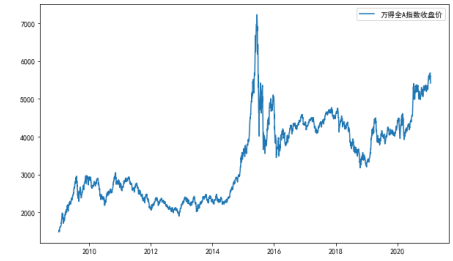
研究专题：股指驱动框架

基于小波变换和支持向量机的指数择时模型

金融工程 数量化投资

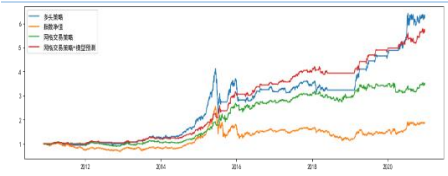
2021.03.20

关键数据（万得全A指数）：



数据来源：wind

策略表现：



本报告的独特之处：

1 在使用SVM对股指驱动模型的研究中首次引入宏观、盈利等大类因子，较为全面地考虑各方面因素对股指的驱动。

2 在多因子模型的因子初步筛选中引入协整性检验以及滞后性检验，较为严谨地考虑了因子暴露值序列与股指收益率间的相关性。

3 首次结合模型预测值以及改进后的网格交易择时策略对股指择时进行优化。

相关报告：

1. 《国信证券-基于小波分析和支持向量机的指数预测模型》
2. 《平安证券-量化择时选股系列报告二：水致清则鱼自现_小波分析与支持向量机择时研究》

·报告摘要

支持向量机（support vector machine，SVM）是量化分析最常用的机器学习算法之一，线性支持向量机可以解决线性分类问题，而核支持向量机则可以有效解决非线性、过拟合等传统回归模型的困难。小波变换（wavelet transform，WT）是目前应用数学和工程学中广泛使用的一种算法，在信号分析、图像识别等方面都取得了显著成果，近年来在股价时间序列滤波方面也渐渐开始应用。

本报告将从实证的角度出发，探究宏观、财务、盈利、行情、技术、估值六大类因子对万得全A指数的影响。模型构建的主要部分包括在因子筛选方面对因子与收益率序列的相关性和协整性进行分析；通过小波变换的降噪功能对万得全A指数进行优化，使其更能反映指数的主趋势；在模型的训练方面，采用能抓住指数序列的短期涨跌趋势的滚动的支持向量机模型；在模型指标评价方面分别对样本内、样本外、交叉验证集计算准确率、AUC、f1 score等常见指标。

基于对万得全A指数的预测值，构造多头、多空、网格交易策略的择时策略，通过年化收益率、夏普比率、最大回撤率等指标对比其异同。

·实证效果

回测的时间段为2010年1月1日至2020年12月31日（250交易日）。模型的样本外准确率为58.1%，基于模型预测值的多头策略年化收益率19.2%、夏普比率1.045，远高于基准收益率曲线的6.2%和0.256。在使用网格交易策略对多头策略进行优化后，我们可以将最大回撤率降至11.04%，并使夏普比率进一步提高。观察结合预测值的网格交易策略效果图，其净值曲线走势平缓，预测值也能准确把握涨跌趋势。

·实证结论

在因子筛选过程中，在考虑滞后因素后，大部分因子与万得全A指数收益率有协整性以及相关性。模型的整体预测准确率较高，基于模型预测值的择时策略在收益、风险控制方面的表现也远优于指数基准净值曲线，说明WT-SVM模型在2010-2020回测区间内具有良好的预测效果。然而在对小波降噪效果评测中，是否使用小波分析对模型的准确率、f1 score、AUC等指标影响不大，对择时策略的年化收益率、夏普比率仅有些许提升，说明小波变换滤波对模型的预测效果影响有限。

内容目录

图表目录

1 基本理论简介

1.1 小波变换（WT）

小波变换，或小波分析是指用有限长或快速衰减的震荡波来表示信号，在图像压缩、数字信号降噪方面有者广泛应用。在金融领域，小波分析可以很好地对金融时间序列中的非平稳性、周期季节性、奇异信号等特性进行分解，通过滤波处理提取主趋势，从而使时间序列的处理变得简单。

小波分析是在傅里叶变换的基础上发展起来的。传统的傅立叶变换的基本原理在于用一系列简单的正弦波和余弦波之和来表示一个复杂的周期函数。然而，传统的傅里叶变换在处理金融时间序列上有者天生缺陷：

a. 只能获取序列总体上包含哪些频率的信号，而无法在时域上对信号出现进行定位。

b. 傅立叶变换只能反映整体信号中包含某一频率分量的平均值，而无法反映信号频率随时间的变换，所以只能处理平稳的时间序列。

所以在时间序列降噪方面，我们更多地使用小波分析。

小波分析的数学原理：

一个离散型信号$x(t)$可以分解成若干个尺度信号$\varphi_{j_0,k}(t)$ 和小波信号$\psi_{j,k}(t)$ 的线性组合：

$$x(t)=\sum_{k} c_{j_0,k}\varphi_{j_0,k}(t)+\sum_{j=j_0}^{\infty}\sum_{k} d_{j,k}\psi_{j,k}(t)$$

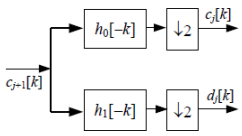
由小波函数多辨率方程，给定小波基函数$\psi(t)$，经过展缩和平移可以得到新函数族 $\psi_{j,k}(t)=2^{\frac{j}{2}}\Psi(2^j t-k);j,k\in Z$，其中j、k分别代表基函数的频域和时域信息。小波信号一般用于表示原信号中的高分辨率精细信息。

尺度函数族也可以由类似的方法定义：$\varphi_{j,k}(t)=2^{\frac{j}{2}}\varphi(2^j t-k);j,k\in Z$，一般用于表示原信号中的低分辨率粗略信息。

小波信号 $\psi_{j,k}(t)$ 可以设计为尺度信号 $\varphi_{j_0,k}(t)$ 的正交信号，即存在$<\varphi_{j,k}(t),\ \psi_{j,k}(t)>\ =\int\varphi_{j,k}(t)\,\psi_{j,k}(t)\,dt=0;j,k,l\in Z$。如图1 和图2所示，理论上我们可以通过尺度信号的若干级分解得到小波变换系数{cj[k]，dj [k]}，也可以由处理后的小波变换系数重构滤波后的新信号

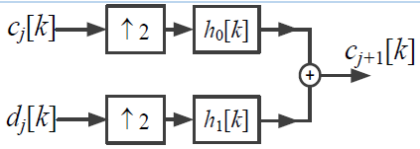
$x(t)$。

图 1：离散小波变换（DWT）一级分解算法



（Mallat算法）

图 2：离散小波变换（DWT）一级重构算法



IDWT 一级重构(synthesis)算法框图

基于小波分析的滤波步骤：

一般来说，包含加性噪声的非平稳时间序列模型为 $s(t) = x(t) + \sigma * \varepsilon(t)$，其中x(t)为有用信号，$\varepsilon(t)$为高斯白噪音信号，$\sigma$为噪声信号的方差。基于小波分析的滤波处理就是抑制分量$\sigma * \varepsilon(t)$，恢复信号x(t)。

a. 选择小波基函数，对信号进行等间隔抽样，得到$c_{j+1}[k]$。基于$c_{j+1}[k]$进行N级DWT，得到小波展开系数$d_{j-i+1}[k]$，i=(1,2•••N)和一级近似展开系数$c_{j-N+1}[k]$。

b. 对各级小波展开系数进行阈值化处理。一般来说噪声部分包含在高频信号中，我们可以对高频系数进行门限阈值量化处理。阈值过高会使信号失真，过低又会使得消噪不完全。在本报告中我们使用无偏风险估计准则（rigrsure）选定阈值：对每个阈值求出对应的风险值，风险值最小的即为所选。

c. 根据处理后的系数和其他未处理的系数进行小波重构。

1.2 支持向量机（SVM）

支持向量机是在分类与回归分析问题中的一种监督性学习相关的机器学习算法，因其在准确率、非线性问题、运算效率上的优异表现而被广泛使用。

类似于其他分类器算法，SVM旨在用N-1个超平面对N维数据进行分类。图3 为一个简易的二维平面支持向量机分类问题。超平面$wx^T + b = 0$是定义在空间内能够完全将点集划分成两份的平面，SVM的核心思想就是找到一个最优超平面，通过平移这个超平面直到与样本点相交，可以得到如图两条虚线（最大边缘超平面），其距离$||w||$是最大化的，即

$$max_{w,b} \frac{2}{||w||}, s.t: y_i(x_i^T w + b) \geq 1$$

其中x,y为输入特征值以及标签值

我们通过以下步骤解决这个最优化问题：

a. 通过拉格朗日乘数法将最优化问题转换为$min_{w,b}max_{\lambda}L(w,b,\lambda)=\frac{1}{2}||w||^2+\sum_{i=1}^{n}\lambda_i[1-y_i(w^Tx_i+b)]\quad s.t:\lambda_i\geq 0$

b. 利用强对偶性将（a）转化为 $max_{\lambda}min_{w,b}L(w,b,\lambda)$，$s.t:\lambda_i\geq 0$

c. 对（b）求w,b的偏导数，可以将问题转换为二次规划：$max_{\lambda}[\sum_{i=1}^{n}\lambda_i-\frac{1}{2}\sum_{i,j=1}^{n}w_iw_j(x_ix_j)]\ s.t:\sum_{i=1}^{n}\lambda_iy_i=0,\ \lambda_i\geq 0$，其中$w_i=\lambda_iy_i$，并用序列最小优化算法（SMO）求出最优超平面。

在支持向量机的实际应用中，往往还应注意以下问题：

a. 软间隔与松弛变量的引入：在实际应用中，完全线性可分的样本是很少的，我们往往会允许部分样本点出现在间隔带内，如图4 所示。此时优化目标拓展为$min_w\frac{1}{2}||w||^2+C\sum_{i=1}^{m}\xi_i,s.t:\ y_i(x_i^Tw+b)+\xi_i\geq 1$，其中C为大于0的常数，表示对错误样本的惩罚程度，ξ_i为松弛变量。

b. 非线性分类与核函数：在处理非线性问题时，核支持向量机可以通过非线性映射，将原始数据变换到高维特征空间，随后再使用线性分类解决。此时二次规划的目标函数变为$max_{\lambda}[\sum_{i=1}^{n}\lambda_i-\frac{1}{2}\sum_{i,j=1}^{n}w_iw_jK(x_ix_j)]$，其中$K(x_ix_j)$为核函数。常见的核函数有线性核、多项式核、高斯核（RBF）。

c. 判别函数：通过训练集训练，我们可以得到支持向量参数w,b。在测试集中，计算$f(x)=\widehat{w^T}x+\hat{b}=sign(\sum_{i=1}^{n}\widehat{w^T}K(x_ix_j)+\hat{b})$，可以得到预测标签值。

d. Gamma值：多项式核，高斯核的核函数参数，决定了原始数据映射到高维空间后的分布情况。Gamma越大，样本在高维空间中的分布越稀疏，支持向量的个数越多，测试集正确率越高，但也更容易导致过拟合。

1.3 模型评价指标

对于二分类问题，根据成功预测正标签（TP）、成功预测负标签（TN）、错误预测正标签（FP）、错误预测负标签（FN）的样本个数，我们可以计算以下评价指标：

a. 准确率（accuracy）=(TP+TN)/(TP+TN+FP+FN)

b. 召回率（recall）=TP/（TP+TN）

c. 精确率（precision）=TP/(TP+FP)

d. 虚报率=FP/(FP+TN)

e. f1 score=(2*precision*recall)/(precision+recall),由于准确

率和召回率常常互相反向影响，我们常常会计算召回率和精确率的调和平均数f1 score，以有效兼顾二者的表现。

ROC曲线：以召回率为纵坐标，虚报率为横坐标的图。ROC曲线下的面积称为AUC值，AUC的值在0.5到1之间，越接近1说明分类器性能越好。ROC的思想是对所有分类阈值（例如划定收益率为0作为标签值划分阈值）可能的取值进行遍历，这样可以避免分类阈值对评价指标的干扰，从而仅关注分类器的性能。

图 3：线性支持向量机的分类超平面

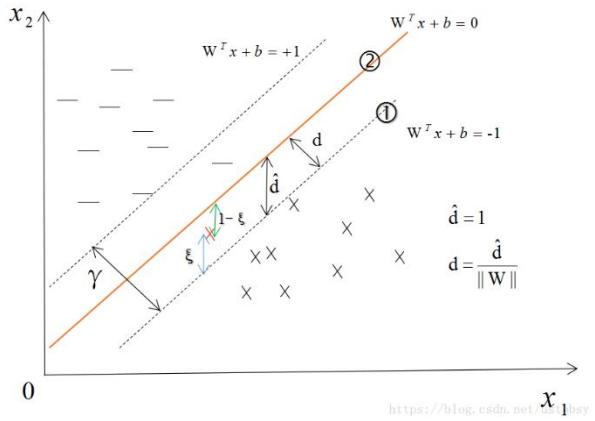
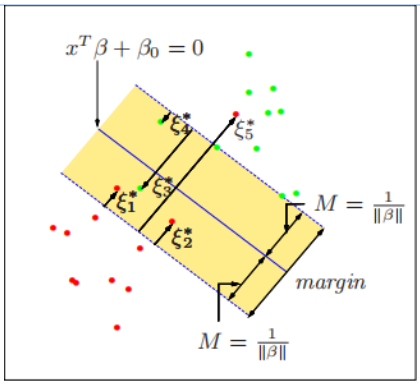


图 4：软间隔与松弛变量



2 WT-SVM 指数预测模型

图 5：WT-SVM 模型构建示意图

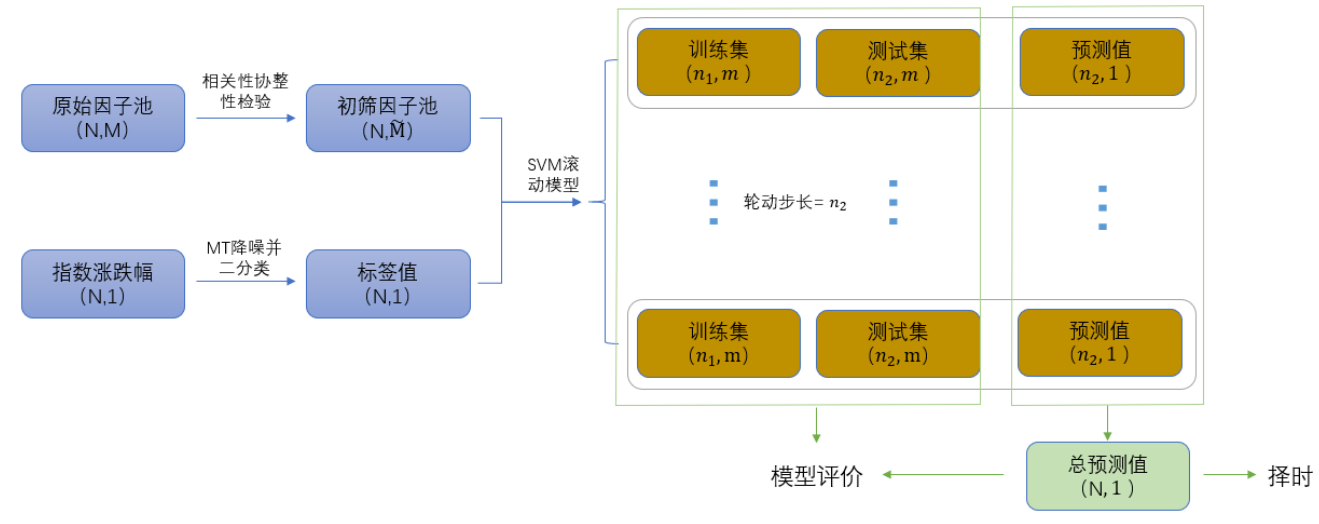
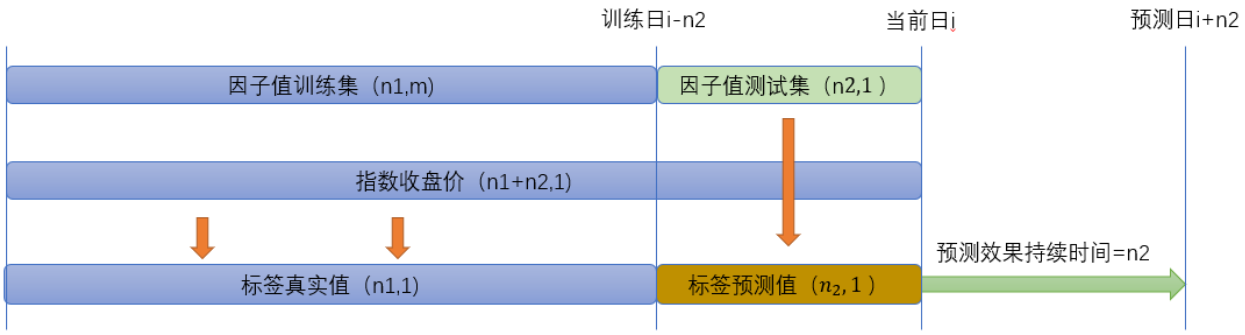


图 6：SVM 滚动训练模型实际意义



如图 5 所示，WT-SVM模型的构建主要分为以下步骤

1. 数据获取：
   a. 原始因子池：如图7 所示，我们从宏观、财务、估值、技术、盈利、行情六大类筛选出约30个可能会对万得全A指数产生影响的因子。得到一个（N, M）的二维数据结构。其中N为总天数，M为因子值个数。
   b. 指数：获取万得全A指数（881001.WI）。
   c. 回测区间：2010-01-01至2020-12-31。由于我们采用滚动回测模型，不需要额外划出测试集。
2. 使用小波变换对指数进行滤波处理：
   a. 每个交易日计算未来一个月后的月涨跌幅 $pch_i=(p_{i-q}-p_i)/p_i-1$，$p_i$为第i日收盘价，q为一个月的天数。
   b. 对pch_i时间序列进行小波分解，小波函数选用db8，分解层数为4，得到一个主趋势和四个细节成分时间序列并对频率较高的细节进行滤波处理。由于涨跌幅频率较高，我们采用无偏风险估计准则来确定阈值。得到处理后的涨跌幅数据$\widetilde{pch}$。
3. 标签提取以及因子值混频数据预处理
   a. 定义标签值 $y_i = 1\ if\ \widetilde{pch_i}$ >=0， $y_i = -1\ if\ \widetilde{pch_i}$ <0。
   b. 混频数据处理：对于月频、季频的数据，我们统一采用后值填充法填充成日频数据。对于绝大多数低频因子我们同时也进行了基于移动平均法的平滑处理。
4. 因子池的初步筛选
   a. 去极值处理：采用三倍标准差法对每个因子暴露值时间序列进行去极值。
   b. 标准化处理：将去极值后的因子暴露值序列减去其现在均值，除以其标准差，得到一个近似服从N(0, 1)分布的时间序列。
   c. 相关性检验：计算 pearson($x_{i-u,j}$，$\widetilde{pch_i}$)及其p-values。其中$x_{i-u,j}$为第i日因子j的暴露值，u为滞后期参数。由于宏观数据等因子对股指的影响很可能有滞后性，此处我们应对原时间序列及其多个滞后期进行相关性和协整性检验。如果p-values<0.1，则通过相关性检验。

d. 协整性检验：先对$x_{i-u,j}$，$\widetilde{pch_i}$ 的单整性进行检验，此处我们采用ADF检验来检验$x_{i-u,j}$，$\widetilde{pch_i}$ 的一阶差分序列的平稳性，以确保满足协整性检验前提。之后检验$x_{i-u,j}$，$\widetilde{pch_i}$ 的协整性。如果协整性统计值显著，则通过协整性检验。

e. 筛选标准：对于因子时间序列$x_{i-u,j}$，如果存在一个及以上的u值使得$x_{i-u,j}$通过相关性检验以及协整性检验，取序列相关性最高的u值$\tilde{u}$作为滞后参数，并更新原时间序列$x_{i,j} = x_{i-\tilde{u},j}$。若都不通过，则舍弃该因子。初步筛选后，剩余因子个数为$\widetilde{M}$。

5. SVM滚动模型

a. 因子值与标签的结合：将初筛因子值与标签序列合并成（N，$\widetilde{M}$+1）的二维数据。

b. 训练集、交叉验证集、测试集的划分：如图5 所示，假如在第k日进行一次训练，我们取第k−n_1日到第k日的数据作为训练集，n_1取一年的天数；第k+1日到第k+n_2日的数据作为测试集，n_2取值为一个月的天数q。同样地我们也会在训练集中进行交叉验证。

c. 模型的实际意义如图6 所示，如果我们在当前日i想知道预测日i+n_2的指数收盘价相较于当前日的趋势，我们可以取i−n_1到 i−n_2的数据作训练，之后将第i日的因子值输入训练后的模型，得到的即为一个月后指数涨跌趋势的预测。

d. 缺失值处理：在某次训练以及测试中，若某因子出现缺失值，则该次训练剔除该因子。

e. 总预测值：对于每次滚动的测试集进行预测。由于测试集时间跨度与步长一致，我们最后可以得到一个长度约为N的总预测值。

f. PCA降维处理：为避免因子共线性，每次测试前都先对训练集进行主成分分析，取累计解释程度大于0.99的前k个成分作为维度转换后的新训练集，之后同样地也对测试集取前k个成分作为新测试集。

g. 样本内训练：使用SVM模型对训练集进行训练，并用GridSearchCV来进行调参。待调的参数有：惩罚系数C，gamma，kernel。模型训练完成后，选择交叉验证集准确率（accuracy）最高的的一组参数作为当期模型的最优参数。

6. 模型评价：

a. 训练集AUC、训练集准确率、训练集f1 score，测试集准确率，测试集f1 score，交叉验证集准确率、交叉验证集f1 score。

b. 各因子对模型的解释程度检验：以一年为周期，分析每个因子与预测值的相关性系数。并且为了检验因子对标签值的区分能力，对每个因子按照标签值y_i=1和y_i=−1分为两组，求其比值。

c. 择时策略净值曲线的夏普比率、年化收益率、最大回撤率、波动率、累计收益率。

研究专题：股指驱动框架

图 7：股指驱动因素总览及其描述

大类因子	具体因子	因子描述	特殊处理
宏观	cpi_month	全国居民消费指数（月）	后值填充，120日移动平均
宏观	pmi	全国制造业采购经理指数（月）	后值填充，60日移动平均
宏观	ind_growth_yoy	全国工业增加值同比增长率（月）	后值填充，60日移动平均
宏观	fixed_assets_inv	固定资产投资完成额累计值（月）	后值填充，60日移动平均
宏观	consistency_idx	一致指数（月）	舍弃
宏观	early_warning_idx	预警指数（月）	后值填充，60日移动平均
宏观	foreign	外汇储备金额（美元）（月）	后值填充，60日移动平均
宏观	m2	货币和准货币供应量（月）	舍弃
宏观	m1	货币供应量（月）	舍弃
宏观	retail_sin	社会消费品零售总额（月）	后值填充，60日移动平均
宏观	gdp_sin	国内生产总值（季）	后值填充，120日移动平均
宏观	commodity_hs_idx	70大城市新建商品住宅销售价格指数均值（月）	后值填充，60日移动平均
行情	中债10年	中债10年到期收益率	60日移动平均
行情	成交额	日成交额	
行情	换手率	日换手率	
行情	自由流通市值	自由流通市值	
估值	PE(TTM)	市盈率PE(TTM)=∑(成分股,总市值2)/∑(成分股,归母净利润(TTM))，总市值2=指定日证券收盘价*指定日当日总股本	20日移动平均
估值	股息率	股息率＝∑近12个月现金股利(税前)／指定日股票市值×100%	60日移动平均
估值	PB(LF)	市净率PB(LF)=∑(成分股,总市值2)/∑(成分股,净资产(最新年报LYR))	120日移动平均
盈利	一致预测净利润同比	一致预测净利润同比=(一致预测净利润_FY1-净利润_FY0〕/净利润_FY0*100	60日移动平均
盈利	一致预测每股收益（FY 1）	截止指定交易日,各机构对该证券的最近预测年度的预测每股收益的算术平均值，有效期180天	60日移动平均
盈利	归属母公司净利润同比增长率	季度数据	后值填充
财务	ROE（平均）	2*∑（成份股当年当期净利润）／∑（成份股期初净资产+期末净资产)*100%，季度	后值填充
财务	EPS(TTM)	归属母公司股东的净利润(TTM)／最新总股本，季度	后值填充
财务	营业收入同比增长率	(∑（成份股当年单季度营业收入）/∑（成份股上年该季度营业收入）-1)*100%，季度	后值填充，60日移动平均
技术	macd	指数平滑异同移动平均线	
技术	kdj	随机指标	
技术	rsi	相对强弱指标	

技术	atr	均幅指标

数据来源：wind，聚宽

3 WT-SVM 模型在万得全 A 指数上的实证

3.1 因子筛选

在因子池的初步筛选中，我们希望寻找能解释未来一个月指数涨跌幅的因素。其中一个最简单的方法就是计算因子暴露值与涨跌幅的相关性系数以及显著性。但在实际应用中，大多数因子序列都是非平稳的，为了防止伪回归的出现，我们应检验二者的同阶单整性以及协整性。同时，很多变量与指数的变化可能存在一定的时滞，所以我们还应对多个时滞因子进行分析。同一个因子若有多个滞后期显著，则优先选择相关性较高的序列

由于SVM模型最终的决策函数取决于少数支持向量，此处我们适当放宽筛选标准，仅要求相关性的p值小于0.1，协整性的p值小于0.1。

对各因子与指数涨跌的测试结果如图8 所示，表格的四列分别为pearson相关性系数，pearson相关性检验p值，一阶差分ADF检验p值与万得全A指数协整性检验的p值。图9 为万得全A指数涨跌的时间序列，经测试，其一阶差分ADF检验的统计值为-8.09，远小于10%置信水平值-2.567。我们可以观察到所有的因子与指数涨跌序列都为一阶单整，满足协整性检验的前提。所有的协整性统计值均显著，故所有的因子序列都与指数涨跌有协整关系。

在相关性检验中，我们舍弃统计值非显著的因子pmi，m1，m2，并根据相关性系数绝对值大小选择滞后期，对于绝大多数因子序列最后都采取原序列或一个月滞后的序列。

图 8：相关性及显著性检验

	corr	p value	1d adf	coint p value
cpi_month(0)	-0.082884868	2.07428E-05	0.010915435	1.19854E-09
cpi_month(-20)	-0.099106305	3.8913E-07	0.015931838	1.34145E-09
cpi_month(-40)	-0.104367728	1.01234E-07	0.011264958	1.25117E-09
cpi_month(-60)	-0.11058204	1.89211E-08	0.018822966	1.27933E-09
cpi_month(-120)	-0.083900645	2.55844E-05	0.046761504	1.07229E-08
pmi(0)	-0.001021505	0.958232727	0	1.80331E-09
pmi(-20)	-0.008111433	0.678664732	0	2.18687E-09
pmi(-40)	0.006775739	0.730293396	0	2.15446E-09
pmi(-60)	-0.017555692	0.373574611	0	2.27538E-09
pmi(-120)	-0.032614704	0.102271493	0	1.26502E-08
ind_growth_yoy(0)	-0.117392863	6.38397E-09	8.18614E-27	2.0294E-08
ind_growth_yoy(-20)	-0.086325664	2.18238E-05	9.12784E-27	2.82648E-08
ind_growth_yoy(-40)	-0.058660835	0.00410523	7.27092E-27	3.37873E-08
ind_growth_yoy(-60)	-0.047838935	0.019805394	8.96091E-27	4.67131E-08
ind_growth_yoy(-120)	-0.046844046	0.024294927	1.80052E-26	9.85154E-08
fixed_assets_investment(0)	0.119446411	3.44874E-09	0	2.37762E-08
fixed_assets_investment(-20)	0.10175534	5.50464E-07	0	3.24932E-08
fixed_assets_investment(-40)	0.054995971	0.007137125	0	2.92841E-08
fixed_assets_investment(-60)	0.032954072	0.108591121	0	4.01893E-08
fixed_assets_investment(-120)	0.126685667	9.78925E-10	0	1.09605E-07
consistency_idx(0)	-0.13342068	7.37848E-12	5.60201E-05	8.76608E-10
consistency_idx(-20)	-0.1290418	4.1682E-11	6.9585E-05	9.31065E-10
consistency_idx(-40)	-0.121806329	5.60709E-10	7.63558E-05	9.76375E-10
consistency_idx(-60)	-0.119661109	1.30028E-09	8.30248E-05	8.98734E-09
consistency_idx(-120)	-0.123448461	6.13841E-10	0.0001099	8.97885E-09
early_warning_idx(0)	-0.130237774	9.84389E-09	8.30396E-07	1.39446E-05
early_warning_idx(-20)	-0.138945346	1.14022E-09	1.58401E-06	1.45263E-05
early_warning_idx(-40)	-0.144964576	2.59138E-10	3.1698E-06	1.51817E-05
early_warning_idx(-60)	-0.14975171	8.15438E-11	2.23187E-06	3.65726E-05
early_warning_idx(-120)	-0.141220308	1.6997E-09	1.50256E-05	5.77345E-05

	corr	p value	1d adf	coint p value
十年国债(0)	-0.043965213	0.022995481	5.50021E-06	1.3583E-09
十年国债(-20)	-0.045271861	0.019681392	1.99181E-06	1.46627E-09
十年国债(-40)	-0.03280264	0.09234264	1.85658E-06	1.68672E-09
十年国债(-60)	-0.011699494	0.549908131	2.12537E-06	2.02899E-09
十年国债(-120)	0.016297541	0.41034754	4.71878E-06	1.64318E-08
成交额(0)	0.003245249	0.866766146	6.86115E-29	1.30815E-09
成交额(-20)	0.325448442	5.19555E-67	6.48649E-29	1.24462E-08
成交额(-40)	0.244054558	2.72583E-37	8.37233E-29	5.37343E-10
成交额(-60)	0.198151551	9.94811E-25	9.49783E-29	5.90912E-11
成交额(-120)	0.184648033	3.56024E-21	1.45654E-28	2.708E-10
换手率(0)	0.018475575	0.339477044	0	1.0055E-09
换手率(-20)	0.410488805	3.2331E-109	4.60365E-30	5.16364E-09
换手率(-40)	0.285957541	4.02324E-51	6.68359E-30	1.18529E-10
换手率(-60)	0.197857287	1.16867E-24	6.76641E-30	1.5001E-11
换手率(-120)	0.17752873	1.14048E-19	7.48253E-30	1.83879E-10
自由流通市值(0)	-0.025588083	0.185827396	1.5823E-17	1.50223E-09
自由流通市值(-20)	0.106009135	3.91296E-08	3.85645E-20	1.72088E-09
自由流通市值(-40)	0.090034648	3.39145E-06	4.86479E-20	1.54866E-09
自由流通市值(-60)	0.101085929	2.00534E-07	5.87635E-20	1.24002E-09
自由流通市值(-120)	0.088635504	6.68217E-06	1.19117E-19	1.44873E-09
PE(TTM)(0)	-0.171593524	3.99513E-19	1.62594E-08	2.39181E-09
PE(TTM)(-20)	0.012927296	0.504008927	1.68814E-08	1.37271E-09
PE(TTM)(-40)	0.165559498	9.13019E-18	7.0193E-09	2.05676E-09
PE(TTM)(-60)	0.172877432	4.04792E-19	8.32998E-11	6.18349E-10
PE(TTM)(-120)	0.159215784	4.44752E-16	2.09656E-08	4.35911E-10
股息率(0)	0.173751023	1.41407E-19	8.34812E-06	6.46514E-10
股息率(-20)	0.126744126	4.78397E-11	9.20321E-06	1.17848E-09
股息率(-40)	0.031034382	0.109947382	1.02122E-05	1.41563E-09
股息率(-60)	-0.063599219	0.001091443	1.14022E-05	1.88533E-09
股息率(-120)	-0.104648134	1.03018E-07	6.81723E-06	1.03074E-09

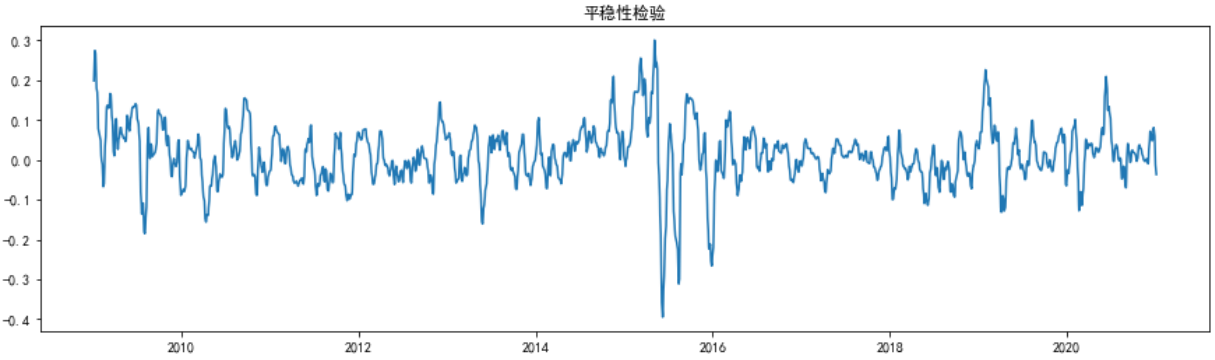
	corr	p value	1d adf	coint p valu
atr(0)	-0.00323154	0.867323718	1.2591E-22	1.37185E-(
atr(-20)	-0.021630012	0.263519313	2.54666E-22	1.62271E-(
atr(-40)	0.142118335	1.90324E-13	2.95767E-22	1.19851E-(
atr(-60)	0.138128635	1.08008E-12	2.1817E-22	1.28673E-(
atr(-120)	0.136871771	3.08923E-12	1.14088E-21	4.58796E-:
ROE(平均)(0)	-0.069862819	0.000350822	0	2.37224E-(
ROE(平均)(-20)	-0.018291204	0.351737632	0	2.08631E-(
ROE(平均)(-40)	0.006675297	0.734980529	0	2.17996E-(
ROE(平均)(-60)	-0.031104902	0.116051993	0	1.49317E-(
ROE(平均)(-120)	-0.067966661	0.000682588	0	1.52006E-(
归母净利润同比增长率(0)	-0.083786456	1.79262E-05	0	1.36802E-(
归母净利润同比增长率(-20)	-0.050286379	0.01042098	0	1.86747E-(
归母净利润同比增长率(-40)	-0.055215288	0.00507707	0	1.94931E-(
归母净利润同比增长率(-60)	-0.033708517	0.088534152	0	1.55177E-(
归母净利润同比增长率(-120)	-0.041339023	0.038987694	0	1.74858E-(
EPS(0)	-0.053864597	0.00587579	0	2.39902E-(
EPS(-20)	0.002969585	0.879839788	0	2.20402E-(
EPS(-40)	0.033443469	0.089811768	0	2.55826E-(
EPS(-60)	-0.009536844	0.629990931	0	1.61877E-(
EPS(-120)	-0.040790843	0.041658086	0	1.64351E-(
营业收入同比增长率(0)	-0.103219069	1.23041E-07	2.15658E-07	1.25575E-(
营业收入同比增长率(-20)	-0.103974892	1.1119E-07	1.7935E-09	1.2532E-(
营业收入同比增长率(-40)	-0.118487061	1.64239E-09	3.43621E-13	1.07738E-(
营业收入同比增长率(-60)	-0.127868888	8.83792E-11	0.000145982	9.28583E-(
营业收入同比增长率(-120)	-0.116245152	5.81073E-09	0.00014683	1.00416E-(

选入因子池的因子

	corr	p value	1d adf	coint p value
foreign(0)	0.15712824	5.22306E-16	4.94991E-11	7.51347E-10
foreign(-20)	0.15254432	4.619E-15	4.92866E-11	8.01103E-10
foreign(-40)	0.140891825	5.84036E-13	8.4352E-11	7.2908E-10
foreign(-60)	0.135968442	4.41853E-12	1.77783E-10	9.31016E-10
foreign(-120)	0.141429104	1.08494E-12	1.42464E-10	7.64287E-09
m2(0)	0.02176589	0.264402694	2.98392E-07	1.7937E-09
m2(-20)	0.017378308	0.374735695	3.88008E-07	2.19897E-09
m2(-40)	0.016682231	0.395989264	4.18268E-07	2.17262E-09
m2(-60)	0.021552495	0.274650584	4.72841E-07	2.34237E-09
m2(-120)	-0.010586656	0.595940839	6.67515E-07	1.48389E-08
m1(0)	0.017002585	0.383335584	6.03147E-08	1.78983E-09
m1(-20)	0.01302012	0.506043453	6.84277E-08	2.1907E-09
m1(-40)	0.016509834	0.400889825	7.74492E-08	2.17595E-09
m1(-60)	0.024996172	0.205150563	8.77687E-08	2.35863E-09
m1(-120)	-0.01303723	0.513758563	1.26358E-07	1.48363E-08
retail_sin(0)	0.056463586	0.00501809	2.97025E-07	1.75847E-08
retail_sin(-20)	0.068875986	0.000649378	3.3076E-07	2.61162E-08
retail_sin(-40)	0.076269882	0.000168747	3.77742E-07	3.05173E-08
retail_sin(-60)	0.029296068	0.150672419	4.31604E-07	3.41286E-08
retail_sin(-120)	0.023393113	0.257175628	6.29779E-07	5.74071E-08
gdp_sin(0)	0.03348901	0.092225361	2.88246E-05	1.38356E-08
gdp_sin(-20)	0.025157524	0.207774091	3.10291E-05	1.41632E-08
gdp_sin(-40)	0.016986821	0.396936118	3.63308E-05	1.95748E-08
gdp_sin(-60)	0.011423325	0.570479793	3.9213E-05	2.18432E-08
gdp_sin(-120)	0.00667729	0.743240278	4.92216E-05	3.54206E-08
commodity_house_idx(0)	-0.134353682	4.53053E-12	1.41812E-13	5.8837E-10
commodity_house_idx(-20)	-0.089360125	4.80599E-06	3.90629E-13	1.11675E-09
commodity_house_idx(-40)	-0.089595557	4.93414E-06	1.59134E-12	1.38286E-09
commodity_house_idx(-60)	-0.062771312	0.001450474	1.40271E-12	1.76472E-09
commodity_house_idx(-120)	-0.048380405	0.015327418	1.44025E-12	1.09841E-08

	corr	p value	1d adf	coint p value
PB(LR)(0)	-0.170709564	6.09045E-19	0.008086226	4.98886E-10
PB(LR)(-20)	-0.149569114	7.57571E-15	0.006579475	6.71909E-10
PB(LR)(-40)	-0.100432359	2.16423E-07	0.008338375	9.74401E-10
PB(LR)(-60)	-0.056068788	0.003995847	0.006665898	1.55989E-09
PB(LR)(-120)	0.104594033	1.04596E-07	0.006824053	1.37597E-09
一致预测净利润同比(0)	-0.040205222	0.037589872	3.38633E-05	1.17744E-09
一致预测净利润同比(-20)	-0.029690089	0.124804264	3.16129E-05	1.28334E-09
一致预测净利润同比(-40)	-0.019186995	0.323112748	3.43481E-05	1.46992E-09
一致预测净利润同比(-60)	0.001865488	0.923761888	3.73219E-05	1.72693E-09
一致预测净利润同比(-120)	0.055421535	0.004914495	4.78941E-05	1.80726E-09
一致预测每股收益(FY1)(0)	-0.000719043	0.970347992	4.92469E-05	1.35787E-09
一致预测每股收益(FY1)(-20)	0.017748475	0.358917466	4.84618E-05	1.37855E-09
一致预测每股收益(FY1)(-40)	0.025579654	0.187710796	5.24248E-05	1.43939E-09
一致预测每股收益(FY1)(-60)	0.042427279	0.02944915	5.67143E-05	1.604E-09
一致预测每股收益(FY1)(-120)	0.081611336	3.39186E-05	7.18238E-05	1.55346E-09
macd(0)	0.012909409	0.504519512	2.10041E-27	9.81001E-10
macd(-20)	0.842189707	0	2.06761E-27	8.68945E-17
macd(-40)	0.385156383	1.35803E-94	2.58896E-27	7.67942E-14
macd(-60)	0.053377992	0.006141196	1.60061E-27	3.11385E-10
macd(-120)	0.032975095	0.094402008	2.44545E-27	2.16731E-09
kdj(0)	0.018597229	0.336305966	1.96249E-26	1.05636E-09
kdj(-20)	0.602743127	2.922E-264	1.67363E-26	9.80741E-12
kdj(-40)	-0.031824008	0.101188303	1.70789E-26	1.88635E-09
kdj(-60)	-0.003763059	0.846928717	1.88558E-26	1.78046E-09
kdj(-120)	-0.066662702	0.000713788	7.26471E-26	2.85362E-09
rsi(0)	0.039880976	0.039158582	2.55153E-27	7.17401E-10
rsi(-20)	0.581734278	5.3219E-242	4.05956E-26	1.68114E-16
rsi(-40)	-0.003759988	0.846479149	4.77723E-26	1.48297E-09
rsi(-60)	0.002863493	0.883217113	4.9891E-27	1.67065E-09
rsi(-120)	-0.070488407	0.000345022	1.35514E-26	2.92551E-09

图 9：万得全 A 一个月后涨幅序列图



3.2 SVM 参数的选择

在每次滚动的SVM训练中，我们选择支持向量机最关键的三个参数加入基于网格搜索（Grid SearchCV）的参数寻优中，惩罚系数C=[0.01,1,3]，gamma=[1e-4,1e-3,1e-2,1]，kernel =[‘RBF’,’linear’]。在每次训练中，基于交叉验证集准确率（accuracy）均值选出最优的参数组合。图10 展示了每个参数的使用频率，我们可以发现模型在大多数情况下使用RBF核以及更为精细的支持向量。这说明传统的线性模型很难对万得全A指数趋势做出预测。

3.3 模型评价指标

在每次滚动训练我们都可以得到训练集（样本内）、交叉验证集、测试集（样本外）的各种评价指标，对其求均值得到模型最终的评价指标。我们选取了图11所示的6个指标来评价模型的解释能力。其中样本外准确率为0.583，远高于随机猜测水平；f1 score表现也良好，说明模型在解释万得全A指数未来一个月涨跌趋势上是基本有效的。由于测试集标签值样本容量太少且分类单一，故不再对样本外AUC进行测试。

图 10：参数频率统计表

C	
0.1	0.317829
1	0.286822
3	0.395349
gamma	
0.0001	0.457364
0.001	0.069767
0.01	0.178295
1	0.294574
kernel	
rbf	0.790698
linea	0.209302

图 11：模型各评价指标结果

样本内准确率	：	0.8222796439850703
样本外准确率	：	0.5813953488372093
交叉验证集准确率	：	0.6209148679876102
样本内 AUC	：	0.7616698688468296
样本内 f1 score	：	0.7800705919102779
样本外 f1 score	：	0.42484254667317295

图 12：准确率时间序列图

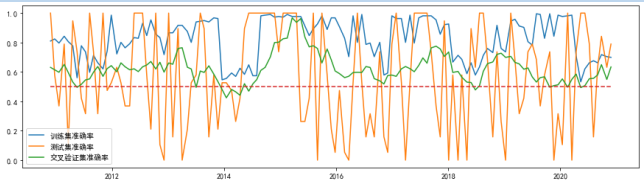
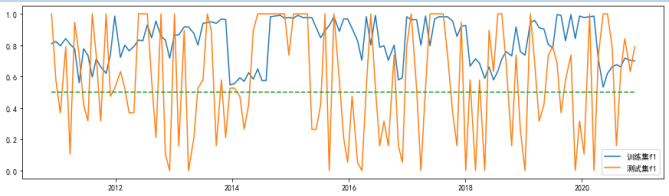


图 13：f1 score 时间序列图



3.4 各因子在模型预测值与真实值上的解释力度

为进一步探究各因子值在预测指数涨跌趋势（标签值）中所起的作用，我们从两个方面构建因子评价体系：

a. 以一年为滚动步长，两年为回看区间，计算因子序列与标签值的相关性系数绝对值。相关性系数应进行独热编码转换的预处理。

b. 以一年为滚动步长，两年为回看区间，对每个因子按照标签值y_i=1和y_i=-1分为两组，分别求均值得到pos_t和neg_t。当期的评价指标$c_t = abs(pos_t - neg_t)$。由于之前因子值均进行过标准化处理，c_t 的大小将直接反映因子暴露值对标签值的区分能力。

图14 和图16 的结果说明了行情、盈利以及部分宏观大类因子与预测标签值关联性较高；在空间维度上2014年前后各因子对标签值的分类能力均高于其他年份，而之后年份的表现均有所下降。同时我们也将因子暴露值序列与实际标签值进行了上述两方面分析，如图15 和图17 所示，其结论也类似。

图 14：因子值与预测标签值相关性系数

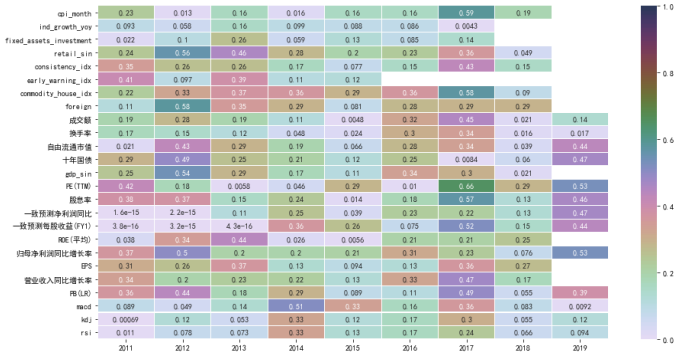


图 15：因子值与实际标签值相关性系数

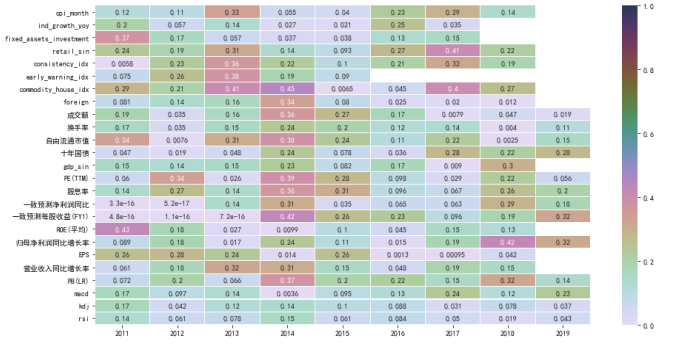


图 16：因子区分预测标签值的能力

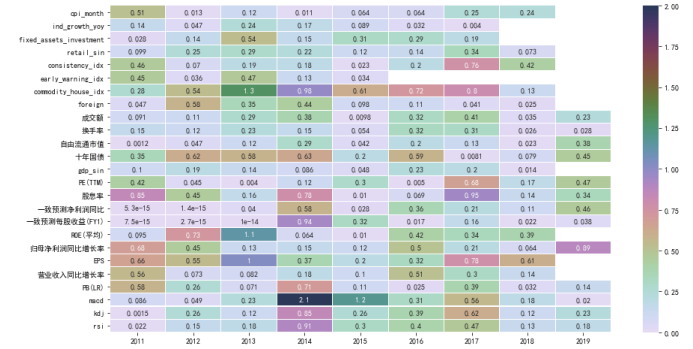


图 17：因子区分实际标签值能力

3.5 小波滤波效果检验

在模型中，我们把万得全A指数时间序列看成是一种特殊的信号，并用小波变换对齐进行滤波处理。我们使用db8小波将指数分四层展开，如图19 所示。我们发现细节3和细节4所包含的高频噪音较多，故对其进行滤波。由于涨跌幅数据类型的频率较高，我们采用无偏风险估计准则（rigrsure）来确定阈值，最终再根据主趋势以及细节一、二和重构后的细节三、四得到滤波后的指数图像，如图18 所示。经过小波滤波处理后，指数的走势明显平滑了。

滤波后万得全A指数在模型评价以及模型择时中的表现如图20 所示。在指标评测方面，二者相差不大，但滤波模型在择时模型中表现略优于未滤波模型。在进一步的探究中，使用小波滤波的基于月频策略的预测上涨、下跌成功概率分别为0.714、0.711，未使用小波滤波的结果是0.753、0.692，较稳健的预测能力使得小波滤波模型带来更高的择时收益。

图 18：小波滤波效果对比图

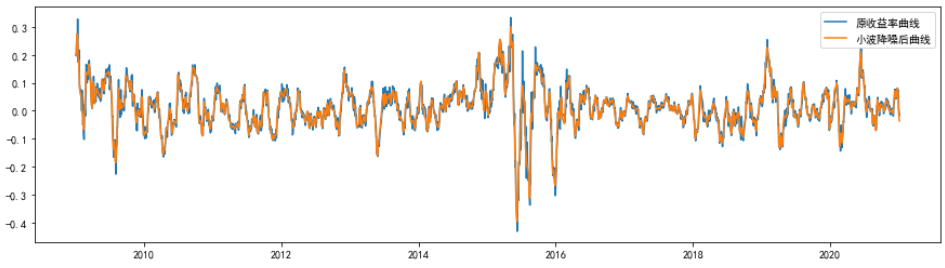


图 19：小波分解细节图

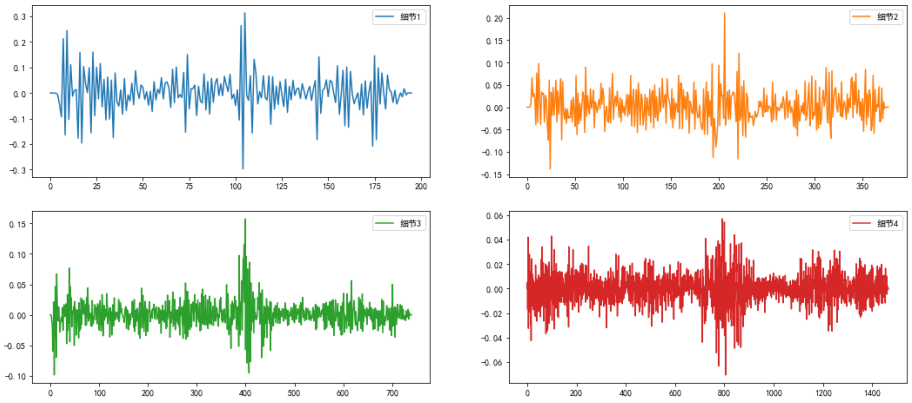


图 20：小波滤波效果指标对比

指标	未使用小波滤波	使用小波滤波
样本内准确率	0.798	0.822
样本外准确率	0.583	0.581
样本内f1 score	0.760	0.780
样本外f1 score	0.440	0.425
样本内AUC	0.731	0.762
样本外AUC	0.495	0.503
多头策略年化	0.157	0.192
多头组合夏普	0.856	1.045

4 WT-SVM 模型与网格交易策略结合的择时模型

在之前的模型中，我们得到了时间跨度约等于原数据跨度的预测值序列。图17为基于月频的预测值预测涨跌效果示意图，在一半以上的时候模型都能准确获取指数的走势。为了测试模型在实盘中的具体表现，我们将测试基于预测值的多头策略、多空策略、网格交易策略。作为对照组，我们也会同时计算指数以及仅网格交易策略的净值曲线。

4.1 基于 WT-SVM 预测值的多头策略择时模型

月频调仓，在调仓当日如果预测值为1，则下一日全仓买入指数，若当前仓位已满则不采取行动；如果预测值为-1，若当前仓位为空，则

不采取任何行动，否则清仓（如果是多空策略则全仓做空指数）。结合图21 的净值曲线以及图25 的评测指标，基于预测值的简单的多头策略就能产生数倍于基准曲线的累计收益，夏普比率也远高于指数基准曲线。然而该策略的最大回撤率高达0.44，在择时模型中我们应该在风险控制方面加以优化。其中较为常见的低风险策略为网格交易策略。

4.2 WT-SVM 模型与网格交易策略的结合

传统的网格交易策略为在基准日设定当前价格p为价值中枢，根据一定规则设定分档网格线。例如根据价格涨跌，设定p，p±1%p，p±2%p共五根网格线。当股价从上往下穿过网格线时买入一定比例仓位，反之卖出。传统的网格交易策略不依赖人为思考，仅仅只是一种机械的抗风险策略，在跟踪大盘趋势、获取高额收益率方面表现不佳。

在结合模型预测值及多头策略基础上，我们可以设计一个稳中求进的改良网格交易策略，步骤如下：

a. 月频调仓，在调仓日i获取当日的ATR值atr_i、当日的预测值$\hat{y}_i$以及收盘价p_i。

b. 如果$\hat{y}_i$=1，则启动网格交易策略，在第i日全仓买入指数。按照$p_i + w_j * atr_i$ 设置六档网格线，其中参数w为网格线距离价值中枢的atr倍数，j=[1, 2, 3, 4, 5, 6]，网格线在周期内不变。如果某日股价触及网格线，则卖出s_j比例的仓位（相较于满仓仓位），每根网格线的触发效果在一个周期内只有一次。

c. 如果$\hat{y}_i$=-1，在第i日抛出所有仓位，在周期内不进行交易。

在可调参数中，经反复对比，我们发现w_j=[-2.5，-1.5，-1，2，3.5，5]，s_j=[1，0.5，0.25，0.25，0.5，1]时模型表现较好，其中若s_j*满仓仓位>=当前仓位，则代表清仓操作。由于我们仅在看涨周期内交易，在w_j设置上我们可以对p_i之上网格线赋予更大的系数。图23展示了看涨周期内顶端和底端两条网格线与指数收盘价变化的时间序列，其网格线很好地覆盖了周期内的价格波幅。

为检验模型效果，我们还设置了不基于预测值的网格交易策略作为对照组。其区别在于在每个周期内都进行交易，调仓周期、w_j、s_j的取值都与前策略一致。

图24 和图25 展示了基准指数、多头策略，多空策略、仅网格交易策略和网格交易策略+WT-SVM预测的净值曲线以及策略评价。我们可以发现采用了网格交易策略后模型的最大回撤率显著下降，尤其是在2015年股灾期间很好地规避了风险。基于WT-SVM预测值的网格交易策略的收益率虽然略低于多头策略，但由于规避了风险，实现了远高于多头组合的夏普比率，也高于仅网格交易策略的对照组，说明无论是网格交易策略还是模型预测值，都有助于我们对万得全A指数进行择时。

图 21： 多头组合与多空组合净值曲线

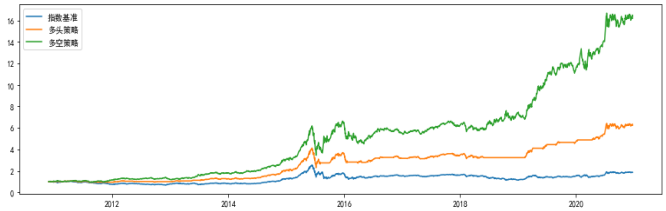


图 22：标签预测值在万得全 A 指数择时中的预测效果

图 23：网格交易策略网格线示意图

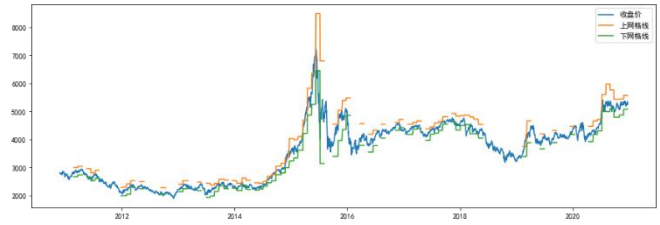


图 24：网格交易策略净值曲线表现对比图

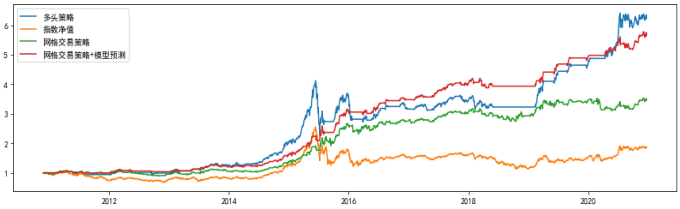


图 25：择时效果指标评测

	万得全A基准	多头组合	多空组合	网格交易策略	网格交易策略+模型预测
年化收益率	0.061676	0.191795	0.306060	0.126261	0.180652
累计收益率	1.866765	6.234014	16.198309	3.456081	5.652173
年化波动率	0.240508	0.183519	0.239802	0.150503	0.116761
夏普比率	0.256439	1.045099	1.276300	0.838928	1.547199
最大回撤率	0.559881	0.444135	0.444135	0.174296	0.110306

5 总结与对未来一个月的投资建议

本文选取了六大类约30个因子，并采用了当前使用较为广泛、准确率也较高的机器学习算法SVM和工程学界的小波降噪模型对股指驱动因素进行探究。本文首先对六大类因子进行协整性和相关性检验，在考虑了滞后因素后初步筛选出对万得全A指数收益率序列有影响的因子。为了把握指数的总体趋势，我们采用了小波变换对指数收益率曲线进行滤波处理。在模型训练以及预测上，我们采用了能及时反映短期变化的滚动SVM模型预测指数的涨跌趋势（“涨”或“跌”），基于月频对SVM的惩罚系数、gamma值、核函数进行网格搜索优化，并计算出每期的准确率、AUC、f1 score等评测指标。模型的样本外准确率达到0.581，说明模型的预测功能良好。由于模型步长与预测区间相等，我们可以得到日频预测数据。

基于预测值的月频调仓多头策略的年化收益率19.1%，夏普比率1.045，最大回撤率44.4%，表现远优于万得全A基准净值曲线，说明模型在择时上能带来数倍于基准收益率的。使用网格交易策略对多头策

略进行改进后，夏普比率提升至1.547，最大回撤率降至0.174，策略在兼顾了风险的同时达到了和多头策略相近的收益率。

不足以及改进之处：模型在一些参数的优化上未进行进一步探讨，如滚动SVM的回测区间以及预测区间。在择时策略上在预测值为-1的时段采取空仓处理，未进一步探索有可能的中短期盈利机会。在后续的研究中我们会对这两个方面着手进行改进。

对未来一个月的投资建议：基于2020年12月31日，模型的预测值为1，说明2021年1月31日的万得全A指数价格会比2020年12月31日高，建议做多。在择时方面建议采取例如网格交易策略等具有风险控制功能的策略，以应对指数短期内的不确定性。