练集和验证集的表现是否接近。如果两者接近，说明模型具备较好的预测性能；如果训

[①] 机器学习填坑：模型参数和超参数之间的区别—CSDN 博客—https://blog.csdn.net/xjp_xujiping/article/details/88556953

练集的表现远优于验证集，说明模型存在过拟合的风险。本文对不同超参数设置下的多个模型进行比较，考察模型在验证集的表现，选择验证集表现最优的超参数作为最终模型的超参数。

2.4.3 交叉验证方法

交叉验证方法主要根据如何划分训练集与验证集来定义：

1. 简单交叉验证——留出法（Hold-out）

从总样本中随机选取一定比例（如 15%）的样本作为验证集。其只需要训练一次模型，效率较高，但是验证集数据从未参与训练，可能削弱模型的准确性，在极端情况下，当验证集中数据本身就是整体数据的“噪点”时，模型的准确度将会大大降低；同时最终的模型评价结果可能还会受到训练集和验证集划分过程中的随机因素干扰。

2. K 折交叉验证

针对上述简单交叉验证的缺陷，本文使用 K 折交互验证（K-fold cross-validation）的方法，随机将全体样本分为 K 个部分（K 在 3~20 之间），轮流将其中 $K-1$ 份作为训练集，1 份作为验证集进行测试，每次测试都得到验证集相应的评价指标数据，最终将得到 K 个验证集的评价指标，取其均值作为验证集的平均表现。

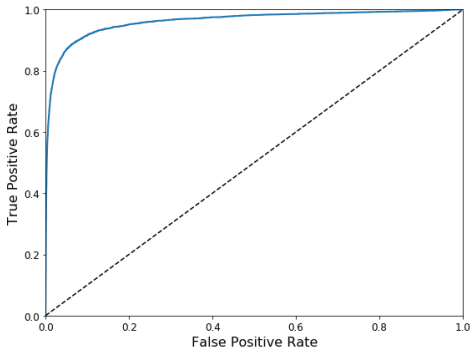
同时为了减少因验证集划分不同而引入的差别，K 折交叉验证通常要随机使用不同划分重复 P 次，最终的评估结果是这 P 次 K 折交叉验证结果的均值，本文使用的就为 “10 次 10 折交叉验证”。

2.4.4 机器学习算法模型过拟合评价标准

衡量模型好坏的常用指标有正确率、召回率、精确率、虚报率和特异度等。但在本文所使用的多因子选股模型中，机器学习算法模型的目的不仅仅是对股票进行正确分类，更多时候是希望选择预测值最高，即上涨可能性最大的一小部分股票进行投资。正确率、召回率、虚报率等并不是稳定的评价指标，所以本文在这里引入接受者操作特征曲线（Receiver Operating Characteristic Curve，ROC 曲线）这一概念对算法模型进行评估[11]。ROC 的思想是通过遍历所有分类阈值（也

就是前文所提的超参数）计算出每一个阈值对应的虚报率和召回率，不断遍历后以虚报率为横轴、召回率为纵轴，将所有点顺次连接可以得到一条曲线，如图 6 所示：

图 6



将 ROC 曲线的形态特征总结为一个指标——ROC 曲线下覆盖的总面积（Area Under Curve，AUC 指标），AUC 指标值衡量标准如下[11]：

a） AUC = 1，是完美分类器，采用这个预测模型时，不管设定什么阈值都能得出完美预测，绝大多数预测的场合，不存在完美分类器。

b） 0.5＜AUC＜1，优于随机猜测。这个分类器（模型）妥善设定阈值的话，能有预测价值。

c） AUC = 0.5，跟随机猜测一样（例：丢铜板），模型没有预测价值。

d） AUC＜0.5，比随机猜测还差；但只要总是反预测而行，就优于随机猜测。

在后续测试中，本文将主要使用 AUC 作为确定超参数的主要依据。

2.4.3 以基于 SVM 算法的多因子选股模型为例的过拟合处理方案

1. SVM 算法中的超参数：惩罚系数 C 和分布系数 γ

对 SVM 算法来说，如果惩罚系数 C 取值过大，分类器容易由于极端样本影响造成过拟合，表面上训练集正确率较高，但是实际上测试集正确率并不高，也就是说有较低的偏差（Bias）和较大的方差（Variance）；如果惩罚系数 C 取值过小，分类器会过于不在乎分类错误，训练集和测试集正确率都将受损，导致较低的偏差和方差。而分布系数 γ 越大则导致样本在高维空间中的分布越稀疏，样本之间更容易被分类边界区分开，所以模型训练集的正确率更高，但是也更容易导致过拟合。实际应用中，过小和过大的 γ 都会使得分类性能受损。

因此，本文将对惩罚系数 C 与分布系数 γ 进行遍历，选择使得交叉验证集 AUC 最高的惩罚系数 C 与分布系数 γ 作为模型最终的参数，完成对 SVM 模型过拟合的分析和研究。

2. 对 SVM 提升多因子股票模型的过拟合处理

1. 数据获取：

a）股票池：全 A 股，每只股票视作一个样本。

b） 样本内区间：2016 年 1 月 1 日至 2018 年 9 月 30 日共 21 个月末截面期。

c） 样本外区间：2018 年 9 月 30 日至 2019 年 3 月 28 日共 6 个月末截面期。

2. 特征和标签提取[3]：每个自然月的最后一个交易日，计算前文中的合成因子（Value，index）暴露度，作为样本的原始特征；计算下一整个自然月的个股收益，作为样本的标签。

3．特征预处理：使用第一章中对矩阵 A 进行预处理的步骤。

4．训练集和交叉验证集的合成[3]：

对于支持向量机模型（以下简称 SVM），在每个月末截面期，选取下月收益排前、后 30 名的股票分别作为正例（$y = 1$）、负例（$y = -1$）。将 21 个月样本合并，随机选取 90%的样本作为训练集，余下 10%样本作为交叉验证集。

5．样本内训练：使用 3 阶多项式核 SVM 对训练集进行训练。同时以本文模型 2 作为统一对照组。

6．交叉验证调参：模型训练完成后，使用该模型对交叉验证集进行预测。本文同时对 C 值和 γ 值进行遍历，选取交叉验证集 AUC 最高（SVM）的一组参数作为模型的最优参数。

使用网格搜索法进行参数寻优：

a) 对惩罚系数 C 和分布系数 γ 的取值按幂次方探索，取

$$\gamma = [e^{-4}, e^{-3}, e^{-2}, e^{-1}, 1], \quad C = [e^{-4}, e^{-3}, e^{-2}, e^{-1}, 1]$$

b) 得到如图 7 所示三阶多项式核 SVM 算法模型网格搜索交叉验证集

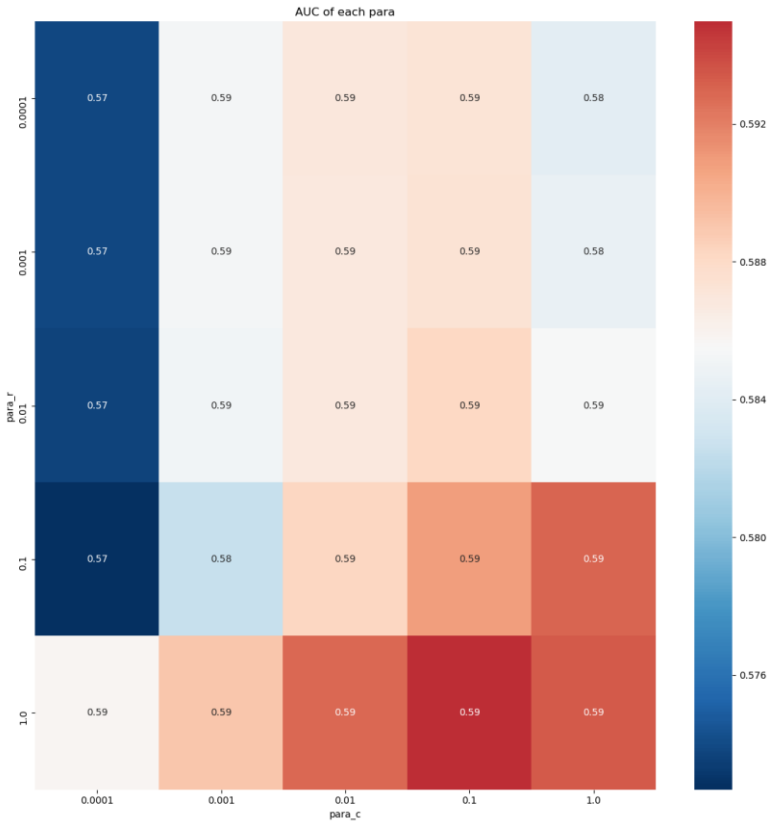


图 7 三阶多项式核 SVM 算法模型网格搜索交叉验证集

可以发现惩罚系数 $C = 0.1$，分布系数 $\gamma = 1$ 时，AUC 最高达到 0.59，分类器的效果较优秀。

7. 样本外测试：确定最优参数后，以 T 月月末截面期所有样本（即个股）预处理后的特征作为模型的输入，得到每个样本的 $T+1$ 月的预测值 $f(x)$（判别函数值，即样本到分类超平面的距离），根据该预测值构建策略组合：

a) 股票池：全 A 股。

b) 回测区间：2018-09-30 至 2019-03-28。

c) 数据处理方法：将支持向量机模型的预测值视作单因子，因子值为空的股票不参与测试。

d) 换仓期：在每个自然月最后一个交易日核算因子值，在下个自然月首个交易日按当日收盘价等资金配置换仓。

8. 过拟合处理后的 SVM 提升模型样本外效果

对过拟合处理后的 SVM-CO 提升模型进行样本外回测测试，测试结果如下图 8 所示：

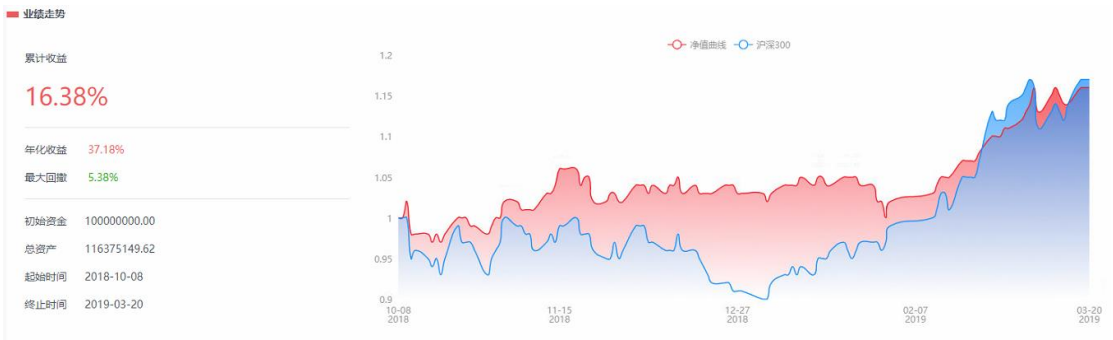


图 8 SVM-CO 提升模型样本外回测净值曲线图

同时策略详细绩效指标如下表 11:

表 11 SVM-CO 提升模型样本外绩效指标表

年化收益率	阿尔法比率	贝塔比率	夏普比率	信息比率	最大回撤
37.18%	0.20	0.45	2.53	0.36	5.38%

从回测结果中可以看到，SVM-CO 提升模型在样本外回测区间，半年间累积收益为 16.38%，并且在达到了年化 37.18%的收益率的同时，只需要付出 5.38%的回撤，模型基准曲线已经跑赢了基准沪深 300 指数收益率曲线，同时夏普比率达到 2.53。

9. 过拟合处理后的 SVM-CO 提升模型样本内效果：

对过拟合处理后的 SVM-CO 提升模型进行样本内回测测试，测试结果如下图 9 所示：

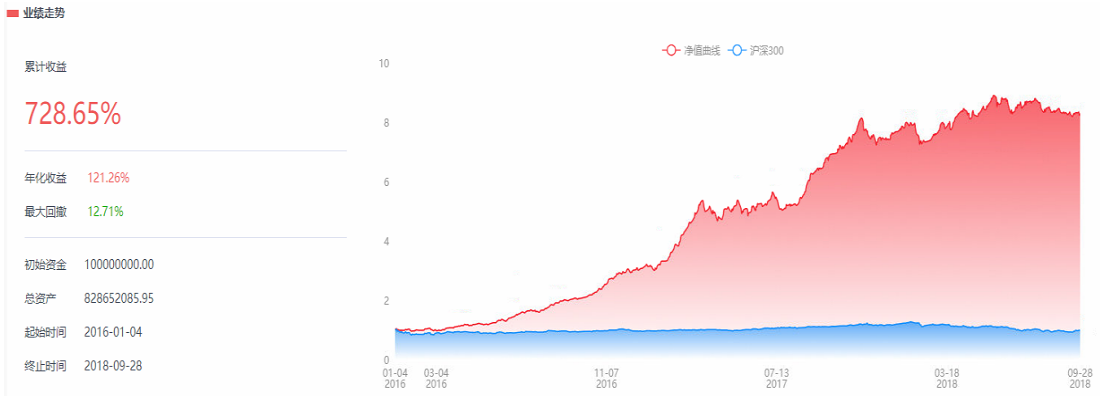


图 9 SVM-CO 提升模型样本内回测净值曲线图

同时策略详细绩效指标如下表:

表 12 SVM-CO 提升模型样本内绩效指标表

年化收益率	阿尔法比率	贝塔比率	夏普比率	信息比率	最大回撤
121.26%	0.82	0.56	5.76	6.29	12.71%

可以看到样本区间内模型净值曲线相比过拟合处理前的算法模型更为平稳，跟过拟合处理之前的模型 5 相比，过拟合处理后的算法模型提供了更高的收益率并降低了最大回撤，夏普比率提高到了 5.76，说明过拟合处理能够提高机器学习算法模型的表现。

3 风险控制

本章针对问题三，利用三种风险对冲方法对第二章得到的 SVM-CO 提升模型进行风险控制，进而建立了新型的风险对冲模型，即 SVM-RC 多因子选股模型。该模型具有高收益、低风险的特点，同时其最大回撤被控制在 10%以内。

3.1 行业、风格中性组合

在多因子模型中，决定策略收益稳健性的另一关键步骤在于每一期股票组合的权重配置。因此，从量化对冲策略追求收益稳定性的角度而言，组合权重优化对多因子模型起着至关重要的作用。本节构建了基于点宽数据库中 29 类行业因子、10 类风格因子的结构化多因子风险模型，通过股票组合的权重优化设计，找到在行业中性和风格因子中性约束下的最优投资组合，对冲基准为沪深 300 指数。

3.1.1 行业中性对冲

行业中性对冲是指，将算法模型中每个截面期所选取的股票投资组合的行业权重与对冲基准的行业权重调整一致，以去除行业因子对策略收益波动的影响，确保策略收益仅与行业内部个股的超额收益有关。

本文对冲基准沪深 300 指数的行业权重分布如下图 10 所示：

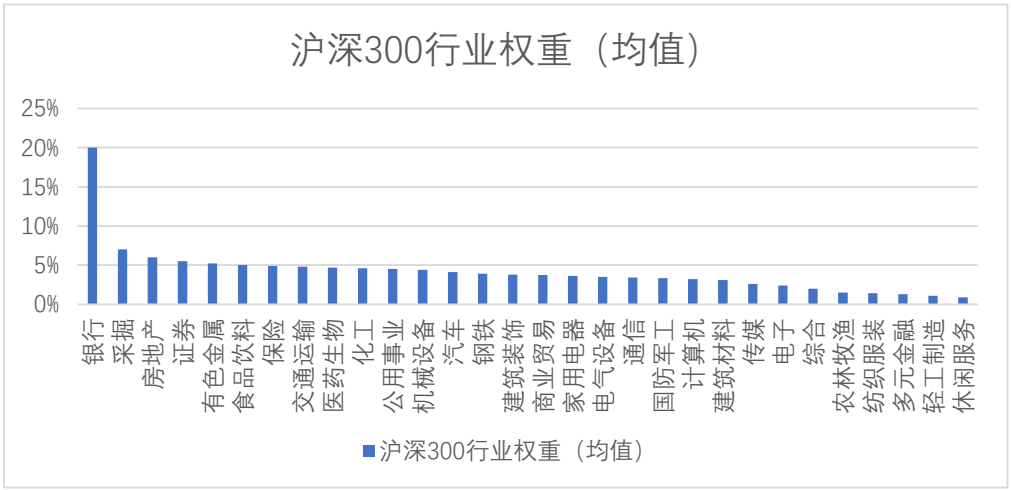


图 10 沪深 300 指数的行业权重分布图[①]

沪深 300 指数中，权重排名前 5 的行业为：银行、采掘、房地产、证券 和

[①] 数据来源：国泰君安证券研究院。

有色金属，权重排名后 5 的行业为：休闲服务、轻工制造、多元金融、 纺织服装和农林牧渔。

假设 H 为股票组合的行业因子哑变量矩阵，h 为沪深 300 指数中各行业的对应权重，那么行业中性的权重 w 满足：

$$w^T H = h^T, \qquad w \geq 0$$

根据公式即可构造每个截面期具有行业中性的股票投资组合。

3.1.2 风格中性

风格中性对冲是指，股票组合的风格因子较之对冲基准的风险暴露为 0。将股票组合的风格特征完全与对冲基准相匹配，使得组合的超额收益不来自于某种风格，从而使投资组合追求稳定的阿尔法收益，而并非市场某种风格的收益，经风格因子中性配置后，组合稳定性将进一步提升。综合来说，风格中性的目的是为了将多头组合的某一风格特点尽量逼近对冲基准，进而规避该风格可能对策略收益产生的波动。

对冲基准沪深 300 指数风险因子敞口如下图 11 所示：

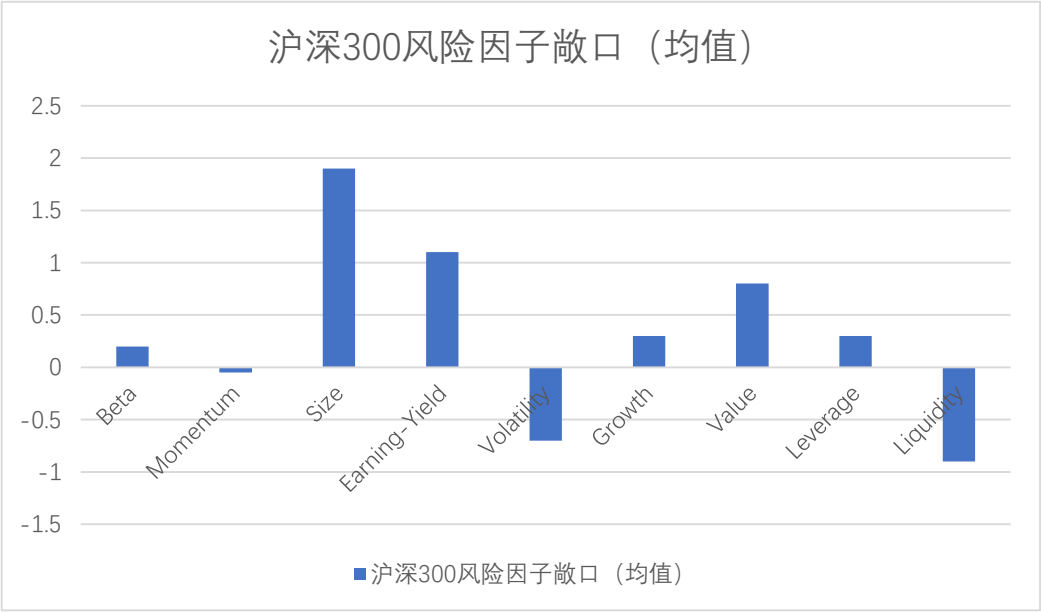


图 11 沪深 300 指数风险因子敞口图[①]

假设 x_k 为股票池第 k 个风格因子的载荷截面，w_{bench} 为沪深 300 指数对应权重，那么因子 k 的风格中性权重 w 满足：

[①] 数据来源：国泰君安证券研究院。

$$(w^T - w_{bench}^T)X_k = 0, \qquad x \geq 0$$

若权重 w 对每个截面期构建的股票组合中任意风格因子满足上述表达式，则称股票组合 w 满足风格因子中性。

3.2 波动率控制风险

在最常见的均值-方差分析中，市场组合的表现与其波动率息息相关。使用月度波动率动态调整市场组合，相比简单的被动持有市场组合，可以获得显著的风险调整后超额收益。所以本文在风险、行业中性的基础上加入了波动率管理，以达到有效对冲市场风险的目的。

3.2.1 波动率模型构建

在经过行业中性和风格中性处理后的股票池模型的基础上，本文统计基准指数（沪深 300 指数）的波动率（收益率标准差），滚动计算过去一段时间（2 个月）的波动率均值作为一个阈值，如果近一周基准指数波动率超过这个阈值就主动降低股票池仓位至原先的 50%。本文通过波动率择时管理投资组合，进一步提升模型的收益风险比。

3.3 模型实证检验

在前文 SVM-CO 提升模型基础之上，利用三种风险控制模型，建立了 SVM-RC 多因子选股模型：

模型（6）：SVM-RC 多因子选股模型
测试样本范围：全市场
测试样本期：2016 年 1 月 1 日至 2018 年 9 月 30 日
模型建立要点：
1. 初始资金为 1000 万元整，手续费为双边千分之三，每月月初调仓
2. 在 SVM-CO 提升模型的基础上进行投资组合处理：
a) 风险、行业中性处理方法：在 SVM-CO 提升模型的基础上，通过 BP 数据库中因子暴露数据得到股票针对 9 个风格因子和 30 个行业因子的暴露值

经过循环遍历求得每个月具有行业中性和风格中性的股票组合池。

b) 波动率控制：即时滚动计算基准指数波动率，当其近一周基准指数波动率超过过去 2 个月的波动率均值，将股票池持仓降低至 50%。

3.4 实证结果

本小节主要对本文最终得到的模型（6）进行回测分析，详细描述了模型优异的市场表现。

3.4.1 模型样本期内回测结果

通过回测 SVM-RC 多因子选股模型在样本期间（2016 年 1 月 1 日至 2018 年 9 月 30 日）内表现，得到业绩表现如表 13 所示。

表 13 SVM-RC 多因子选股模型回测绩效指标表

累计收益率	年化收益率	阿尔法比率	贝塔比率	夏普比率	信息比率	最大回撤
470.49%	92.32%	0.65	0.25	8.73	5.83	4.69%

模型在 2 年多的样本期间内获得了高达 470.49%的累积收益率，年化收益为 92.32%，远超同时间内沪深 300 指数的业绩表现。同时模型最大回撤仅为 4.96%，夏普比达到了 8.73，完全体现了模型中对冲方案带来的效果。得到的模型收益率曲线及每日盈亏如下图 12 所示（红色线为模型累积收益率，蓝色线为沪深 300 基准收益率曲线）：

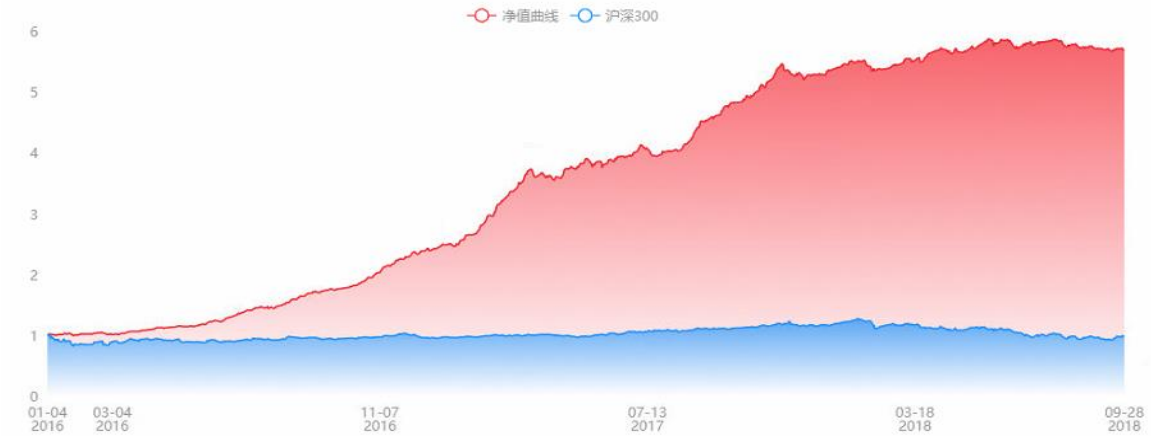


图 12　SVM-RC 多因子选股模型净值曲线图

该策略在获得大量收益的同时只付出微小的回撤代价，说明模型在样本期内

的效果极为显著，但是仅用样本期间的回测结果来说明模型在股票市场投资中的表现将不具有代表意义，所以接下来将对模型进行长跨度时期内的回测检验。

3.4.2 模型长跨度时期回测检验

为了对模型进行长时间跨度回测，本文选取自 2013 年 1 月 1 日起至 2019 年 3 月 1 日合计 6 年多时间内为回测期间，得到如下表 14 所示业绩表现：

表 14 SVM-RC 多因子选股模型长时间跨度回测指标表

累计收益率	年化收益率	阿尔法比率	贝塔比率	夏普比率	信息比率	最大回撤
1312.84%	56.22%	0.42	0.20	5.65	2.37	9.88%

模型在 6 年多的回测期间内获得了高达 1312.84%的累积收益率，年化收益为 56.22%，远超同期沪深 300 指数的业绩表现。同时最大回撤仅为 9.88%，反映了模型对冲后带来的低风险效果，其夏普比也达到了 5.65，取得极高的收益风险比。得到的模型收益率曲线及每日盈亏如下图 13 所示（红色线为模型累积收益率，蓝色线为沪深 300 基准收益率曲线）：

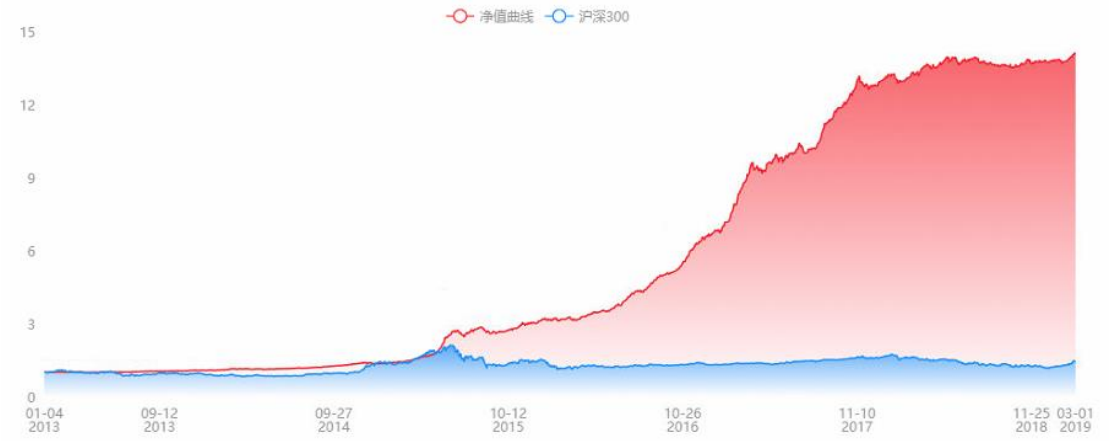


图 13 SVM-RC 多因子选股模型长时间跨度回测净值曲线图

3.4.3 分行情实证分析

本节将对 SVM-RC 多因子选股模型进行更全面的验证分析，以说明其在不同行情下都具有稳定的收益表现。从上证综合指数 2011-2019 年的走势图（详见图 14）可以看出，从 2014 年至今，指数走势经历了从大涨到大跌、然后有一个缓慢盘整的过程。具体分析来看：2014 年到 2015 年上半年股票市场处于明显的

牛市行情，指数从 2000 点一路上扬至 5300 点左右。但从 2015 年 6 月开始，股票市场发生股灾，短短大半年内指数从 5300 点左右的峰值大跌至 2700 点左右，亏损几乎达到 50%。到 2016 年 5 月以后，股票市场开始有了一些缓慢上升，但是期间又发生了较为明显的跌幅，导致指数仍然在 3000 点左右震荡。由此本文可以将 2014 年至今的股票行情时间段切割为 3 个不同走势的部分，即“牛市”、“熊市”、“震荡市”，其具体时间分段如下：

a) 牛市：2014 年 4 月 22 日至 2015 年 6 月 12 日。

b) 熊市：2015 年 6 月 12 日至 2016 年 5 月 2 日。

c) 震荡市：2016 年 5 月 2 日至 2019 年 4 月 12 日。

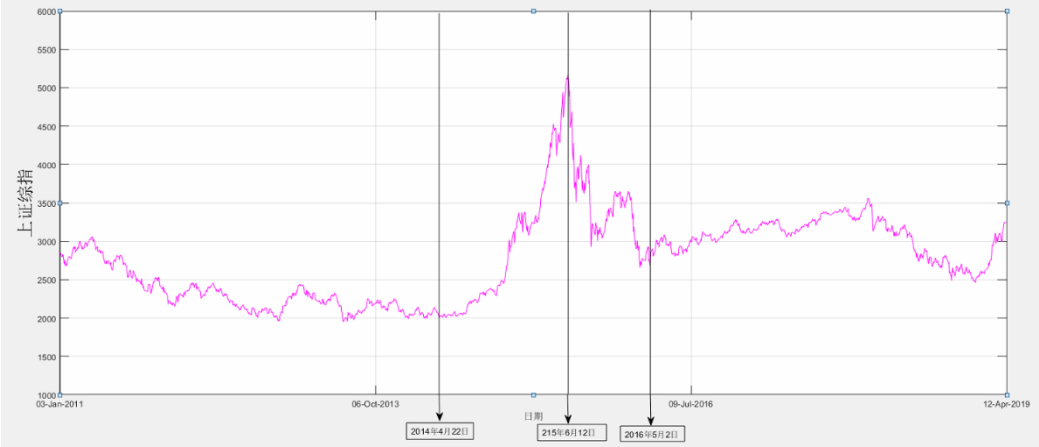


图 14 上证综合指数历史走势

将 SVM-RC 多因子选股模型分别进行这 3 个时间段的回测，收益情况详见表 15。通过观察对比发现，模型在牛市、熊市和震荡市的年化收益率分别为 133.49%、36.58%和 64.17%，其收益率表现都显著优于基准指数收益率（详见图 15、16、17）。其中模型在熊市中受行情影响，收益率较低、回撤较大，相对于基准收益率的提升，没有在其它两个行情中优异。但是在三种行情中，模型都不仅获得了超过市场平均收益水准的收益，而且同时完全体现了对冲方案对模型的风险管理效果，模型在每个行情的最大回撤都不超过 10%。

表 15 分行情实证表现对比

市场表现	累计收益率	夏普比率	年化收益率	最大回撤
牛市	133.49%	10.81	113.93%	3.49%
熊市	36.58%	3.89	42.67%	6.63%
震荡市	305.02%	7.61	64.17%	4.42%

图 15 牛市中收益走势

图 16 熊市中收益走势

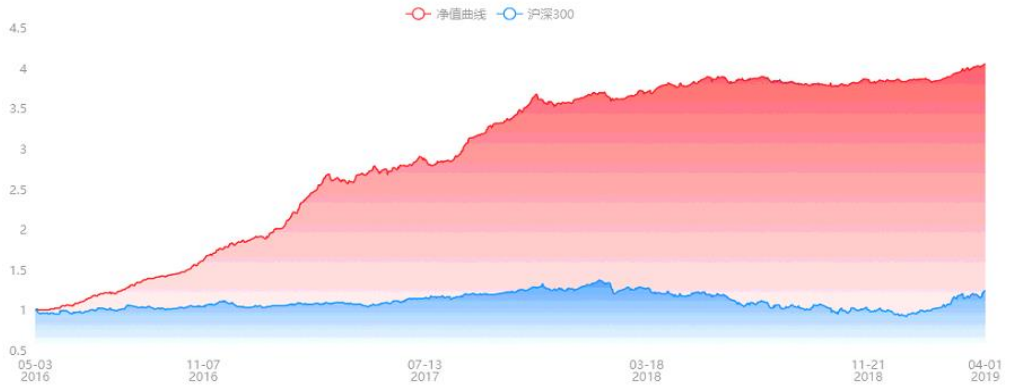


图 17 震荡市中收益走势

4 结语

本文首先对全市场 3000 多支股票和 BP 因子库中 500 多个因子，对 2016 年 1 月 1 日至 2018 年 9 月 30 日的数据进行完备化预处理，并通过 IC 显著性分析和单因子选股模型的绩效指标筛选出市场表现优异的六个因子：MktValue、NegMktValue、LFLO、LINEARREG_INTERCEPT、HT_TRENDLINE、KAMA。之后在完成题目要求的基础上进一步利用主成分分析和聚类分析进行延申，得到了能分别反映价值类和技术指标类信息的两个主成分因子，并基于此建立了多因子打分选股模型。该模型能够获取远超证券市场基准指数的收益，得到较为可观的收益率曲线，并且具有较低的回撤，从而风险性相对较小。

本文进一步分别使用随机森林算法、AdaBoost 算法与支持向量机算法这三种较为成熟的机器学习算法，对已建立的多因子打分选股模型在收益率与风险方面进行提升。通过对比提升模型的收益率曲线，发现支持向量机算法的提升效果最优。在完成题目所要求的模型比较后，以支持向量机算法为例，针对多因子选股模型提出了消除机器学习过拟合现象的处理方案：通过 10 次 10 折交叉检验和基于 AUC 指标的超参数网格搜索，建立了过拟合处理后的最优算法模型。该模型不仅能够取得较好的收益，同时其夏普比率达到了 4.10，说明本方案能有效降低模型的风险。而 AUC 值也达到了 0.59，说明对于模型的过拟合现象也做出了很好的处理。本文所提出的过拟合处理方案与机器学习算法的选取无关，因而可以适用于不同类型的机器学习算法，进而为各种算法模型的过拟合处理提供了一个非常有效的手段。

本文再进一步使用风险中性、行业中性和波动率主动管理模型，对每个截面期多因子选股模型选出的股票投资组合进行风险管理和资金配置，并在前文最优算法模型的基础上建立了新型的风险对冲模型，即 SVM-RC 多因子选股模型。该模型是一个结构完整的量化投资策略，通过实证分析发现，本策略不仅能够取得较高的收益率，而且可以保证策略的低风险性。

本文最后利用真实的市场数据对 SVM-RC 多因子选股模型进行了多角度的实证分析，不局限于题目中所要求的测试年份。首先利用 2016 年 1 月 1 日至 2018 年 9 月 30 日两年数据对该模型进行回测，在短期内获得了高达 470.49%的累积收益率和 8.73 的夏普比，同时模型很好的控制了短期内风险，使其最大回

撤不到 5%；其次利用 2013 年 1 月 1 日至 2019 年 3 月 1 日长时间跨度的数据对该模型进行回测，发现长时间内模型仍然可以稳定的保持 56.22%的年化收益率和低于 10%的最大回撤，较长的投资周期更使模型累积收益率达到了 1312.84%；同时在分行情实证分析中发现模型不管在哪种行情走势内都能获得远超过市场平均收益的表现，并且只有 6.63%的最大回撤。由此可见，本文所构建的 SVM-RC 多因子选股策略不仅能够在短期投资中获得极大收益，在长期投资中稳健获利，而且能够极好的适应各种市场行情，在多角度分析中选股策略的最大回撤都控制在 10%以内。

参考文献

[1] T M Mitchell. Machine Learning [M]. New York: McGraw-Hill, 1997.

[2] 李航. 统计学习方法 [M]. 北京: 清华大学出版社, 2013.

[3] 林晓明. 华泰人工智能系列之三: 人工智能选股之支持向量机模型 [R]. 广东: 华泰证券研究所, 2017.

[4] 林晓明. 华泰人工智能系列之六: 人工智能选股之 Boosting 模型 [R]. 广东: 华泰证券研究所, 2017.

[5] 林晓明. 华泰人工智能系列之五: 人工智能选股之随机森林模型 [R]. 广东: 华泰证券研究所, 2017.

[6] M Banko, E Brill. Mitigating the Paucity-of-Data Problem: Exploring the Effect of Training Corpus Size on Classifier Performance for Natural Language Processing [J]. Proceedings of the First International Conference on Human Language Technology Research, 2001: 1-5.

[7] A Halevy, P Norvig, F Pereira. The Unreasonable Effectiveness of Data [J]. IEEE Intelligent Systems, 2009, 24(2): 8-12.

[8] A Geron. 机器学习实战 [M]. 北京: 机械工业出版社, 2018.

[9] 吕凯晨, 闫宏飞, 陈翀. 基于沪深 300 成分股的量化投资策略研究 [J]. 广西师范大学学报, 2019(1): 1-12.

[10] 王淑燕, 曹正凤, 陈铭芷. 随机森林在量化选股中的应用研究 [J]. 运筹与管理, 2016(3): 163-168.

[11] 苏红果. 机器学习在土木工程施工安全管理中的应用研究 [D]. 辽宁: 大连理工大学, 2018.