专题报告-金融期货

量化投资策略之机器学习应用（1）

基于 SVM 模型的期货择时交易策略

报告日期：2016 年 5 月 9 日

刘宇 助理分析师（指数）

从业资格号：F3010181

Tel: 8621-63325888-3907

Email: yu.liu@orientfutures.com

★摘要：

SVM 属于监督学习算法，对于求解小样本、非线性、高维度问题具有优秀的泛化学习能力，而择时交易策略则属于利用 SVM 解决二元分类问题。构建 SVM 模型的过程可以简要概括为寻求支持向量与超平面函数间隔的最大化，从而优化求解模型参数。SVM 在求解非线性问题时使用核函数将数据映射到高维空间，以寻求超平面进行分类，同时在低维空间进行内积运算。

SVM 择时策略模型对数据进行归一化和降维处理，然后选取粒子群算法和遗传算法进行参数优化。将带有涨跌标签的普通量价数据和技术指标作为训练集数据源，将 SVM 模型训练成一个可以预测涨跌的分类器。

回测结果显示 SVM 择时策略模型预测准确率超过 50%，并且对下跌趋势的预判能力较高。模型具有良好的累计授予率与夏普比率，其中技术指标类特征量得出策略模型最大回撤控制在 10%以内。

金融期货

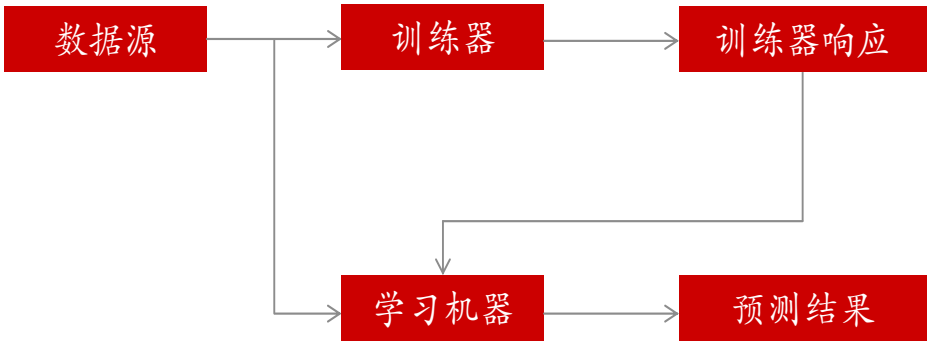
目录

图表目录

1、 机器学习简述

根据 Tom Michael Mitchell 对机器学习的定义，假设有任务 T、执行结果衡量标准 P 以及从中获取的经验值 E，计算机程序在反复执行相关任务（T）后的成绩（P）会随着经验（E）的积累而不断提高和完善，这个过程被统称为机器学习，对任务求解的途径称为学习方法。从实例中学习的模型主要包含三个部分：数据源（数据发生器）、训练器以及学习机器。数据源的特性决定训练器和学习环境，将数据$\vec{x}$输入至训练器中，返回响应值$\vec{y}$，学习机器通过观测训练集$((x_1,y_1),...(x_n,y_n))$，构造相应算法用于预测其他特定数据源 x_i 在训练器中的响应 y_i，并以预测结果$\bar{y}$适当地逼近响应值为目标。依据不同的训练数据集，机器学习可以分为监督学习、无监督学习、半监督学习以及强化学习，其中监督学习是指每一个有效的数据输入都对应一个输出（响应值）；依据不同的输出值，机器学习又可以分为分类问题（输出值离散分布）、回归问题（输出值连续分布）以及结构化问题（输出值隐性决定）。

图表 1：机器学习过程



资料来源：东证期货研究所

2、 SVM 择时策略

SVM（support vector machine）属于机器学习中的监督学习算法，以统计学习理论为基础，在最小化样本误差的同时寻求结构风险最小化，以提高学习机器的泛化能力。SVM 算法的优点在于通过参数寻优以降低泛化出错率，同时使用核函数在高维度特征空间中进行学习以解决非线性分类和回归问题。

SVM 择时模型的本质属于分类器，构建的流程包括：1. 收集数据，例如基本的历史行情信息（开盘价、收盘价、最高价以及最低价等）；2. 分析数据，对数据的所有特征项进行梳理，删除重复项以降低数据维度；3. 训练算法，使用粒子群优化算法（PSO）以及遗传算法（GA）实现参数调优；4. 测试算法，在设置止损点位的基础上对择时策略进行回测检验，统计模型准确率和收益率。总体来说，我们将 SVM 分类算法与其他参数优化算法相结合，使其通过学习带有涨跌标签的历史交易数据，成为预测未来某一时间区间涨跌方向的分类器，并据此进行为期货交易决策。

3、 线性分类问题

以篮球比赛中对球员的定位问题为例，根据2015至2016赛季NBA(美国职业篮球协会)在编450名球员的比赛场均数据，包括得分（两分球/三分球)、助攻、篮板、盖帽以及抢断等特征量，构建SVM模型，其中后场球员标签为+1，前场球员标签为-1。此模型的学习过程是通过以上球员的样本数据(训练集)，能够准确地对球员的定位进行分类。若有新球员进入联盟，我们可以将其大学和高中比赛的数据输入模型，根据预测结果对其定位进行分类。若仅考虑二维特征量，图表2中分类器可以用函数(1)表示，其余数据点的分布用函数(2)表示

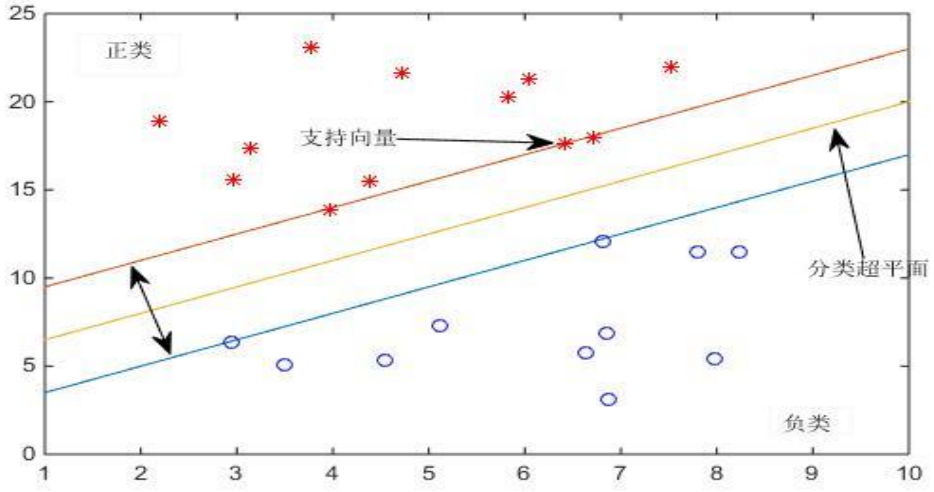
$$w^T x + b = 0 \qquad (1)$$

$$f(x) = w^T x + b \qquad (2)$$

$$h(x) = g(w^T x + b), \quad y = g(x) \in (-1, +1) \qquad (3)$$

其中x表示二维特征量$\vec{x}(x_1, x_2)$，函数(1)用来表示分类超平面，若$f(x) < 0$，则数据点在超平面左边，标签值为-1，若$f(x) < 0$，则数据点在超平面右边，标签值为+1，若$f(x) = 0$，则数据点在超平面上，不属于任何一类。函数$g(x)$类似于Logistics函数，将$f(x)$的值映射到标签集(-1,+1)上。$f(x)$能够表示数据点到超平面距离的远近，我们可以用函数间隔$\hat{\gamma} = y(w^T x + b)$的正负性来验证分类的置信度和准确性，为此我们致力于寻找能够使正负类中函数间隔最小值最大化的超平面。

图表2：支持向量机二维示图



资料来源：东证期货研究所

图2中两个支撑着中间间隙的超平面到分类超平面的距离相等，即求得的最大函数间隔$\hat{\gamma}$，在这两个支撑超平面上的点称为支持向量，它们满足$y(w^T x + b) = \hat{\gamma}$；对于不是支持向量的数据点，则满足$y(w^T x + b) > \hat{\gamma}$；若$y(w^T x + b) < 0$，则表示数据点分类错误。SVM模型可以通过参数寻优找到分类超平面，以最大限度分隔正负类数据并使支撑超平

面之间的距离最大化。

4、 非线性分类问题

鉴于 SVM 模型更多被运用在解决小样本的非线性分类问题，可以通过核函数和惩罚因子来推广上述的线性支持向量机。在核函数能够计算对应输入特征数据的内积的前提下，可以通过使用恰当的核函数来替代内积，在将非线性的数据映射到高维空间的同时不增加需调参数的个数，从而在高维空间中寻找响应的超平以进行分类。根据问题性质和数据类型，我们选择不同类型的核函数。常用的核函数有如下几类：

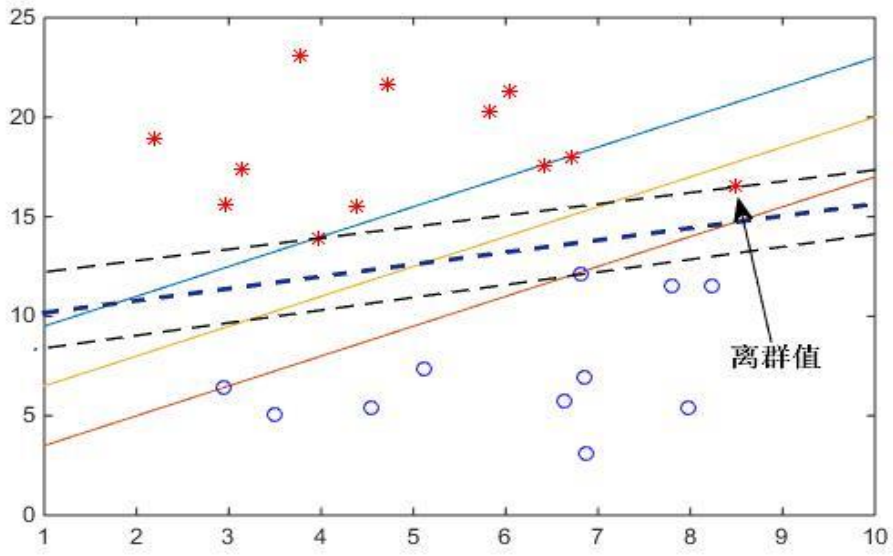
1) 多项式核：$\mathrm{k}(x_i, x_j) = (\langle x_i, x_j \rangle + R)^d$；

2) 径向基核：$\mathrm{k}(x_i, x_j) = exp\{-\gamma |x_i - x_j|^2\}$

3) 线性核：$\mathrm{k}(x_i, x_j) = \langle x_i, x_j \rangle$

4) Sigmoid 核：$\mathrm{k}(x_i, x_j) = \tan a(x_i, x_j) + b$

图表 3：干扰最优分类超平面的离群值



资料来源：东证期货研究所

虽然通过映射将原始数据映射到高维空间之后，线性分隔的可操作性增加，但是因为数据特性的原因，例如噪音，而造成与均值或正常位置偏离度较大的数据点难以准确分类。一般来说，超平面由少数几个支持向量组成，如果支持向量里存在离群值，则会对非线性分类器的效果产生负面影响。如图 3 所示，若不考虑被圈起的数据点则分类超平面与支撑超平面为三条实线，若考虑此离群值则超平面为三条间隔变小的虚线，此时其他测试集的分类可信度将会降低，并且随着离群值往左下方偏离，分类超平面的构造将失效。

为获得最佳分类效果，SVM 模型允许数据点在一定程度内偏离超平面，通过引入松弛变量ξ和惩罚因子 C，使得$C\sum_{i=1}^{n}\xi_i$最小，其中ξ表示数据点允许偏离的函数间隔的量，C 用于控制分类函数中函数间隔最优与数据点总体偏差量最小之间的权重。例如，根据历史球员比赛数据，后场球员的场均抢断和助攻数据要优于前场球员，而盖帽和篮板数据呈现相对劣势，若出现盖帽数据优于联盟平均值的后场球员，则会干扰 SVM 模型的判断。惩罚因子的作用在于消除离群值的影响。

5、 特征量筛选方案

数据分析中对原始数据源的清理会极大提高优化算法的效率，例如数据源中不同特征向量反映同类特性时对其进行合并。在 SVM 择时策略模型中我们采用 PCA(Principle Component Analysis)算法对数据源进行降维处理，将基础交易数据项（开盘价、收盘价、最高价、最低价等）与合成数据项（MACD、RSI 等）中存在相关性的特征进行剔除，并最大化保留具有代表性的特征项。

PCA 可以分为以下步骤：

1) 将原始数据按维度（特征量）进行标准化处理，减去均值并处以标准差；

2) 计算协方差矩阵的特征向量和特征值，选取贡献度达到 90%的特征值所对应的特征向量；

3) 使用选取的特征向量将标准化矩阵转换到新空间，产生新的样本数据，降维完成。

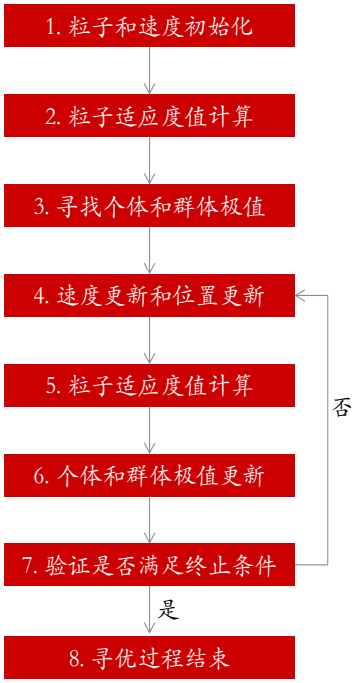
我们使用经过处理的数据样本构建模型，预测变量的个数缩减使得计算过程简化，变量之间相互独立但可以最大化反映原始特征信息。

6、 参数寻优过程

参数寻优的过程主要针对 SVM 模型中核函数参数和惩罚因子，我们选择两种方法来对核函数中γ以及惩罚因子 C 进行动态寻优，分别是粒子群算法和遗传算法。

PSO 算法从生物种群行为中得到启发，采用跟随式寻优步骤，每个粒子代表一个实验个体，也对应一个潜在最优解。适应度函数决定粒子的适应度值，而粒子的速度决定粒子移动的方向和距离，速度随自身及其他粒子的移动情况而动态调整，从而实现个体在可解空间中的寻优。

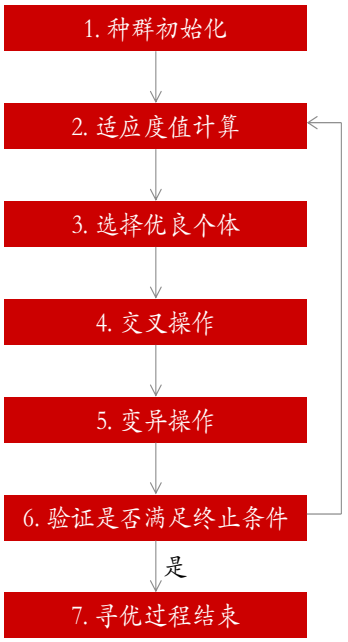
图表 4：粒子群优化算法流程



1. 在可行解空间初始化一群粒子，每个粒子代表极值优化问题的一个潜在最优解，类似于实验个体。
2. 适应度由适应度函数计算得到，其值大小表示粒子的优劣。
3. 个体和群体极值分别代表个体粒子和种群中所有粒子搜索到的适应度最优位置。
4. 粒子每更新一次位置就计算一次适应度值，并且通过比较新粒子适应度值以及个体极值、群体极值的适应度值更新个体极值和群体极值位置。

资料来源：东证期货研究所

图表 5：遗传算法流程



1. 通过编码将问题可行解表示成遗传空间个体。
2. 适应度由适应度函数计算得到，用于区分群体中个体的好坏。
3. 选择操作从旧群体中以一定概率选择优良个体组成新的种群，以繁殖得到下一代。个体被选中的概率跟适应度值呈正相关。
4. 交叉操作从种群中随机选择两个个体，通过交换组合，将父代的优秀特征遗传给子代，产生新个体。
5. 变异操作的主要目的是维持种群多样性，从种群中随机选择个体，并针对其某一特征进行变异以产生更优秀的个体。

资料来源：东证期货研究所

遗传算法借鉴生物界自然选择和遗传机制，从随机产生的初始解开始搜索，通过选择、交叉、变异操作逐步迭代以产生新解。同 PSO 算法，群体中的每个个体代表一个潜在最优解，其好坏由适应度来衡量，根据适应度从上一代中选择一定数量的优秀个体，通过交叉、变异形成下一代群体。遗传算法的优点在于将问题编码后进行优化，而不针对参数本身，从而不受函数约束条件限制，采用并行搜索的方式也可以减少陷入局部最优的可能性。

粒子群算法比遗传算法具有更高效的信息共享机制，更新群体极值使得信息实现全局范围共享，但遗传算法通过交叉和变异拥有比粒子群算法更有效的逃离局部最优解的概率。所以我们通过 SVM 模型中分别使用两种算法进行参数寻优，最终选择交叉验证结果中准确率最高的算法以及相应参数。

7、 SVM 模型实践

7.1、 特征量选择

我们旨在构建一个日度级别的交易策略，所以选取能够快速反应市场信息的交易数据作为 SVM 的测试集。选择的样本属性集合分为两类，特征量集合 A 包括基本交易数据，特征量集合 B 包括技术指标，在后续模型构建中将对这两类集合进行对比。

图表 6：交易数据特征量集合 A

集合 A 的特征量		
收盘价	最低价	最高价
昨日成交额	前 5 日平均涨跌幅	前 20 日平均涨跌幅
当日涨跌幅	前 5 日平均成交金额	前 20 日平均持仓量
当日成交额	前 5 日平均持仓量	前 20 日平均成交金额
昨日涨跌幅		

资料来源：东证期货研究所

图表 7：技术指标特征量集合 B

集合 B 的特征量	
MACD	指数平滑异同平均
MTM	动力指标，反映价格变动的能量和速度
PRICEOSC	价格震荡指标，反映收盘价的周期变动
DMI	趋向指标
VR	成交量比率，通过成交量比值反映市场买卖情绪
RSI	相对强弱指标，反映市场买卖强度变化
KDJ	随机指标，通过最高价、最低价以及收盘价反映价格趋势的强弱
WR	威廉指标，反映收盘价的趋势分布
VOSC	成交量震荡指标，反映成交量长短期运动趋势
WAD	威廉聚散指标
CVLT	佳庆离散指标，反映价格的波动率

资料来源：东证期货研究所

7.2、 SVM 择时模型流程

1) 选取过去 n 天的训练集数据，将其进行归一化，并使用 PCA 进行降维处理，得到新的训练集；
2) 使用遗传算法和粒子群优化算法选取最佳参数，获取的标准依据各自算法中适应度值以及交叉验证的准确率来判断，并从两个算法中选择结果较优的参数；
3) 将得到的最优参数代入 SVM 模型，选择出构建模型的最佳时间窗口长度，判断的标准为滑窗内的最佳预测准确率；
4) 将滑动窗口以及最优参数代入 SVM 模型，使用当日的训练数据进行分类，预测明日的上涨或下跌，并根据预测结果设置合适的止损，开盘建仓，收盘平仓。

7.3、 SVM 择时策略实证分析

首先针对 SVM 择时策略回测，数据源、交易参数以及模型参数设定如下：

图表 8：择时策略要素

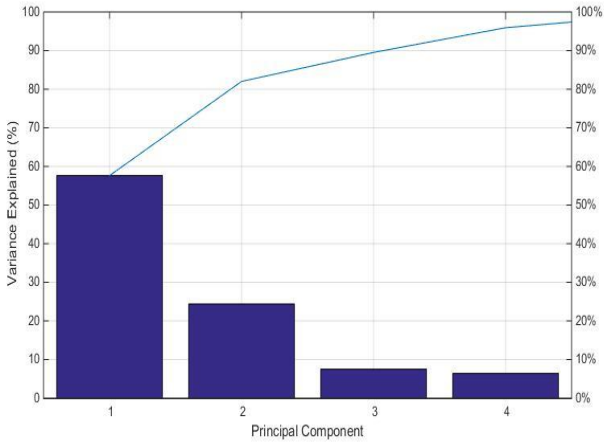
SVM 择时策略	
数据标的	沪深 300 股指期货当月合约
时间区间	2011 年 1 月 4 日至 2016 年 4 月 29 日
策略	根据模型输出值信号，进行开盘建仓，收盘平仓的被动策略
持仓量	1 手
日内止损线	30 个点
遗传算法参数设定	最大迭代次数 200，种群规模 20，交叉验证折数 5，惩罚系数取值范围 0-100，核函数γ的取值范围 0-1000
粒子群优化算法参数设定	最大迭代次数 200，种群规模 20，交叉验证折数 3，惩罚系数取值范围 0-100，核函数γ的取值范围 0-1000

资料来源：东证期货研究所

通过对数据归一化处理，我们将原本规模或者单位不同的特征量数据统一转换至 0 到 1 的区间内，避免出现某一项特征量自身绝对数值过大而影响分类效果。降维处理后特征量的维数降低，但保留的维数已经能够覆盖原始数据 90%的方差特征。

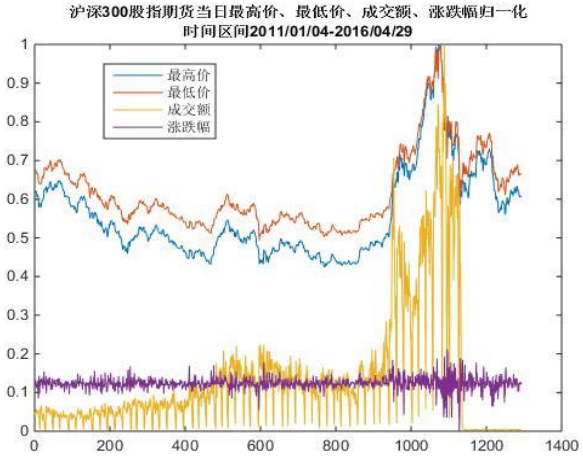
对比粒子群优化算法和遗传算法优化参数的结果，我们可以看出遗传算法在特征量结合 A 和 B 的测试中，交叉验证的准确率相对较高，于是我们选择由遗传算法优化得到的参数。使用遗传算法和技术指标数据源可以达到接近 65%的交叉验证准确率。

图表 9：主成分分析降维结果（基础价量）



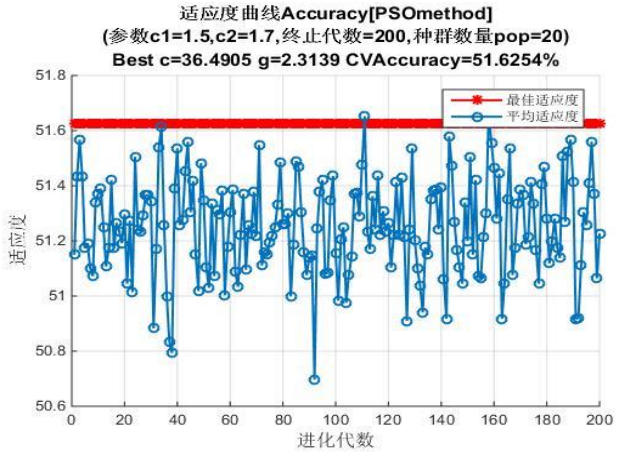
资料来源：东证期货研究所

图表 10：沪深 300 股指期货交易数据归一化



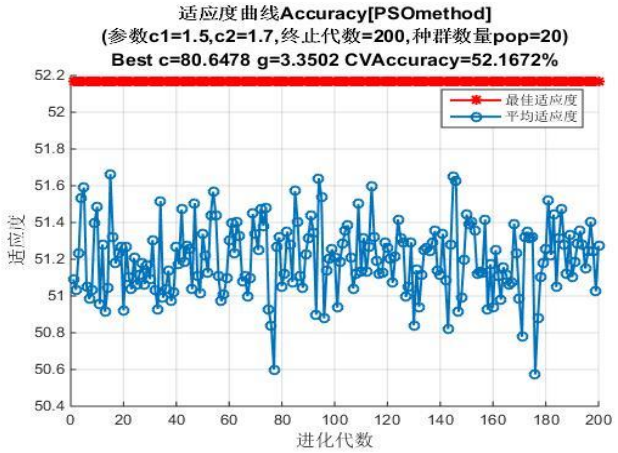
资料来源：Wind，东证期货研究所

图表 11：PSO 算法参数适应度曲线（基础价量）



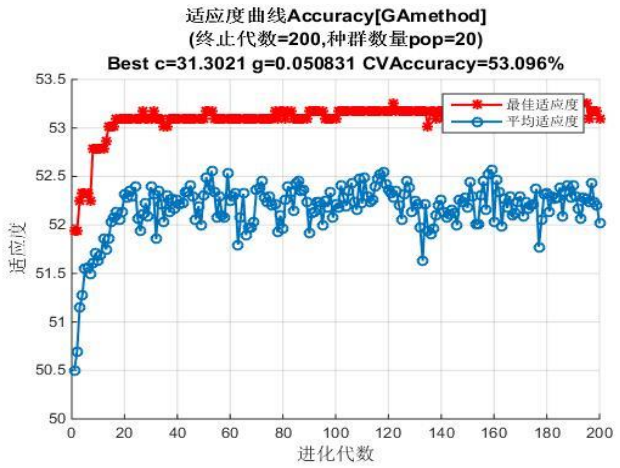
资料来源：Wind，东证期货研究所

图表 12：PSO 算法参数适应度曲线（技术指标）



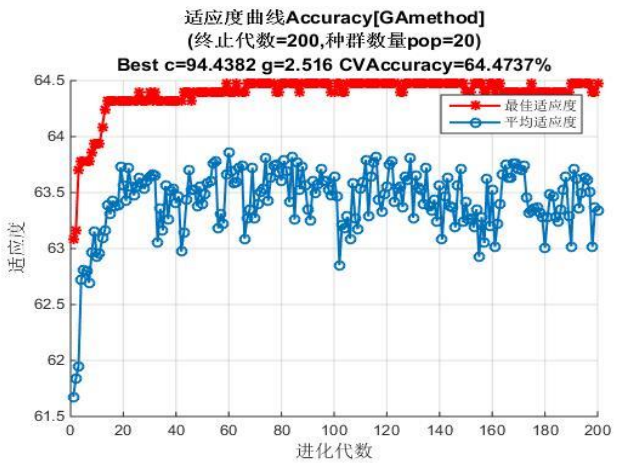
资料来源：Wind，东证期货研究所

图表 13：GA 算法参数适应度曲线（基础价量数据）



资料来源：Wind，东证期货研究所

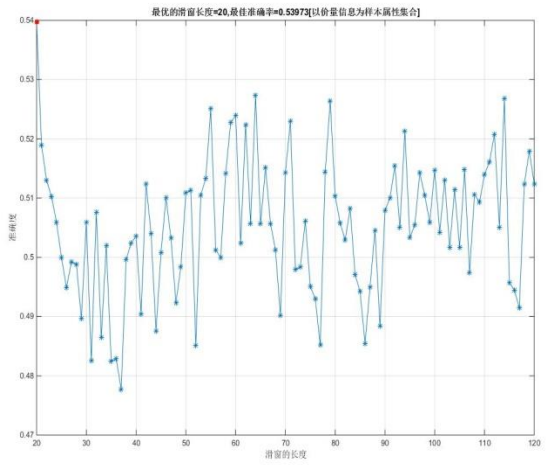
图表 14：GA 算法参数适应度曲线（技术指标）



资料来源：Wind，东证期货研究所

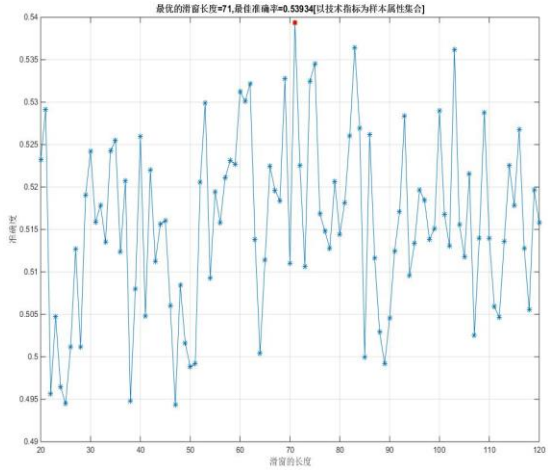
在得到的最佳窗口长度基础上，我们可以使用得到的 SVM 模型预测交易信号，并对比实际买卖信号。

图表 15：以基础价量为样本的最优滑窗长度



资料来源：东证期货研究所

图表 16：以技术指标为样本的最优滑窗长度



资料来源：东证期货研究所

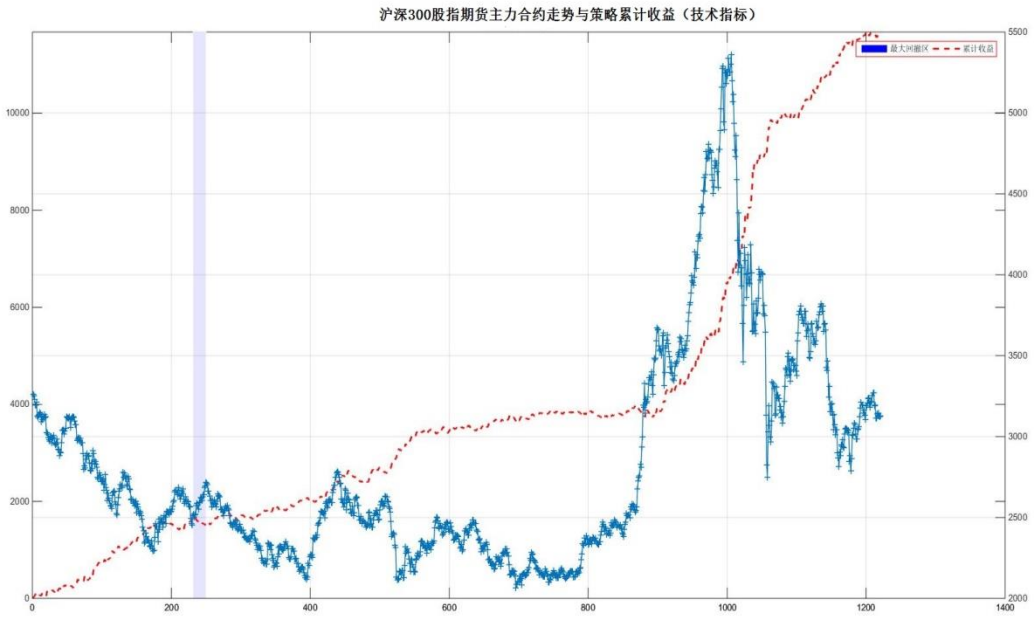
7.4、 SVM 择时策略回测分析

图表 17：特征量集合 A 与集合 B

指标	集合 A	集合 B
预测精度	53.97%	53.93%
最大回撤	0.22924	0.072618
上涨预测准确度	52.35%（301/575）	52.59%（244/464）
下跌预测准确度	55.32%（385/696）	54.76%（414/756）
累计收益率	424.24%	350.79%
最大连盈次数	10	9
最大连亏次数	7	10
年化夏普比率	2.8858	3.2774
信息比率	0.092964	0.098278
最佳滑动窗口	20	71

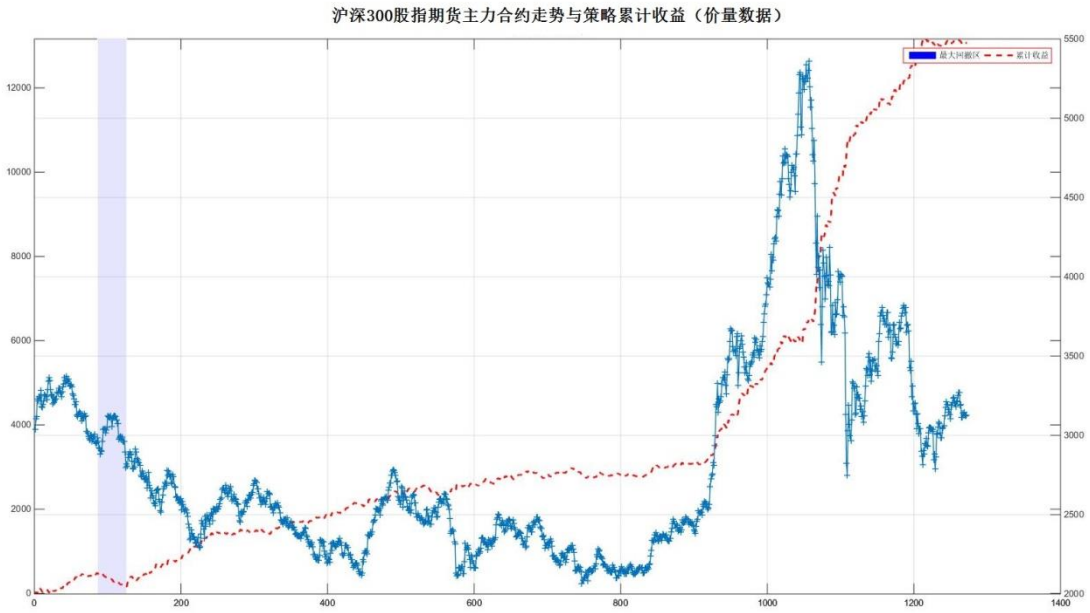
资料来源：东证期货研究所

图表 18：沪深 300 股指期货主力合约走势与策略累计收益（技术指标）



资料来源：Wind，东证期货研究所

图表 19：沪深 300 股指期货主力合约走势与策略累计收益（价量数据）



资料来源：Wind，东证期货研究所

根据回测数据我们可以看到，价量数据集合 A 与技术指标集合 B 在预测下跌时胜率均比预测上涨时要更精确，并且策略的总体预测精度均超过 50%。结合最大回撤、累计收益率、夏普比率以及滑动窗口，可以看出集合 A 对市场的消息反映更快，收益率累计高，但波动率也相对较高，相比而言集合 B 更加稳健，拥有低回撤的同时也有较强的夏普比率和信息比率。双方的连盈、连亏次数相当，反映两个特征集合对市场趋势的把握能力相同。

8、 总结

通过 SVM 择时策略模型，我们通过数据清理与参数优化两个重要步骤来调试 SVM 模型，结合窗口滑动进行动态测试交易，在明确交易信号、止损标准清晰、建仓不考虑冲击成本、流动性因素的基础上，构建日度开盘建仓、收盘平仓的被动策略，保证一定程度的精度预测，同时也获得良好的收益和风险表现。但仍有一下几个方面需要改善：

1) 特征量是训练 SVM 模型的根基，是机器学习的“参考书”。继续挖掘能反映市场情绪、量价关系、动能转换以及趋势变化的优秀特征量是必不可少的。另外，日度交易策略也需要测量在不同的市场趋势与环境中哪些特征指标是最能有效反映市场信息。

2) 数据频率的升级可以提升数据源信息丰度，但遗传算法的迭代使得目前算法仍只能适用于日度级别。如何优化算法以进一步优化计算效率是研究的重点。

3) PCA 线性正交降维的劣势可以得到解决，可以用更少的维度来反映相同量的特征信

息。

4) 开发操作简便、指令完善以及交互体验优良的模型参数调试 GUI（图形用户界面），使数据清理优良程度以及参数优化结果能以图形的方式直接展示，方便直观地了解模型构建与调试过程。

5) 丰富交易策略，包括开发更灵活、指令清晰的择时模型，以及将 SVM 分类回归功能运用至市场波动率研究和选股策略。

走势评级	短期（1-3 个月）	中期（3-6 个月）	长期（6-12 个月）
强烈看涨	上涨 15%以上	上涨 15%以上	上涨 15%以上
看涨	上涨 5-15%	上涨 5-15%	上涨 5-15%
震荡	振幅-5%-+5%	振幅-5%-+5%	振幅-5%-+5%
看跌	下跌 5-15%	下跌 5-15%	下跌 5-15%
强烈看跌	下跌 15%以上	下跌 15%以上	下跌 15%以上

上海东证期货有限公司

上海东证期货有限公司（简称东证期货）是东方证券股份有限公司全资子公司，注册资本达5亿元，系国内四家期货交易所的结算会员。

东证期货专注于金融期货和商品期货的研究与服务，提供权威、及时的研发产品服务和投资策略；专注于信息技术的创新，创建安全、快捷的交易通道，开发多样化、个性化的交易系统；专注于构筑全面的风险管理和客户服务平台。

东证期货管理团队管理经验丰富，业绩出众，在业内享有盛誉。人才管理及激励机制完善，公司拥有硕士学历以上人员占比30%，具有海外证券和期货经历的高端人才占比10%。

2010年，东证期货发展迅猛，成绩斐然，成为业内进步最快、最受瞩目的期货公司之一。2011年初，东证期货荣获2010年度中国金融期货交易所年度会员金奖，同时获投资者教育奖、客户管理奖、技术管理奖和功能发挥奖等四项单项大奖；荣获上海期货交易所优胜会员第七名，铜、橡胶和燃料油三项企业服务奖；荣获大连商品交易所优秀会员第九名；东证期货研究所荣获大连商品交易所、和讯网第二届全国“十大期货研发团队”农产品团队全国第二名、化工团队全国第五名；荣获郑州商品交易所行业进步奖等。

东证期货全年无风险事故，充分体现了公司稳健经营，稳步发展的经营宗旨。

分析师承诺

刘宇

本人具有中国期货业协会授予的期货执业资格或相当的专业胜任能力，以勤勉的职业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接接收到任何形式的报酬。

免责声明

东证期货研究所

地 址： 上海市中山南路 318 号东方国际金融广场 2 号楼 39 楼
联系人： 周冰沁
电 话： 8621-63325888-3914
传 真： 8621-33315862
网 址： www.orientfutures.com
Email： research@orientfutures.com