文章编号:1003－207(2023)12－0096－11
DOI:10.16381/j.cnki.issn1003－207x.2021.2308

基于机器学习预测股票收益率的两步骤M-SV投资组合优化

张 鹏[1],党世力[1],黄梅雨[2],李璟欣[1]

(1.华南师范大学经济与管理学院,广东 广州 510006;
2.华中科技大学管理学院,湖北 武汉 430074)

摘 要:由于准确预测股票收益序列能够提高投资组合优化模型的表现,相比于传统的计量经济预测模型,机器学习在处理非线性和非平稳特征的问题上更具优势。因此,本文提出了一种基于机器学习方法的两步骤多元化投资组合优化模型。具体而言,该模型包括以下两个步骤:步骤1是股票选择,即通过机器学习方法极端梯度提升法(extreme gradient boosting,XGBoost)、支持向量回归(support vector regression,SVR)、K近邻算法(K－nearest neighbor,KNN)选择具有较高预测收益率的股票,并对模型进行评估和选择。步骤2是投资组合优化,在考虑交易成本、上下界约束的现实约束条件下,采用均值－下半方差(mean semi－variance,M－SV)模型、均值－方差模型和等比例模型确定所选股票的投资比例。最后,以沪深300指数成分股作为研究样本,实证结果表明,XGBoost＋M－SV模型在收益和风险指标上均优于其他模型和沪深300指数。

关键词:均值－下半方差;旋转算法;机器学习;股票收益预测
中图分类号:O159 **文献标识码**:A

1 引言

1952年Markowitz[1]提出了均值－方差理论,标志着现代投资组合理论的开端,该投资组合模型以均值和方差来度量投资组合的收益和风险。但方差不能区别偏离均值的正负值。所以,有些学者认为,只有当实际收益低于期望收益率时,才存在损失。针对该问题,Markowitz[2]用半方差度量投资组合的风险,并提出了均值－下半方差(mean semi－variance,M－SV)模型。随后,国内外学者在M－SV模型的基础上提出了很多拓展模型。如Zhang Peng和Dang Shili[3]提出了具有加权上下界可容许M－SV模型。姚海洋等[4]利用下半方差和下半偏差作为风险度量,提出了不允许卖空的均值－下方风险投资组合模型。而Yan Wei等[5]将M－SV模型拓展到多阶段情形。张鹏等[6]提出了多阶段均值－标准下半方差模糊投资组合模型。以上拓展模型进一步丰富了现代投资组合理论,并为学者提供了更多的研究视角。然而,大多相关研究更关注如何改进均值－方差(mean variance,MV)模型,忽略了股票的初步选择,即最优投资组合形成前的阶段。在实际投资过程中,优质的股票投入是最优投资组合形成的关键保证,而股票的预期收益率会影响优质股票的选择。因此,股票收益预测是投资决策过程中的一个关键因素[7]。

由于股票市场本质上是动态的、复杂的、进化的、非线性的,通过正确选择股票来构建投资策略是一项具有挑战性的任务。到目前为止,关于股票价格预测方面的研究主要集中在两个方面:一是计量经济学模型,如线性自回归、自回归条件异方差和差分自回归移动平均模型,这些统计模型要求金融时间序列满足一定的假设以保证结果的可靠性;二是机器学习和深度学习模型,如支持向量机(support vector machine,SVM)、极端梯度提升算法(extreme gradient boosting,XGBoost)和长短期记忆人工神经网络(long－short term memory,LSTM)等,这些模型与计量经济学模型相比,没有严格的假设条件,且具有更强的处理非线性和非平稳特征问题的能力[8]。例如,Lin Chiming等[9]使用Elman神经网络来预测股票价格,实证结果表明,动态投资组合选择模型优于向量自回归模型。Thenmozhi和Sarath Chand[10]使用SVM算法研究了1999—

收稿日期:2021－11－08;**修订日期**:2022－03－04
基金项目:国家自然科学基金资助项目(71271161);广东省社会科学项目(GD19CGL32)
通讯作者简介:张鹏(1975—),男(汉族),江西吉安人,华南师范大学经济与管理学院,教授,博士生导师,研究方向:投资组合、金融工程,E-mail:zhangpeng300478@aliyun.com.

2011年由全球市场指数组成的样本,结果表明,SVM算法的预测能力超过回归模型和技术分析指标。因此,很多国内外学者尝试使用机器学习和深度学习来预测资产收益和选择投资标的。如Krauss[11]通过深度神经网络、决策树、随机森林算法等方法来预测标普500指数的涨跌情况。Freitas等[12]提出了一种基于自回归移动参考神经网络预测股票收益的投资组合模型,实证结果表明,该模型优于原始投资组合模型及基准指数。

之后,许多学者将深度学习和机器学习预测方法与投资组合优化方法相结合,如Day和Lin Jianting[13]开发了具有不同机器学习和深度学习预测方法的人工智能顾问,并利用投资组合模型的预测结果来帮组投资者做出决策。Fischer和Krauss[14]、Lee和Yoo[15]基于神经网络预测模型构建投资组合,实证结果表明,该投资策略能够获得较高的收益。虽然深度学习方法在股票预测的准确性上显示出一定的优势,但是这些方法需要大量的数据和更多的时间[16]。因此,许多学者仍选用机器学习作为预测模型,如Paiva等[17]构建了SVM＋MV的投资组合模型,并进行算法训练和投资组合优化。而SVR算法在原理上与SVM算法相同,最终目标是最小化误差,也被广泛用于股票收益率的回归预测模型中,并与投资组合优化方法相结合(Ayala等[18];Ma Yilin等[19])。此外,Chen Wei和Guestrin[20]提出了一种新的机器学习模型——XGBoost算法,该优化算法可以有效缓解过拟合的问题,而且运行速度快,准确度高。Hongjoong[21]、Chen Wei等[22]进一步采用XGBoost算法对股票进行预测,并与MV模型相结合,实证结果不仅展现了该算法较强的预测能力,也验证了投资组合决策优异的样本外表现。

基于以上研究发现,准确预测股票收益序列能够提高投资组合优化模型的表现,从而凸显了股票收益预测在构建投资组合模型优化中的优势。然而,研究大部分是将预测模型与经典的MV模型相结合,主要考虑了预测收益率对投资决策的影响,而忽略了风险度量这一因素。因此,本文采用下半方差的方法来度量投资组合的风险,克服方差不能区别偏离均值的正负值的局限性。此外,由于XGBoost算法和SVR算法在股票预测中频繁使用且表现出了较强的预测能力,本文采用这些模型对股票的收益率预测,并且考虑到K近邻(KNN)算法具有精度高、对异常值不敏感、无数据输入假定、实现简单等优点,将其作为机器学习预测能力评价的一个参考模型。具体地,本文沿着股票收益预测与投资组合模型优化相结合的发展方向,提出了一种基于机器学习预测方法的两步骤投资组合优化方法。该模型包括两个步骤:股票收益预测和投资组合优化。在步骤1中,本文分别采用XGBoost算法、SVR算法和KNN算法预测$t+1$天股票的收益率,并选择股票收益率排在前k位的股票进行投资。在步骤2中,本文考虑交易成本、上下界约束的现实约束条件下,运用M－SV模型、MV模型和等比例(1/N)模型确定所选股票的最优投资比例。接下来,本文采用2013年1月1日—2021年12月31日期间沪深300指数(China Securities 300 Index,CSI300)成分股进行实证分析,以探究如下问题:一是XGBoost算法、SVR算法和KNN算法中,哪种算法会取得最好的预测效果?二是基于机器学习的投资组合模型的样本外表现是否能超过随机选股模型和沪深300指数?三是M－SV模型的累积收益是否优于MV模型和1/N模型的?

2　预测方法

本部分简单介绍了所采用的机器学习预测模型,即XGBoost算法、SVR算法和KNN算法。

2.1　极端梯度提升算法

本文引用华盛顿大学陈天奇博士提出的XGBoost[20]算法来预测股票的收益。XGBoost算法如下:

$$\hat{y}_i=\sum_{t=1}^{T}f_t(x_i),f_t\in F \tag{1}$$

其中,$\hat{y}_i$表示对i个样本的预测值,f_t属于F集合范围内,$f_t(x_i)$表示通过第t棵树对i个样本进行预期。

目标函数可以表示为:

$$Obj^{(t)}=\sum_{i=1}^{n}l(y_i,\hat{y}_i^t)+\sum_{t=1}^{t}\Omega(f_i)=$$

$$\sum_{i=1}^{n}l(y_i,\hat{y}_i^{t-1}+f_t(x_i))+\Omega(f_t)+cons\tan t \tag{2}$$

其中,$l(y_i,\hat{y}_i)$表示损失函数,$\Omega(f_t)$表示第t棵树的复杂程度。

惩罚Ω计算为:

$$\Omega(f_t)=\gamma T+\frac{1}{2}\lambda\|\omega_j\|^2 \tag{3}$$

其中,γ和λ为超参数;T为叶子节点个数;ω_j为叶子节点值。

对于样本x_i首先初始化,假定第0棵树为$f_0(x_i)$,预测值为$\hat{y}_i^{(0)}=0$,在第0棵树的基础上得到第1棵树的预测值$\hat{y}_i^{(1)}$,依次类推,具体如下:

$$\hat{y}_i^{(0)}=0 \tag{4}$$

$$\hat{y}_i^{(1)}=\hat{y}_i^{(0)}+f_1(x_i) \tag{5}$$

$$\hat{y}_i^{(t)}=\sum_{i=1}^{t-1}f_i(x_i)+f_t(x_i)=\hat{y}_i^{(t-1)}+f_t(x_i) \tag{6}$$

在训练第t棵树决策树时,则XGBoost的目标函数可以转为:

$$Obj^{(t)}=\sum_{i=1}^{n}[2(\hat{y}_i^{t-1}-y_i)f_t(x_i)+f_t(x_i)^2]$$
$$\Omega(f_k)+cons\tan t \tag{7}$$

接下来,对(2)式进行二阶泰勒展开,对目标函数进行近似转换为:

$$Obj^{(t)}\approx\sum_{i=1}^{n}\left[g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)\right]+\Omega(f_t) \tag{8}$$

其中,$g_i=\partial_{\hat{y}^{t-1}}l(y_i,\hat{y}^{t-1})$,$h_i=\partial^2_{\hat{y}^{t-1}}l(y_i,\hat{y}^{t-1})$。

2.2 支持向量回归算法

对于给定的训练样本$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$,$y_i\in R$,通过构造类似(9)式的回归模型,使得$f(x)$与$y$尽可能接近。

$$f(x)=\omega^T x+b \tag{9}$$

其中,ω^T和b是待确定的模型参数。

假设SVR[23]算法能容忍$f(x)$与y之间最多有ϵ的偏差,即仅当$f(x)$与y之间的差别绝对值大于ϵ时才计算损失,这相当于以$f(x)$为中心,构建了一个宽度为2ϵ的间隔带,若训练样本落入此间隔带中,则认为是被正确预测的。于是,SVR算法可形式化为:

$$\min_{\omega,b}\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{m}\ell_\epsilon(f(x_i)-y_i) \tag{10}$$

其中,C为正则化常数,ℓ_ϵ是ϵ的不敏感损失函数

$$\ell_\epsilon(z)=\begin{cases}0, & |z|\leqslant\epsilon\\ |z|-\epsilon, & \text{其他}\end{cases} \tag{11}$$

引入松弛变量ξ_i和$\hat{\xi}_i$,可将(10)式重写为:

$$\min_{\omega,b,\xi_i,\hat{\xi}_i}\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{m}\ell_\epsilon(\xi_i+\hat{\xi}_i)$$
$$\text{s.t.}\begin{cases}f(x_i)-y_i\leqslant\epsilon+\xi_i\\ y_i-f(x_i)\leqslant\epsilon+\hat{\xi}_i\\ \xi_i\geqslant 0,\hat{\xi}_i\geqslant 0,i=1,2,\cdots,m\end{cases} \tag{12}$$

其中,ξ_i和$\hat{\xi}_i$为松弛变量,C为正则化常数,ω为权重的系数向量。

2.3 K近邻算法

KNN算法[24]可用于回归。具体是找到未知样本的k个最近邻近,根据距离与权重成反比的原则,为各个最近邻居的属性赋予不同的加权值,得到未知样本的属性。

设特征空间χ为p维实数向量空间R^P,$x_i,x_j\in\chi$,$x_i=(x_{i1},x_{i2},\cdots,x_{ip})^T$,$x_j=(x_{j1},x_{j2},\cdots,x_{jp})^T$,$x_i$,$x_j$的$L_r$距离定义为:

$$L_r(x_i,x_j)=\left(\sum_{i=1}^{n}|x_{il}-x_{jl}|^p\right)^{1/p},p\geqslant 1 \tag{13}$$

计算未知样本与邻近样本的距离有多种方法。其中,当$r=2$时,称为欧式距离;当$r=1$时,称为曼哈顿距离;当$r\to\infty$时,称为切比雪夫距离。

根据给定的距离度量,在训练集T中找出与x最近邻的k个点,涵盖这k个点的领域记为$N_k(x)$,然后得到KNN算法回归函数为:

$$h(x)=\frac{1}{k}\sum_{i\in N_k(x)}Y_i \tag{14}$$

3 均值-下半方差投资组合模型

3.1 问题描述与符号说明

假设投资者将初始财富投资于金融市场中的n种风险资产。为了便于表述,假设R_i是风险资产i的随机收益,r_i是R_i的数学期望,$r_i=E(R_i)$;协方差矩阵是$G=(\sigma_{ij})_{n\times n}$,$\sigma_{ij}=COV(r_i,r_j)$,$i,j=1,2,\cdots,n$;$x_i$是风险资产$i$的投资比例,$x=(x_1,x_2,\cdots,x_n)'$,$x_1+x_2+\cdots+x_n=1$;$r_0$为投资者的预期收益;$x_{i0}$是风险资产投资的初始比例;$r_p$是投资组合$x$的期望收益率;$r_N$是投资组合$x$的净收益率;$r_f$是无风险借贷利率;$u_i$是$x_i$的上界约束;$c_i$是风险资产$i$的单位交易成本;文中(′)表示矩阵的转置。

3.2 收益、风险和约束条件

假设风险资产i的不确定收益率为r_i,则投资组合x的期望收益率为:

$$r_p=\sum_{i=1}^{n}r_i x_i \tag{15}$$

设投资组合x的交易成本函数是V形函数,初始投资比例为$x=(x_{01},x_{02},\cdots,x_{0n})'$,则投资组合$x$的总交易成本为:

$$C_i=\sum_{i=1}^{n}c_i|x_i-x_{0i}| \tag{16}$$

因此,投资组合x的净收益为:

$$r_N=\sum_{i=1}^{n}r_i x_i-\sum_{i=1}^{n}c_i|x_i-x_{0i}| \tag{17}$$

令$r_i^+=\max(r_i-\bar{r}_i,0)$,$r_i^-=\min(r_i-\bar{r}_i,0)$,其中,$r_i$为第$i$种资产的收益率,$\bar{r}_i=E(r_i)$,以$\bar{r}_i$为基准,将$r_i$的上、下波动部分分离,$r_i^+$和$r_i^-$对应的上、下半方差分别记为$\sigma^2(r_i)^+$和$\sigma^2(r_i)^-$,则$\sigma^2(r_i)=\sigma^2(r_i)^++\sigma^2(r_i)^-$。

其中,$\sigma^2(r_i)^+=E((r^+)^2)$,$\sigma^2(r_i)^-=E((r^-)^2)$

定义1[25] 设r_i和r_j分别为风险资产i和j的收益率,$\sigma^2(r_i)$和$\sigma^2(r_j)$为方差,$\sigma^2(r_i)^+$、$\sigma^2(r_i)^-$、$\sigma^2(r_j)^+$和$\sigma^2(r_j)^-$为半方差,则其上半协方差和下半协方差分别为:

$$G^+=COV(r_i,r_j)^+=\frac{1}{\sigma^2(r_i)\sigma^2(r_j)}\left(\sigma^2(r_i)^+\sigma^2(r_j)^+ +\frac{\sigma^2(r_i)^-\sigma^2(r_j)^+}{2}+\frac{\sigma^2(r_i)^+\sigma^2(r_j)^-}{2}\right)COV(r_i,r_j) \tag{18}$$

$$G^{-}=COV(r_i,r_j)^{-}=\frac{1}{\sigma^2(r_i)\sigma^2(r_j)}\left(\sigma^2(r_i)^{-}\sigma^2(r_j)^{-}+\frac{\sigma^2(r_i)^{-}\sigma^2(r_j)^{+}}{2}+\frac{\sigma^2(r_i)^{+}\sigma^2(r_j)^{-}}{2}\right)COV(r_i,r_j) \tag{19}$$

则投资组合的下半方差:

$$\sigma^2(p)^{-}=x'G^{-}x \tag{20}$$

3.3　投资组合优化模型

用均值和下半方差分别衡量投资组合的收益和风险,考虑交易成本、上下界约束的情况下,构建均值－下半方差投资组合模型:

$$\begin{aligned}&\min\ x'G^{-}x/2\\&\max\ \sum_{i=1}^{n}r_i x_i-\sum_{i=1}^{n}c_i|x_i-x_{0i}|\\&\text{s.t.}\begin{cases}0\leqslant x_i\leqslant u_i & (a_1)\\ \sum_{i=1}^{n}x_i=1 & (a_2)\end{cases}\end{aligned} \tag{21}$$

在(21)式中,有2个约束条件:(a_1)是每种风险资产的投资比例的上下界约束;(a_2)是n种风险资产的投资比例和为1。

接下来,在(21)式中引用风险厌恶系数θ来描述投资者的预期收益和风险之间的关系:

$$\begin{aligned}&\min\ \theta\frac{1}{2}x'G^{-}x-(1-\theta)\left(\sum_{i=1}^{n}r_i x_i-\sum_{i=1}^{n}c_i|x_i-x_{0i}|\right)\\&\text{s.t.}\begin{cases}0\leqslant x_i\leqslant u_i & (a_1)\\ \sum_{i=1}^{n}x_i=1 & (a_2)\end{cases}\end{aligned} \tag{22}$$

其中,θ表示投资者的风险厌恶系数,$\theta\epsilon[0,1]$,当$\theta=0$时,表示投资者是风险偏好型;当$\theta=1$时,表示投资者是风险厌恶型。

4　实证分析

4.1　样本选择及数据来源

本文选取2013年1月1日—2021年6月30日沪深300指数成分股为研究样本,并在研究期限内删除暂停上市或者未上市的股票,共选择75支股票数据。在附录中,附表1展示了所选用股票的详细信息。

同时,考虑到时间序列的连续性,本文将采用滚动窗口的方法进行模拟,借鉴文献[26]的方法,以2年作为一个训练和测试的周期,其中前1年半作为训练期,后半年作为测试期。将2013年1月1日—2021年6月30日的样本数据分为14个周期。如图1所示,其中,黑色代表1年半的训练集,白色代表半年的测试集。

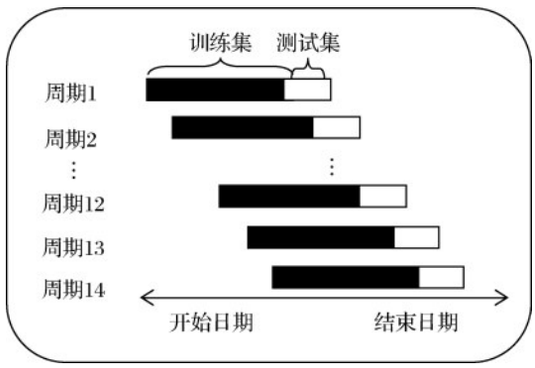

图1　样本的滚动窗口

4.2　机器学习+投资组合模型

(1)指标因子

由于输入变量的选择对于时间序列预测任务是非常重要的,本文借鉴文献[8,14,22]的研究成果,使用包括开盘价格(open price,OP)、收盘价格(close price,CP)、最高价格(high price,HP)、最低价格(low price,LP)作为指标因子的输入参数,同时,还包括Ture－range、MA均线、MTM动量线和RSI相对强弱指标等4个技术指标来反映金融市场的有效特征。在附录中,附表2展示了19个指标因子的详细信息。

(2)指标因子重要分析

由于XGBoost预测模型在构建提升树之后,计算输入变量的重要性分数相对简单,因此,本文根据XGBoost预测模型中变量的重要性分数,对整个样本时间的变量进行分析,具体结果如图2所示。一般来说,重要性分数表明每个变量在模型内构建提升决策树时的有用性和价值。显然,变量r_{t-2}的重要性最高为87,变量Ture－range的重要性最低为4。与文献[8]的结论一致,这与文献[8]的结论一致。相对于这4个技术指标(Ture－range、MA、MTM和RSI)来说,以开盘、收盘、最高和最低四种价格计算的15个滞后收益观察值对收益预测准确性的重要程度更高一些。也就是说,各种交易数据(尤其是价格因素)能更好地反映股票市场信息,表明了金融时间序列预测总是用滞后观测来解释。同时,技术指标通常与未来市场变化相关,很多投资者会遵循技术指标进行投资,所以不能忽略技术指标在预测金融时间序列方面的影响。

(3)实证过程

本文的实证过程主要分为两个步骤:一是股票选择(通过机器学习预测方法选择较优质的股票);二是投资组合优化(在考虑交易成本和上下界约束的情况下,构建最优M－SV投资组合)。在附录中,附图1展示了基于机器学习的投资组合框架信息。

步骤1:股票选择。模型预测的输入指标如附

录中的附表2所示。本文为了构建一个稳健的多元化投资组合，分别采用XGBoost算法、SVR算法和KNN算法对第$t+1$天的股票收益率$\hat{r}_i$进行预测并排序，最后选择排在前k位的股票进行投资。

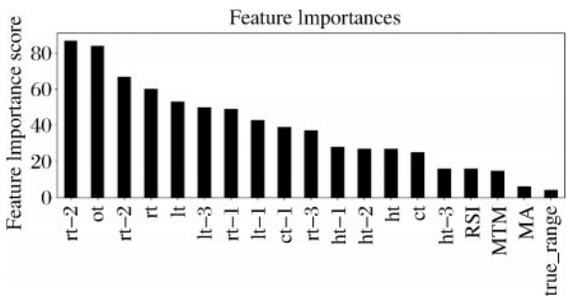

图2 特征选择

步骤2：投资组合优化。根据步骤1选择出较优质的股票后，在选定的股票中分散一定数量的财富。本步骤的目标是使用M－SV模型确定每项资产的投资比例。具体而言，M－SV模型可转化为具有线性等式和线性不等式组的二次规划问题，本文采用张鹏[27]提出的旋转算法求解最优投资组合策略。

此外，借鉴其他学者[28]的研究，将运用MV模型、1/N模型以及随机模型用于比较。具体地：①机器学习(XGBoost算法、SVR算法和KNN算法)＋M－SV、MV和1/N模型之间的比较。采用不同的机器学习方法是为了确定预测方法是否会影响最优投资组合的有效性。在步骤1中，该模型使用XGBoost算法、SVR算法和KNN算法这三种不同的机器学习方法，选择第$t+1$天预测收益率$\hat{r}_i$的前k只股票。在步骤2中，采用M－SV、MV或1/N模型确定所选股票的最优投资比例。②与随机选股(Random)＋M－SV、MV和1/N模型进行比较。随机选股方法是随机进行的，不依赖于任何预测。具体来说，本文从所有样本中随机选择一些股票，并采用M－SV、MV或1/N模型来确定最优的投资组合比例，从而检验使用机器学习预测第$t+1$天股票收益率$\hat{r}_i$的必要性。

5 实证结果

5.1 模型调参和评价

(1)模型参数的设置

超参数的设置对机器学习模型的表现具有直接影响。本文采用交叉验证与网格搜索法相结合的方法寻找机器学习模型的最优超参数。上文所描述的三种机器学习模型的超参数值的设置如附录中附表3所示，利用经过参数寻优的机器学习模型对股票的收益率进行预测。

(2)评价选股模型的表现

经过上一部分的参数优化，确立了最终的机器学习选股模型。以附录中的附表2的指标因子作为模型的输入指标，根据预测股票收益率排在前k位的股票进行投资。本部分借鉴文献[22]的均方误差(mean square error，MSE)、均方根误差(root mean square error，RMSE)、平均绝对误差(mean absolute error，MAE)、H_R(hit ratio)作为模型误差的评价指标。具体公式如下：

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(r_i-\hat{r}_i)^2 \tag{23}$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i-\hat{r}_i)^2} \tag{24}$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|r_i-\hat{r}_i| \tag{25}$$

$$H_R=\frac{1}{n}\sum_{i=1}^{n}a_i \tag{26}$$

其中，r_i和$\hat{r}_i$分别表示$t+1$时第i个股票的实际收益和预测收益，n表示预测时间；如果$r_i\cdot\hat{r}_i>0$，则$a_i=1$，否则$a_i=0$。

表1总结了2013年1月—2021年6月各模型的预测结果。从MAE、MSE、RMSE和H_R误差评价指标的结果来看，XGBoost算法相对于SVR算法和KNN算法来说，具有更高的预测效果，SVR算法预测效果次之。

表1 不同模型预测的表现

模型		*MAE*	*MSE*	*RMSE*	H_R
XGBoost	Mean	0.0295	0.0012	0.0364	50.92%
	Std	0.0230	0.0014	0.0241	0.1010
SVR	Mean	0.0340	0.0025	0.0440	47.85%
	Std	0.0190	0.0029	0.0247	0.1126
KNN	Mean	0.0267	0.0014	0.0330	45.05%
	Std	0.0182	0.0017	0.0185	0.0794

5.2 投资组合的结果

本部分将讨论三种模型在不同交易成本和上界约束情况下的投资组合表现，以比较所提出的模型在实际股票市场中的盈利表现。此外，为了进一步检验所提出的投资组合模型的有效性，本文引入了MV模型、1/N模型和沪深300指数作为对比模型。具体地，本部分将通过实证对投资组合的最优投资股票数目、数据特征、累积收益率、日度收益率的箱体图和不同周期的夏普比率进行分析。其中，假设投资者是风险厌恶型，风险厌恶系数$\theta=0.9$，单位交易成本$c=0$或0.0005，无风险年利率$r_f=0.03$，上界约束$u_i=0.5$或0.35。

(1)在一定条件下，对不同模型的投资组合规

模进行分析

在金融市场中,由于资产规模、交易成本和人力资源等方面的原因,投资者不可能投资每一支股票,更多地是根据自己的投资偏好和风险厌恶程度,投资于数个风险资产和无风险资产,因此,本文借鉴文献[8]对投资组合规模的研究成果,考虑最优投资股票个数$k\in\{5,6,7,8,9\}$的情况下来构建投资组合模型。此外,本文选择年度的收益率、标准差、夏普比率和索提诺比率来评价投资组合规模的年度表现。其中,$c=0$,$u_i=0.5$。

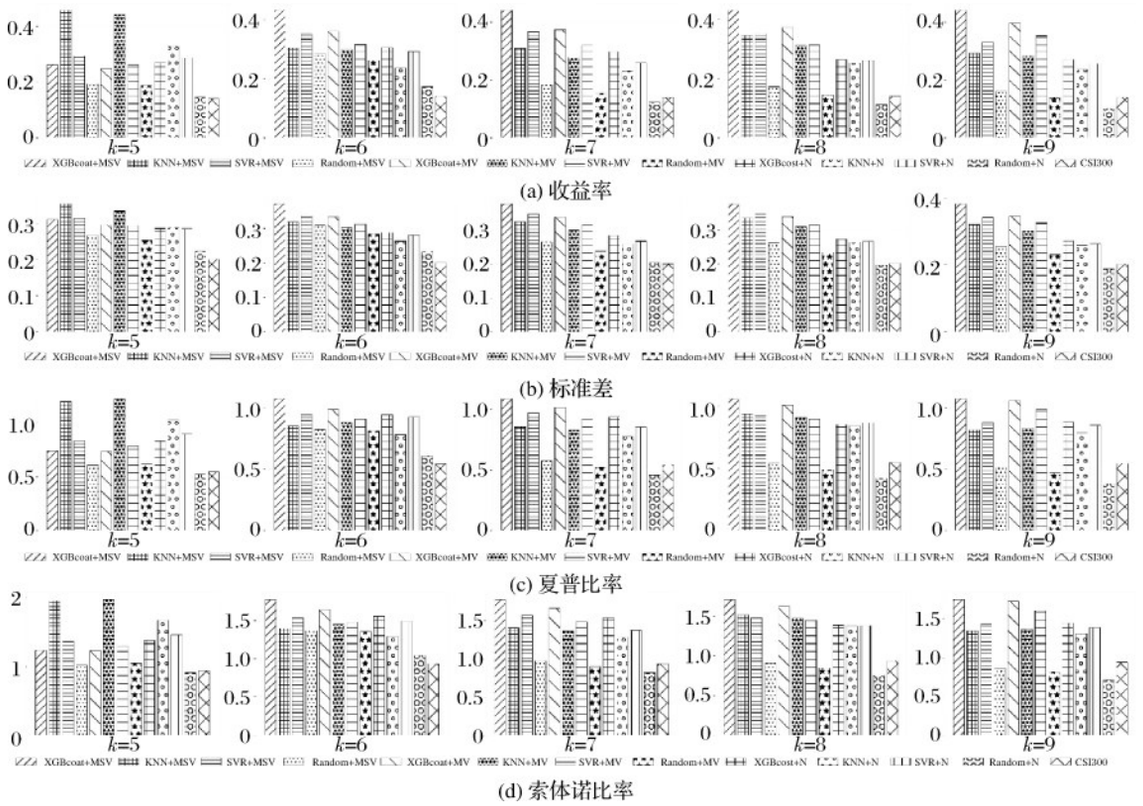

图3　当k取不同值时,不同投资组合的年化表现

从图3可以看出,不同投资组合模型在k取不同值的年度表现,其中,四个子图的纵轴分别表示年度的收益率、标准差、夏普比率、索提诺比率,横轴表示不同模型的投资组合规模。研究结果表明,无论投资组合规模k取何值,XGBoost＋M－SV模型在年化收益率、夏普比率和索提诺比率这三个维度上都表现出其他策略更高的性能。

①在收益率方面,当投资组合规模$k\in\{6,7,8,9\}$时,XGBoost＋M－SV模型的年化收益率显著优于其他投资策略,尤其当$k=6$时,XGBoost＋M－SV模型表现最优,其年化收益率为0.4316。但当$k=5$时,基于KNN＋M－SV模型的投资策略更胜一筹。

②在标准差方面,当投资组合规模$k\in\{5,6,7,8,9\}$时,各个模型之间的差异并不明显,即大部分结果在0.3上下波动。因此,需要进一步采用收益风险指标(夏普比率和索提诺比率)进行投资策略的评价。

③在夏普比率方面,当投资组合规模$k\in\{6,7,8,9\}$时,XGBoost＋M－SV模型的夏普比率显著优于其他投资策略,即意味着该投资组合表现较好。尤其是$k=6$时,XGBoost＋M－SV模型表现最优,其年化夏普比率为1.0776。但当$k=5$时,基于KNN＋M－SV模型的投资策略相对较好。这表明,XGBoost＋MSV模型的风险指标(标准差)没有显著优于其他模型,但其单位风险回报总体表现优于其他模型。

④在索提诺比率方面,同样获得了与夏普比率一致的结论。

总的来说,当最优投资组合个数$k=6$时,这四个指标在每个模型中的整体表现优于k取其他值。因此,本文选择$k=6$作为最优投资组合的数目进行后续的分析。

(2)在不同的约束条件下,对投资组合模型的数据特征进行统计性分析

本文将证券交易的印花税、佣金和过户税都视为单向交易成本,通过借鉴文献[22]的研究成果,在不同交易成本和上界约束条件的条件下,对投资组合模型的数据特征进行统计性分析,具体地,分以下A、B和C三个板块,其分别表示为日度收益特征、日度风险特征和年度风险收益指标。其中,$k=6$,$c=0$或0.0005,$u_i=0.5$或0.35。不同模型数据

的统计性分析,结果如附录中附表4～附表6所示。

在日度收益特征上,运用日度收益率的均值、标准差、最小值和最大值来分析投资组合的日度收益特征,可以得到:①附表4的板块A中可以看出,XGBooat+M－SV模型日度平均收益率的最高值为0.0017,SVR+M－SV模型紧随其后,其平均收益率为0.0015。②在改变交易成本和上界约束后,如附表5和附表6的板块A所示,XGBooat+M－SV模型的日均收益排名还是最高的。此外,还可以看出,1/N模型和CSI300的标准差相对较小。

在日度风险特征上,使用1%VaR、5%VaR、1%CVaR和5%CVaR来衡量投资组合的风险。可以看到:①在附表4的B板块中,XGBoost+M－SV模型的风险略高于其他模型,这与实际投资情况相吻合,即较高的收益通常面临较高的风险水平。②在改变交易成本和上界约束后,从附表5和附表6的B板块可以看出,当考虑交易成本后,各投资模型面临的风险增加,此时减小上界约束,即增加投资的分散化程度,可以降低风险。

在年度风险收益指标上,使用收益率、标准差、夏普比率、索提诺比率、下行风险和最大回撤分析投资组合的年度风险收益特征,可以得到:①在附表4的C板块中,XGBoost+M－SV模型的各个指标整体表现优于其他模型。其中,年化收益率为0.4434、夏普比率为1.0879、最大回撤为0.4147、下行偏差为0.2445、索提诺比率为1.7844、标准差为0.3800。②在改变交易成本和上界约束后,从附表5和附表6的C板块同样可以得出一致的结论。此外,当考虑交易成本后,各投资策略的整体表现下降,而减小上界约束虽然可以起到分散风险的作用,但同时也会减小投资策略获得的收益。

总的来说,在不同交易成本和上界约束的情况下,一方面,基于机器学习(XGBoost算法、SVR算法和KNN算法)的投资组合模型在日度收益特征、日度风险特征和年度风险收益指标方面整体优于Random模型和沪深300指数,尤其是XGBoost模型表现突出;另一方面,M－SV模型整体也优于MV模型和1/N模型。

(3)在不同约束条件下,对投资组合的累计收益率进行分析

本文采用机器学习(XGBoost算法、SVR算法和KNN算法)+M－SV、MV和1/N模型与随机选股(Random)+M－SV、MV和1/N模型,以沪深300指数为基准,对投资组合的累积收益率进行分析,从而验证模型和算法的有效性。其中,k=6,c=0或0.0005,u_i=0.5或0.35。

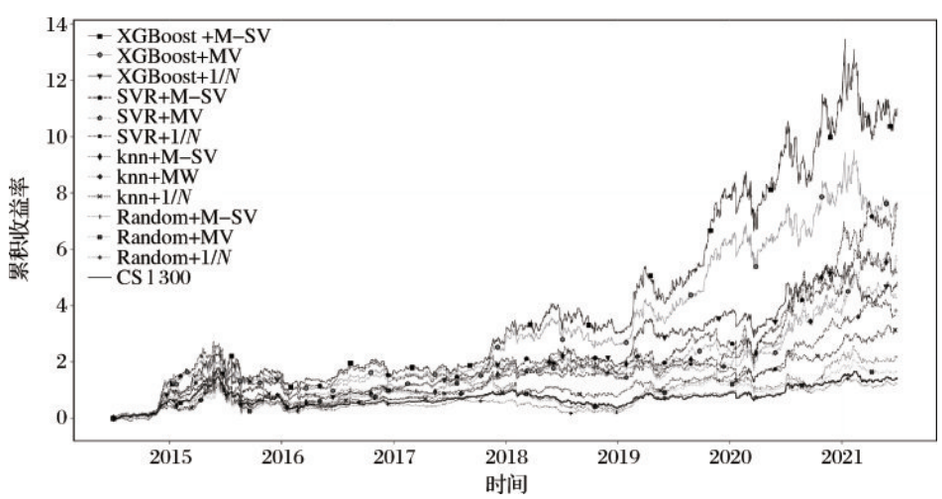

图4 当c=0,u_i=0.5时,不同模型的累积收益率

从图4～图6可以看出不同模型策略随时间的性能。具体来说:①当交易成本c=0,上界约束u_i=0.5时,从图4可以看出不同投资组合模型的累积收益率,其中,XGBoost+M－SV模型的累积收益率为1099.4%,XGBoost+MV模型和XGBoost+1/N模型的累积收益率分别为763.4%和484.5%,在数值上显著低于XGBoost+M－SV模型,这表明,本文所考虑的MSV模型比MV模型和1/N模型具有更好的样本外表现;此外,SVR+M－SV模型、KNN+M－SV模型、Random+MSV模型和CSI300的累积收益率分别为744.4%、524.1%、214.8%和141.6%,在数值上也显著低于XGBoost+M－SV模型。

②当交易成本c=0.0005,上界约束u_i=0.5时,从图5可以看出不同投资组合模型的累积收益率相比图4的有了明显的下降。其中,XGBoost+M－SV模型的累积收益率为411.5%,而XGBoost+MV模型、XGBoost+1/N模型、SVR+M－SV模

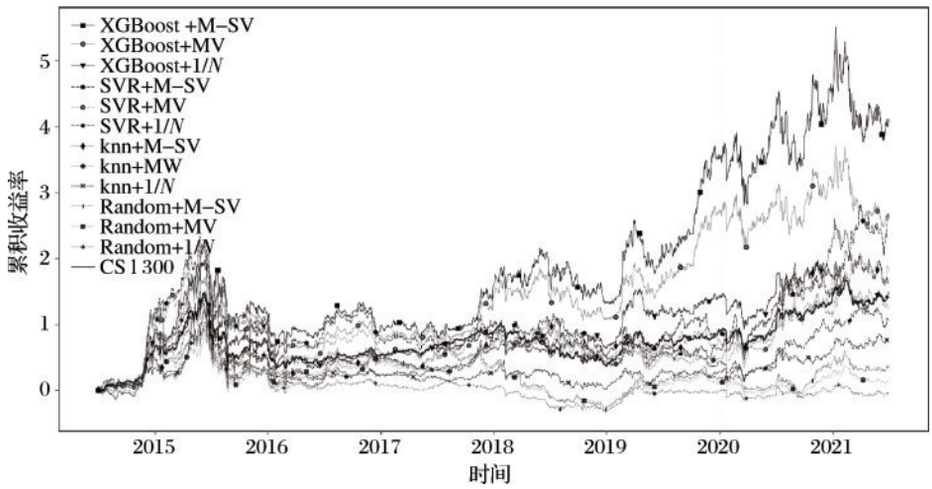

图5　当c=0.0005，u_i=0.5时，不同模型的累积收益率

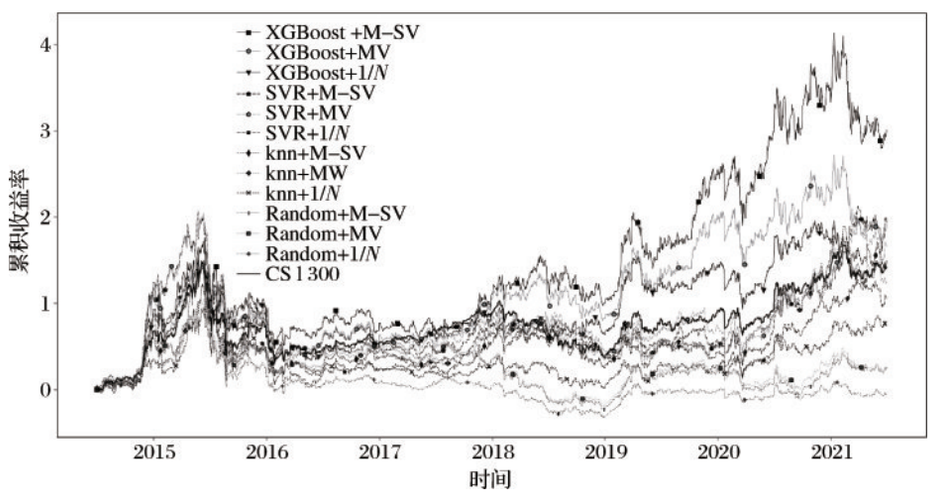

图6　当c=0.0005，u_i=0.35时，不同模型的累积收益率

型、KNN＋M－SV模型、Random＋M－SV模型和CSI300的累积收益率分别为268.2%、149.2%、260.0%、166.1%、34.2%和141.6%。

③当交易成本c=0.0005，上界约束u_i=0.35时，从图6可以看出不同投资组合模型的累积收益率相较于图5有了明显的降低。其中，XGBoost＋M－SV投资组合的累积收益率为300.9%，而XGBoost＋MV模型、XGBoost＋1/N模型、SVR＋MSV模型、KNN＋M－SV模型、Random＋M－SV模型和CSI300的累积收益率分别为191.5%、149.2%、193.7%、141.7%、23.6%和141.6%。

综上，可以发现，不同交易成本和上界约束对投资组合模型的累积收益有着显著的影响，即考虑交易成本和减小上界约束会降低累积收益。另一方面，在不同的约束条件下，XGBoost＋M－SV模型的累积收益远高于其他投资组合模型和CSI300的累积收益，这表明，XGBoost算法比其他选股算法更为有效，同时，M－SV模型比其他优化模型在样本外的表现更好。

(4)在不同的约束条件下，对不同模型日度收益率的箱线图进行分析

为了描述不同投资组合模型策略的结果，本文对机器学习(XGBoost算法、SVR算法和KNN算法)＋M－SV、MV和1/N模型与随机选股(Random)＋M－SV、MV和1/N模型，以及沪深300指数的日度收益率分布进行分析，得到了不同模型日度收益率的箱线图。其中，k=6，c=0或0.0005，u_i=0.5或0.35。

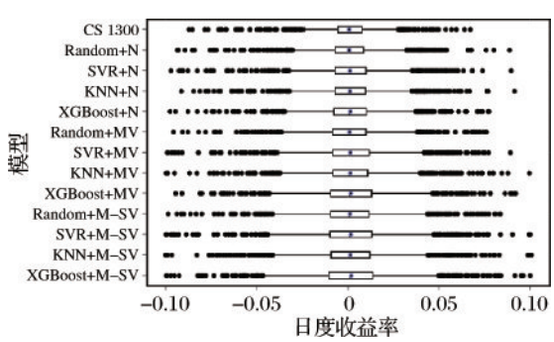

图7　当c=0，u_i=0.5时

从图7～图9中，可以看出不同模型日度收益率的箱线图。具体来说：①在c=0，u_i=0.5的情况下，从图7可以看出XGBoost＋M－SV模型的波动率

最低(不考虑CSI300),其次是SVR+M-SV模型。②在改变交易成本和上界约束的情况下,从图8和图9可以看出(XGBoost算法、SVR算法和KNN算法)+MSV、MV和1/N模型的波动范围明显扩大,离散点较少。而随机选股(random)+M-SV、MV和1/N模型的波动范围较小(不考虑CSI300),但离群点较多。

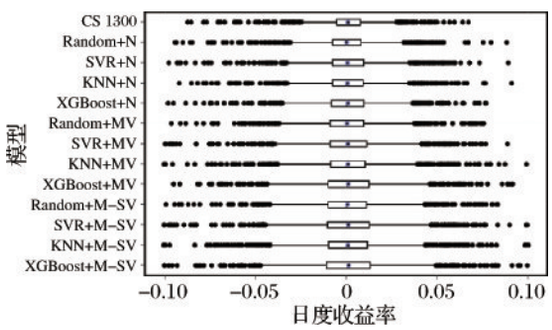

图8 当c=0.0005,u_i=0.5时

(5)在不同的约束条件下,对不同周期投资组合模型的夏普比率进行分析

为了进一步分析机器学习(XGBoost算法、SVR算法和KNN算法)+M-SV、MV和1/N模型与随机选股(random)+M-SV、MV和1/N模型的良好性能是否只发生在某一时期内,本文以半年为一个周期来比较不同模型的夏普比率。其中,k=6,c=0或0.0005,u_i=0.5或0.35。

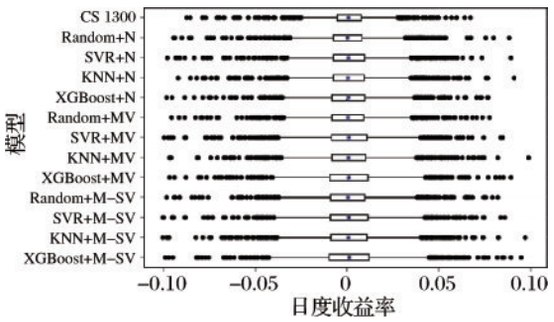

图9 当c=0.0005,u_i=0.35时

图10~图12展示了不同交易成本和上界约束情况下的夏普比率,可以看出:①在大部分时间段,机器学习(XGBoost算法、SVR算法和KNN算法)+M-SV、MV和1/N模型可以实现正的夏普比率(除了2015年下半年的股灾、2018年的中美贸易摩擦以及2020年的新冠疫情等特殊情况),特别是XGBoost+M-SV模型的夏普比率在相应时期的结果优于其他模型。②在14个投资周期中,基于机器学习选股模型构建的投资组合的夏普比率多数优于随机选股构建的投资组合和沪深300指数,尤其是图11的XGBoost+M-SV模型的夏普比率在6个对应周期的结果优于其他模型。

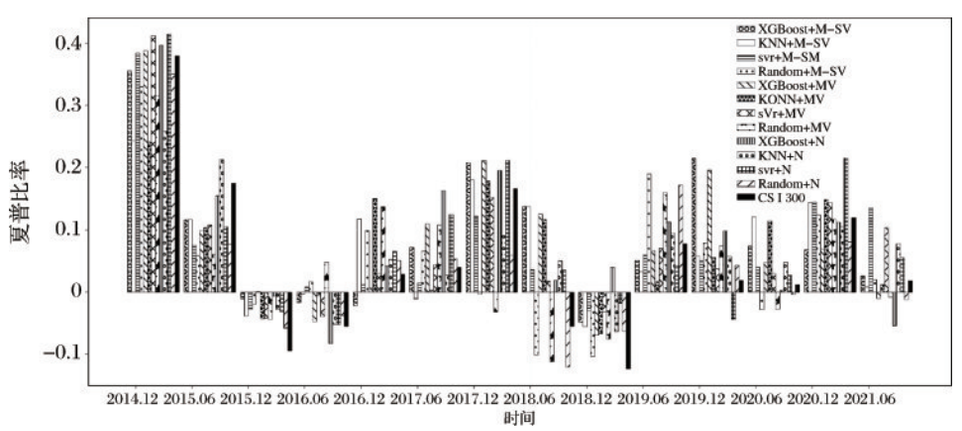

图10 当c=0,u_i=0.5时,不同周期的夏普比率

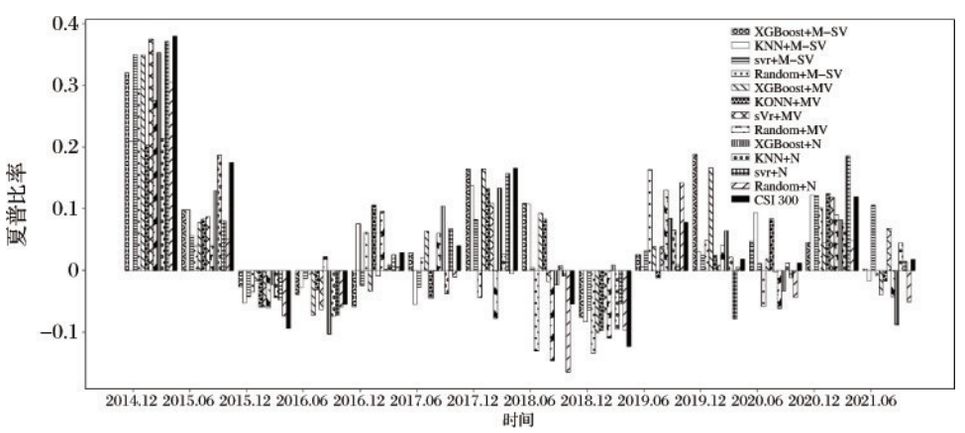

图11 当c=0.0005,u_i=0.5时,不同周期的夏普比率

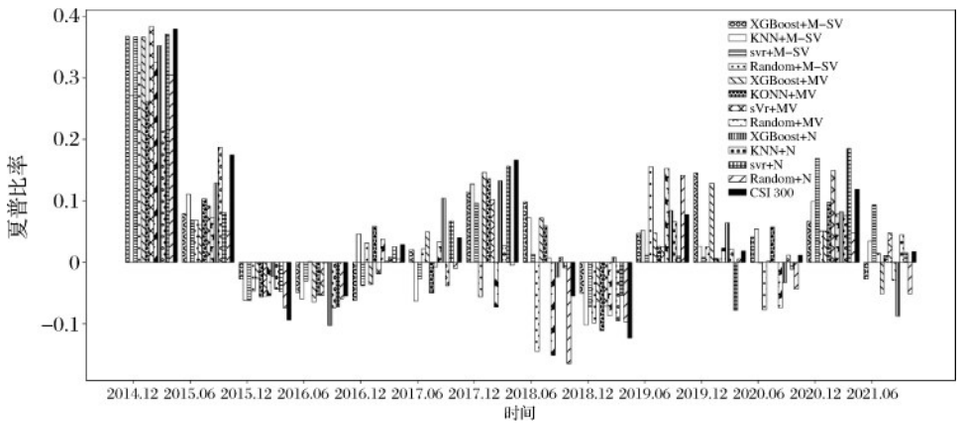

图 12　当 $c=0.0005$，$u_i=0.35$ 时，不同周期的夏普比率

6　结语

本文在考虑交易成本、上下界约束的基础上，提出了一种基于机器学习的两步骤多元化投资组合优化模型。具体来说，该模型涉及以下两个步骤：步骤1是股票选择，即通过机器学习（XGBoost算法、SVR算法、KNN算法）选择具有较高预测收益率的股票，并对模型进行评估和选择；步骤2是投资组合优化，其根据步骤1选择预测收益较高的股票后，采用M－SV模型、MV模型和1/N模型确定所选股票的投资比例。最后，本文采用沪深300指数成分股作为样本数据进行实证研究，并得到以下几点重要结论。

（1）在预测模型评价上，从MAE、MSE、RMSE和H_R四种误差评价指标的实证结果来看，XGBoost算法相对于SVR算法和KNN算法来说在股票收益预测上具有更好的预测效果。

（2）对不同模型的投资组合规模进行分析，研究结果表明：当最优投资组合个数$k=6$时，投资组合模型年度的收益率、标准差、夏普比率和索提诺比率这四个指在每个模型中的整体表现优于k取其他值。

（3）在不同交易成本和上界约束条件的条件下，本文分别对不同投资组合模型的样本外表现进行分析，研究结果表明：一是XGBoost＋M－SV模型在日度收益特征、日度风险特征和年度风险收益指标方面整体优于其他投资组合模型，并且交易成本的考虑导致投资组合的表现变差，而上界约束的减小在起到分散风险的同时会降低收益。二是XGBoost＋M－SV的累积收益远高于其他投资组合模型，进一步验证了XGBoost选股模型比其他选股模型更为有效，M－SV模型比其他优化模型在样本外的表现更好。三是在投资组合模型的日度收益率分布方面，XGBoost＋M－SV模型获得了相对较低的波动率，并且在不同交易成本和上界约束的情况下，各投资组合模型的收益波动范围和离群点会发生相应的变化。

（4）在某一周期的内，对不同模型的夏普比率进行分析，研究结果表明：一是在样本外的大部分时间段，（XGBoost算法、SVR算法和KNN算法）＋M－SV、MV和1/N模型可以得到正的夏普比率；二是投资组合模型良好的表现发生在不同的周期，且在大多数时间段，基于机器学习选股模型构建的投资组合的夏普比率优于随机选股构建的投资组合和沪深300指数。

参考文献：

[1] Markowitz H. Portfolio selection[J]. Journal of Finance, 1952,7(1):77－91.

[2] Markowitz H M. Portfolio Selection: Efficient Diversification of Investments [M]. New York, Wiley: 1959.

[3] Zhang Peng, Dang Shili. The weighted lower and upper admissible mean downside semi－variance portfolio selection[J]. International Journal of Fuzzy Systems, 2021,23(6):1775－1788.

[4] 姚海祥，姜灵敏，马庆华．不允许买空时的均值－下方风险投资组合选择——基于非参数估计方法[J]．数理统计与管理，2015，34(6)：1077－1086.
Yao Haixiang, Jiang Lingmin, Ma Qinghua. Mean－downside risk portfolio selection without short selling: based on nonparametric estimation[J]. Journal of Applied Statistics and Management, 2015, 34(6):1077－1086.

[5] Yan Wei, Miao Rong, Li Shurong. Multi－period semi－variance portfolio selection model and numerical solution [J]. Applied Mathematics and Computation, 2007, 194(1): 128－134.

[6] 张鹏，李影，曾永泉．现实约束下多阶段模糊投资组合的时间一致性策略研究[J]．中国管理科学，2022. DOI: 10.16381/j.cnki.issn1003－207x.2020.1759.
Zhang Peng, Li Ying, Zeng Yongquan. Time－consistent strategy for the multiperiod possibilistic portfolio selection with real constraints[J]. Chinese Journal of Management Science, 2022. DOI: 10.16381/j.cnki.issn1003－207x.2020.1759.

[7] Guerard J B, Markowitz H, Xu Ganlin. Earnings forecasting in a global stock selection model and efficient portfolio construction and management [J]. International Journal of Forecasting. 2015,31(2):550－560.

[8] Wang Wuyu, Li Weizi, Zhang Ning, et al. Portfolio formation with preselection using deep learning from long－term financial data [J]. Expert Systems with Applications. 2020, 143: 113042.

[9] Lin Chiming, Huang J J, Gen M, et al. Recurrent neural network for dynamic portfolio selection[J]. Applied Math-

ematics and Computation, 2006, 175(2):1139－1146.
[10] Thenmozhi M, Sarath Chand G. Forecasting stock returns based on information transmission across global markets using support vector machines[J]. Neural Computing and Applications, 2016, 27(4):805－824.
[11] Krauss C. Deep neural networks, gradient－boosted trees, random forests: statistical arbitrage on the S&P 500[J]. European Journal of Operational Research, 2017(2): 689－702.
[12] Freitas F D, Souza A, Almeida A. Prediction－based portfolio optimization model using neural networks[J]. Neurocomputing, 2009, 72(10－12):2155－2170.
[13] Day M Y, Lin Jianting. Artificial intelligence for ETF market prediction and portfolio optimization[C]//Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Vancouver, Canada, August 27－30, 2019, IEEE, 2019: 1026－1033.
[14] Fischer T, Krauss C. Deep learning with long short－term memory networks for financial market predictions [J]. European Journal of Operational Research, 2018, 270(2): 654－669.
[15] Lee S I, Yoo S J. Threshold－based portfolio: the role of the threshold and its applications [J]. Journal of Supercomputing, 2018, 76(10): 8040－8057.
[16] Wong S. Stock price prediction model based on the short－term trending of KNN method[C]//Proceedings of the 2020 7th International Conference on Information Science and Control Engineering, Changsha, China, June 7, 2020, IEEE: 1355－1360.
[17] Paiva F D, Cardoso R T N, Hanaoka G P, et al. Decision－making for financial trading: a fusion approach of machine learning and portfolio selection[J]. Expert Systems with Applications, 2019, 115: 635－655.
[18] Ayala J, M García－Torres, Noguera J, et al. Technical analysis strategy optimization using a machine learning approach in stock market indices[J]. Knowledge－Based Systems, 2021, 225(6):107119.
[19] Ma Yilin, Han Ruizhu, Wang Weizhong. Portfolio optimization with return prediction using deep learning and machine learning [J]. Expert Systems with Applications. 2021, 165: 113973.
[20] Chen T, Guestrin C. XGBoost: a scalable tree boosting system [C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, August, 13－17, 2016: 785－794.
[21] Hongjoong K. Mean－variance portfolio optimization with stock return prediction using XGBoost[J]. Economic Computation and Economic Cybernetics Studies and Research, 2021, 55(4): 5－20.
[22] Chen Wei, Zhang Haoyu, Mehlawat M K, et al. Mean－variance portfolio optimization using machine learning: based stock price prediction[J]. Applied Soft Computing, 2021, 100: 106－943.
[23] Cortes C, Vapnik V. Support－vector networks. Mach Learn[J], 1995, 20(3):273－297.
[24] Pan Jia, Manocha D. Bi－level locality sensitive hashing for k－nearest neighbor computation[C]//Proceedings of the 2012 IEEE 28th International Conference on Data Engineering, Washington D C, America, April 01－05, 2012, IEEE, 2012: 378－389.
[25] 张鹏. 可计算的投资组合模型与优化方法研究[D]. 武汉: 华中科技大学, 2006.
Zhang Peng. The studying on the models and optimal methods of the computable selection [D]. Wuhan: Huazhong University of Science and Technology, 2006.
[26] Yang Li, Shami A. On hyperparameter optimization of machine learning algorithms: theory and practice [J]. Neurocomputing, 2020, 415: 295－316.
[27] 张鹏. 不允许卖空情况下均值－方差和均值－VaR投资组合比较研究[J]. 中国管理科学, 2008(4): 30－35.
Zhang Peng. The comparison between mean－variance and mean－VaR portfolio models without short sales [J]. Chinese Journal of Management Science, 2008, 16(4):30－35.
[28] Ballings M, Van den Poel D, Hespeels N, et al. Evaluating multiple classifiers for stock price direction prediction[J]. Expert Systems with Applications, 2015; 42(20):7046－7056.

Two-stage Mean Semi-variance Portfolio Optimization with Stock Return Prediction Using Machine Learning

ZHANG Peng[1], DANG Shi-li[1], HUANG Mei-yu[2], LI Jing-xin[1]

(1. School of Economics & Management, South China Normal University, Guangzhou 510006, China;
2. School of Management, Huazhong University of Science & Technology, Wuhan 430074, China)

Abstract: Since accurately predicting stock return sequences can improve the performance of portfolio optimization models, the results have indicated that machine learning methods have a greater capacity to confront problems with nonlinear, nonstationary charateristics than econometric models. Consequently, a novel two-stage method is proposed for well-diversified portfolio construction based on stock return prediction using machine learning, which includes two stages. To be specific, the purpose of the first stage is to select diversified stocks with high predicted returns, where the returns are predicted by machine learning methods, i. e. eXtreme Gradient Boosting(XGBoost), support vector regression(SVR), K-Nearest Neighbor(KNN), and evaluate and select the model. In the second stage, considering the constraints such as transaction costs and threshold constraints, the predictive results are incorporated into the mean semi-variance (M-SV) model, mean-variance model and equally weighted model to determine optimal portfolio. Finally, using China Securities 300 Index component stocks as study sample, the empirical results demonstrate that the XGBoost＋MSV model achieves better results than similar counterparts and market index in terms of return and return-risk metrics.

Key words: mean semi-variance; pivoting algorithm; machine learning; stock price prediction