DOI:10.19473/j.cnki.1008-4940.2020.06.004

倒置误差组合优化算法的沪深 300 指数预测研究

谢承燕，方宏彬，郭梦洁，杨梦卓

（安徽大学　经济学院，安徽　合肥，230601）

[摘　要] 针对多变量预测股指开盘价问题，为了提高预测精度，提出一种基于倒置误差的 GXS 组合模型，对沪深 300 指数每日开盘价进行回归预测。运用网格搜索（GridSearchCV）算法和 10 折交叉验证法，对极度梯度提升树（XGBoost）模型与基于径向基 RBF 核函数的支持向量回归（SVR）模型进行参数优化，取修正的预测误差进行误差倒数法赋权，搭建 GXS 组合模型。实证结果表明，基于修正误差 MAE 赋权的 GXS 组合模型对沪深 300 指数开盘价预测效果最优。

[关键词] 沪深 300 指数；XGBoost 模型；SVR 模型；误差倒数法；回归预测

[中图分类号] F830.91　　[文献标识码] A　　[文章编号] 2096-3300（2020）06-0014-09

大数据时代下，随着机器学习和深度学习等算法模型在金融领域的广泛应用，量化投资引起了国内外的研究热潮。西蒙斯提出的“壁虎式”交易策略通过捕捉市场短暂且微小异常的变化，以期在短期内随时进行买入卖出的高频量化交易，依靠活跃获利。中国股票市场的量化投资发展相对西方较晚，2005 年 4 月 8 日由沪深证券交易所联合发布的沪深 300 指数反映中国证券市场股票价格变动，为指数化投资和指数衍生产品创新提供基础条件，推动了量化交易的发展。机器学习和深度学习等算法模型能较好地处理金融数据的复杂关系，在金融科技领域得到了极大的重视。较多学者运用组合学习算法对金融资产价格进行分类预测[1-3]。Chen 提出的 XGBoost 算法在近两年的量化投资中取得了较好的应用[4-6]。研究发现支持向量机在金融资产价格回归预测中运用较为广泛[7-10]。

一、文献综述

金融资产价格预测大致可以分为分类和回归两大类。现有量化投资研究中，分类预测研究主要对下一时刻价格涨跌问题进行二分类预测。Thakur[1]689-702利用随机森林和支持向量机的组合学习算法，对美国股市的多指数涨跌变动进行预测，先通过随机森林提取有效特征并降维，再用改进的支持向量机模型对指数价格涨跌进行预测，实证发现该方法提高了预测精度；Krauss[2]238-243将随机森林、梯度提升树和人工神经网络三种机器学习算法等权重相加，形成新的组合算法，对标普 500 指数进行下一日价格涨跌预测，实证分析表明，买入预测上涨概率前十的股票，根据收益结果显示，该组合算法的预测效果最佳；谢琪[3]238-243提出一种基于长短记忆神经网络集成学习的金融时间序列预测模型，构建出六层长短记忆神经网络，通过集成学习

收稿日期：2020-05-06

基金项目：国家社科基金资助项目“基于马克思劳动生产力理论的技术进步测度方法研究”（19BJL004）。

作者简介：谢承燕（1996-），女，安徽六安人，硕士研究生，研究方向：量化投资；

方宏彬（1971-），男，安徽池州人，副教授，博士，研究方向：数据挖掘与粒度计算。

中 Bagging 方法组合 8 个长短记忆神经网络，选取中国股市的 6 个指数组成金融时间序列数据集，对下一个交易日的涨跌情况进行分类预测，实验结果表明该模型具有较好的预测效果；黄卿[4]297-307等采用沪深 300 股指期货 1 分钟高频数据作为研究对象，对比分析神经网络、支持向量机和 XGBoost 对股指期货下 1 分钟价格的变动方向的预测能力，研究发现三种机器学习的预测能力都较好，但 XGBoost 的预测能力要优于传统的神经网络和支持向量机；王芊[5]27-30基于高效机器学习算法 XGBoost 建立了一套量化研究的模型，通过与随机森林、支持向量机等多种机器学习方法进行对比，实证发现运用 XGBoost 模型对沪深 300 中 300 支成分股进行涨跌预测的准确率最高。

XGBoost 算法中有两种增强树：回归树和分类树。黄卿[4]304-305、王芊[5]33-35运用 XGBoost 算法进行价格涨跌分类预测；王燕[6]202-207等通过网格搜索算法对 XGBoost 模型进行参数优化构建 GS-XGBoost 的金融回归预测模型，对中国平安、中国建筑、中国中车、科大讯飞和三一重工等五只股票的日收盘价进行短期回归预测，对比分析 GBDT 模型、SVM 模型与改进的 XGBoost 模型在评价指标 MSE、RMSE 与 MAE 上的预测效果，得出 GS-XGBoost 金融预测模型在股票短期预测中具有更好的拟合性能。

金融资产价格波动较大，用传统的数学模型对其进行预测的准确率较低，国内学者利用机器学习算法在股票市场、期货市场上进行回归预测的研究较少，主要利用支持向量机模型进行价格回归预测。王芳[7]18-42构建了通过遗传算法优化的支持向量机，对沪深 300 指数的每日开盘价进行回归预测，在波动区间上，进一步建立了基于模糊信息粒化的支持向量机回归预测模型，对沪深 300 指数进行深入研究；郑明[8]517-524等将模糊信息粒化和支持向量机相结合，对未来 5 天股票数据的变化趋势作出预测，实证表明该方法达到预期效果；赛英[9]35-39等首次提出用支持向量机对股指期货进行回归预测，发现基于粒子群算法的优化线性核函数支持向量机对沪深 300 股指期货每日开盘价具有较好的预测效果；魏勤[10]123-126等运用 SVM 人工神经网络算法对以沪深 300 股指期货为代表的期货市场进行回归预测，结果表明 SVM 是价格预测的较好方法，能充分反映期货价格时间序列的变动。

二、模型原理及 GXS 模型搭建

(一) XGBoost 模型

XGBoost 算法是集成学习算法的延伸，以决策树为基学习器，在集成学习过程中，每一轮学习加入一个决策树，预测结果由各轮决策树运行结果串行相加而成。

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

式中，$f_t(x_i)$ 为第 t 棵决策树的输出结果，$\hat{y}_i^{(t)}$ 为第 t 轮学习模型的预测结果。模型引入惩罚项 $\Omega(f_t)$ 进行 $L2$ 正则化以防止过拟合，为了避免决策树过于庞大，惩罚项对决策树叶子个数的权重进行限制：

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

式中，T 为决策树叶子个数；γ 为权重系数，以限制叶子个数，让每一轮学习加入的决策树模型尽可能精简；$\frac{1}{2}\lambda \sum_{j=1}^{T}\omega_j^2$ 为 $L2$ 正则化项，以降低模型过拟合风险。

模型的目标函数：

$$\mathrm{Obj}^{(t)} = \sum_{i=1}^{n} \mathrm{L}\left(y_i, \hat{y}_i^{(t)}\right) + \Omega(f_t) + \mathrm{Constant} \tag{3}$$

式中，$\mathrm{L}\left(y_i, \hat{y}_i^{(t)}\right)$ 为损失函数，评价模型预测值和真实值之间的差异。让损失函数 $\mathrm{L}\left(y_i, \hat{y}_i^{(t)}\right)$ 和惩罚项 $\Omega(f_t)$ 尽量小，以达到优化目标函数目的。

XGBoost 模型采用泰勒展开式来近似目标函数，区别于传统的决策树模型使用梯度下降的方法优化目标函数，并采用遍历叶子节点代替遍历样本容量以找到能优化目标函数的决策树 $f_t(x_i)$，所以目标函数为：

$$\mathrm{Obj}^{(t)} \approx \sum_{j=1}^{T}\left[G_j\omega_j + \frac{1}{2}\left(H_j + \lambda\right)\omega_j^2\right] + \gamma T \tag{4}$$

式中，$\begin{cases} g_i = \dfrac{\partial \mathrm{L}\left(y_i, \hat{y}_i^{(t)}\right)}{\partial \hat{y}_i^{(t-1)}} \\ h_i = \dfrac{\partial^2 \mathrm{L}\left(y_i, \hat{y}_i^{(t)}\right)}{\partial \hat{y}_i^{(t-1)}} \end{cases}$　$\begin{cases} G_j = \sum_{i \in I_j} g_i \\ H_j = \sum_{i \in I_j} h_i \end{cases}$

式中，$\omega_q(x_i)$ 为第 t 棵决策树 $f_t(x_i)$ 的权重输出结果，I_j 为第 j 个叶子节点范围。

公式（4）目标函数 $\mathrm{Obj}^{(t)}$ 对 ω_j 的一阶偏导为零，实现最小化目标函数:

$$\mathrm{Obj}=-\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda}+\gamma T \tag{5}$$

将目标函数 Obj 看成决策树的结构分数，信息增益函数 Gain 是未进行结构分割的树原始分数与进行一次分割后左、右子树分数之差:

$$\mathrm{Gain}=\mathrm{Obj}-(\mathrm{Obj}_L+\mathrm{Obj}_R)$$

经过公式（5）运算得到:

$$\mathrm{Gain}=\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right]-\gamma \tag{6}$$

遍历所有特征取值，选择使信息增益函数 Gain 最大化的特征值为分裂点，说明通过该特征分割的左、右子树结构分数 $\mathrm{Obj}_L+\mathrm{Obj}_R$ 达到最小，决策树结构最优。

（二）SVR 模型

支持向量机的核心思想是寻找一组支持向量，满足由其所确定的分类间隔最大化，使分类的确信度最高，提高模型的推广能力。图 1 为线性可分情况下支持向量机的基本思想。

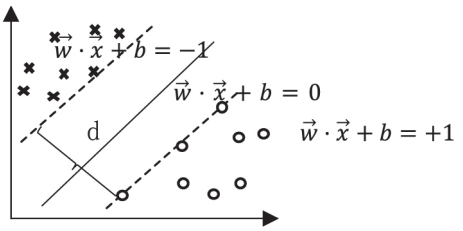

图 1 线性可分 SVM

Fig. 1 Linearly separable SVM

分类间隔 $d=\frac{f(x)}{\|\omega\|}$，最大化分类间隔等价于最小化 $\frac{1}{2}\|\omega\|^2$，则目标函数为:

$$\min\frac{1}{2}\|\omega\|^2,$$

$$s.t.\ y_i(\omega^T x_i+b)\geq 1,\ i=1,2\cdots,n \tag{7}$$

引入非负拉格朗日乘子 α 融合到目标函数中，得到拉格朗日函数 $L(\omega,b,\alpha)$，求解出的鞍点即为目标函数的最优解:

$$L(\omega,b,\alpha)=\frac{1}{2}\|\omega\|^2-\sum_{i=1}^{n}\alpha_i(y_i(\omega^T x_i+b)-1) \tag{8}$$

若数据中存在噪声，引入松弛变量 ξ_i 调整约束条件（9）为:

$$\min\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}\xi_i,$$

$$s.t.\ y_i(\omega^T x_i+b)\geq 1-\xi_i,$$

$$\xi_i\geq 0,\ i=1,2\cdots,n \tag{9}$$

式中，C 为事先确定的正常数，作为控制错分类样本的惩罚力度，称为惩罚参数。Vapnik[11] 提出的支持向量回归（SVR），在用于解决分类问题的支持向量机上实现了回归应用。SVR 核心思想也在于惩罚参数和核函数参数。本文使用灵活性高的径向基 RBF 核函数对非线性数据进行回归预测:

$$k(x,x_i)=\exp(-\gamma\|x-x_i\|) \tag{10}$$

γ 为本文要进行优化的核函数参数。

（三）GXS 模型搭建

网格搜索法需要在实验前设定好参数的取值范围和搜索过程的变化步长，在模型训练过程中会按设定好的步长遍历所有的参数取值，最终通过模型预测评价指标来判断最优参数值。在步长设置过程中，当遍历步长设置过大时，可能导致训练过程变动幅度过大，直接跳过了最优值；步长设置过小会消耗大量训练时间，加大了训练成本。通过多次实验权衡后，比较模型的运行时间和模型拟合效果，找到最佳步长设置和参数取值范围。为了修正模型预测误差，运用网格搜索算法和 10 折交叉验证法，对 XGBoost 模型与 SVR 模型进行参数优化。

误差倒数法赋予权重:

$$\omega_i=\frac{\varepsilon_i^{-1}}{\sum_{i=1}^{n}\varepsilon_i^{-1}} \tag{11}$$

式中，ε_i 表示第 i 种模型的预测误差，n 为预测模型的个数，ω_i 表示第 i 种模型在预测模型中权重。使用修正的预测误差进行误差倒数法赋权，根据公式（11）得:

$$\omega_1=\frac{\varepsilon_2}{\varepsilon_1+\varepsilon_2} \tag{12}$$

$$\omega_2=\frac{\varepsilon_1}{\varepsilon_1+\varepsilon_2} \tag{13}$$

$$f=w_1f_1+w_2f_2 \tag{14}$$

式中，f_1为 XGBoost 的预测值，f_2为 SVR 的预测值，f为组合模型预测值。

通过公式（11）（12）（13）可知，误差倒数法能保证对误差较小的预测模型赋予较大的权重，从而减小组合模型的预测误差，提升组合模型的整体预测精度。搭建 GXS 模型对沪深 300 指数每日开盘价进行预测，与原始单一模型、参数优化单一模型预测结果进行比较，最后根据模型预测评价指标进行验证。

1. 实验步骤

（1）爬取沪深 300 指数历史数据，处理缺失值并将数据进行归一化，划分数据为训练集与测试集，训练集为前 80%数据，测试集为后 20%数据；（2）构建单一模型：XGBoost 模型、SVR 模型，使用训练集数据对模型在初始化参数下进行训练，再用测试集进行预测；（3）网格搜索算法优化单一模型参数，修正模型预测误差，训练改进的单一模型并进行预测；（4）使用修正的预测误差进行误差倒数法赋权，搭建 GXS 模型对沪深 300 指数每日开盘价进行预测；（5）比较原始单一模型、改进单一模型、组合模型的预测结果差异。

2. 实验组合模型计算方法

GXS 组合模型计算方法如图 2 所示。

算法： GXS

输入： GridSearchCV 算法改进后单一模型及修正误差

输出： GXS 组合模型

改进 XGBoost 模型预测值及预测误差←f_1、ε_1

改进 SVR 模型预测值及预测误差←f_2、ε_2

改进 XGBoost 模型在组合预测模型中权重：

$$\omega_1\leftarrow\frac{\varepsilon_2}{\varepsilon_1+\varepsilon_2}$$

改进 SVR 模型在组合预测模型中权重：

$$\omega_2\leftarrow\frac{\varepsilon_1}{\varepsilon_1+\varepsilon_2}$$

GXS 组合模型预测值 $f\leftarrow\omega_1f_1+\omega_2f_2$

end

输出： GXS 组合模型

图 2　GXS 组合模型计算方法

Fig. 2　Calculation method of GXS combination model

三、实验分析

（一）实验环境和实验数据

本次实验使用的 python 开发环境为 Jupyter Notebook，python 版本为 3.7.4，实验用到 python 中的 numpy、pandas、sklearn、xgboost、matplotlib、numba 等包。选取 2005 年 4 月 8 日-2014 年 6 月 11 日的沪深 300 指数[引 19]，共 2 228 行有效数据，指标选取采用常规基本面分析指标：每日开盘价、收盘价、最高价、最低价、成交量、成交价，数据爬取自聚宽平台①。使用 pandas 包中的 dropna 函数，将节假日等缺失空白数据滤除。为消除指标不同量纲对预测的影响，对数据进行归一化处理，再将数据集按照 80%进行划分：训练集为前 1 782 条数据，测试集为后 446 条数据。沪深 300 指数归一化每日开盘价涨跌趋势如图 3。

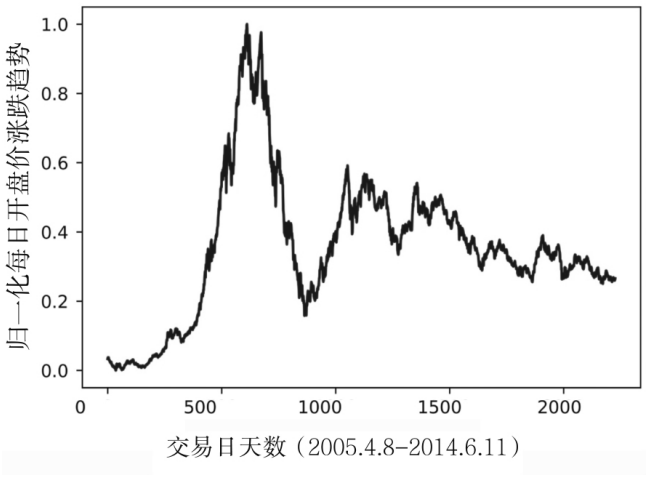

图 3　沪深 300 指数归一化每日开盘价涨跌趋势

Fig. 3　Normalized daily opening prices of the CSI 300 Index

（二）模型回归预测评价指标

选取拟合优度（R^2）、平均绝对误差（MAE）和均方根误差（RMSE）三个评价指标对实验结果进行对比分析。

$$R^2=1-\frac{\sum_{i=1}^{n}(y^{(i)}-\hat{y}^{(i)})^2}{\sum_{i=1}^{n}(y^{(i)}-\bar{y})^2} \tag{14}$$

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|y^{(i)}-\hat{y}^{(i)}| \tag{15}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y^{(i)}-\hat{y}^{(i)})^2} \tag{16}$$

式中，$y^{(i)}$ 为真实值，$\hat{y}^{(i)}$ 为预测值，$\bar{y}$ 为真实均值。

（三）单一模型实验结果与分析

实验一，搭建单一模型：XGBoost 模型、SVR 模型。使用模型初始化参数进行训练，再用测试集进行预测。本文选取的各模型关键参数初始化值，SVR 模型中：错误分类惩罚参数（C）取值为 1，核函数系数（γ）取值为 0.2；XGBoost 模型中：最大树深（max_depth）取值为 3，树的棵数（n_estimators）取值为 100，学习率（learning_rate）取值为 0.1，最小叶子权重（min_child_weight）取值为 1。单一模型预测结果见图 4。

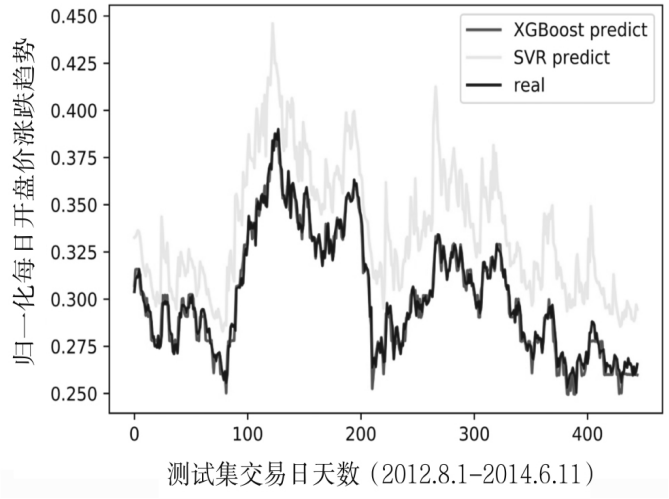

图 4 初始化参数下单一模型对沪深 300 指数开盘价预测

Fig. 4 Forecast of the opening price of the CSI 300 Index by a single model under the initialization parameters

实验二，对单一模型参数进行网格搜索，修正预测误差。在参数初始值的基础上，对取值设定范围。SVR 模型中设置 C、γ 在 $[10^{-2}, 10]$，步长设置为 0.01，进行 10 折交叉验证网格搜索寻优。设置 XGBoost 模型参数取值如表 1。

表 1 XGBoost 模型参数网格搜索取值范围

Tab. 1 Grid search value range of XGBoost model parameters

范围	max_depth（最大树深）	n_estimators（树的棵数）	learning_rate（学习率）	min_child_weight（最小叶子权重）
步长	1	50	0.01	1
取值范围	[1，5]	[100，300]	[0.01，0.1]	[1，5]

经过网格搜索确定 XGBoost 模型参数为：最大树深（max_depth）设置为 4，树的棵数（n_estimators）设置为 300，学习率（learning_rate）设置为 0.08，最小叶子权重（min_child_weight）设置为 1；SVR 模型参数为：错误分类惩罚参数（C）设置为 1.38，核函数参数（γ）设置为 0.63。改进单一模型与原始模型预测结果比较见图 5、图 6。

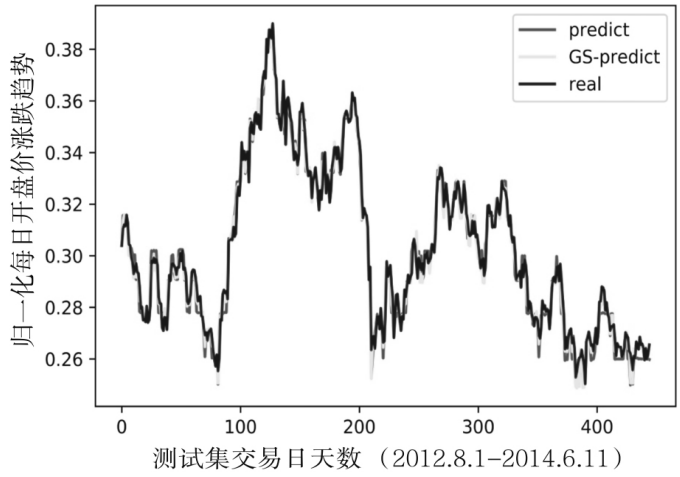

图 5　参数优化前后 XGBoost 模型对沪深 300 指数开盘价预测

Fig. 5　Prediction of the opening price of the CSI 300 Index by the XGBoost model before and after parameter optimization

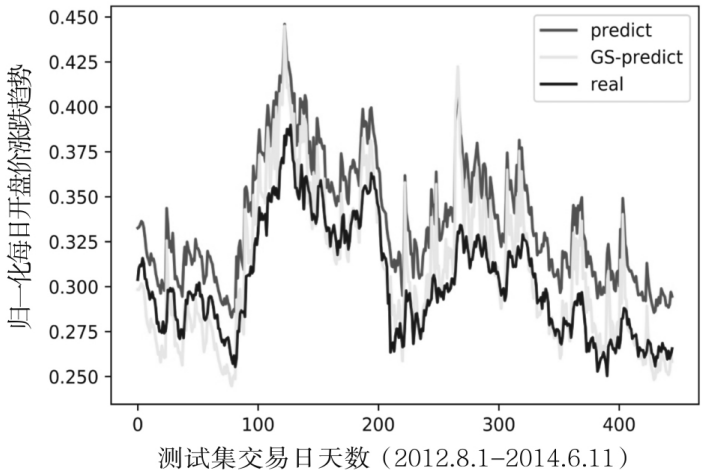

图 6　参数优化前后 SVR 模型对沪深 300 指数开盘价预测

Fig. 6　Prediction of the opening price of the CSI 300 Index by the SVR model before and after parameter optimization

经过参数优化后，模型预测误差得到改善，误差修正结果如表 2。

表 2　单一模型预测结果对比

Tab. 2　Comparison of single model prediction results

预测模型	模型预测评价指标		
	R^2	MAE	RMSE
XGBoost	0. 978 133	0. 003 591	0. 004 596
SVR	0. 278 523	0. 025 704	0. 000 697
改进 XGBoost	0. 984 014	0. 003 000	0. 003 930
改进 SVR	0. 849 622	0. 008 348	0. 000 145

将 XGBoost 模型使用网格搜索算法对参数优化后的修正预测误差记为 GS-MAE、GS-RMSE，图 7 为 XGBoost 模型改进前后预测误差比较。

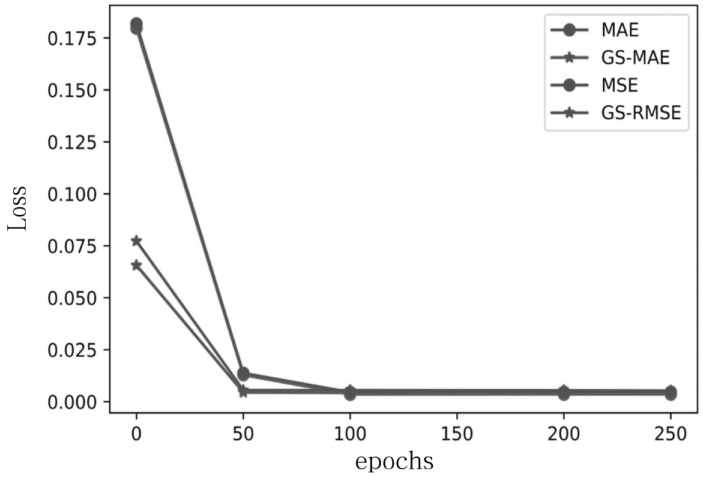

图 7 XGBoost 模型改进前后预测误差

Fig. 7 Prediction error before and after the improvement of XGBoost model

经过网格搜索算法对单一模型参数进行优化后，改进的 XGBoost 模型、SVR 模型拟合效果得到进一步提升，预测误差 MAE 分别减小 16.46%、67.52%，RMSE 分别减小 14.49%、79.20%，预测误差得到有效修正。

（四）组合模型实验结果与分析

考虑到组合模型在设置权重时应该发挥各个模型的优势，当权重出现严重倾向时，会导致个别模型在组合模型中的作用被放大，削弱了其他模型在组合模型中独有的学习能力，不利于组合模型的预测效果。根据公式（11），基于表 2 中改进 XGBoost 模型、改进 SVR 模型用修正预测误差 MAE，分别计算出在组合模型中的权重ω_1为 0.735 6、权重ω_2为 0.264 4；改进 XGBoost 模型、改进 SVR 模型用修正预测误差 RMSE，分别计算出在组合模型中的权重ω_1为 0.035 6、权重ω_2为 0.964 4，发现该组合模型中权重分配存在严重偏向，对改进 SVR 模型的学习程度远大于改进 XGBoost 模型。通过实验进一步考察基于修正预测误差 MAE、RMSE 的两个组合模型对沪深 300 指数开盘价的预测能力。

图 8 为两种修正预测误差 MAE、RMSE 下，GXS 组合模型的回归预测结果。

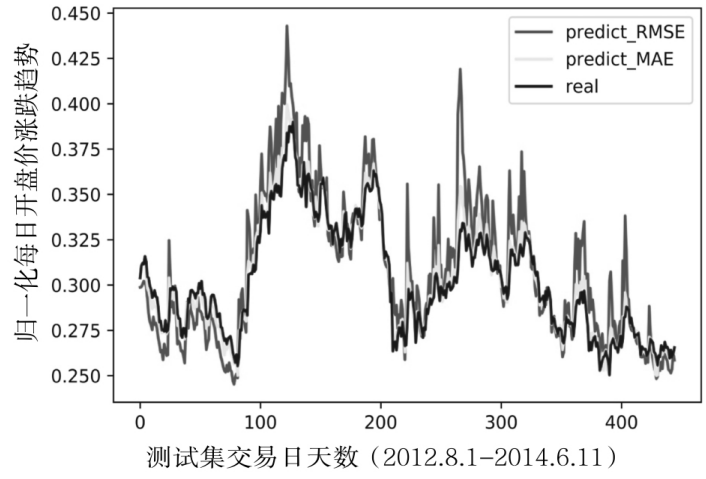

图 8 GXS 组合模型对沪深 300 指数开盘价预测

Fig. 8 GXS portfolio model forecast of the opening price of the CSI 300 Index

将在两种修正预测误差 MAE、RMSE 下分别赋权的组合模型与改进的单一模型预测效果进行比较（见表 3）。基于修正误差 MAE 赋权的 GXS 组合模型预测结果较改进 XGBoost 模型在拟合优度、模型预测误差上都有所提升。与改进 XGBoost 模型相比，MAE 下降 16.37%，RMSE 下降 91.30%；与改进 SVR 模型相比，MAE 下降 69.94%，RMSE 上升 57.60%，拟合优度上升 17.53%。基于修正误差 RMSE 赋权的 GXS 组合模型预测结果较改进 SVR 模型在拟合优度、模型预测误差上都有所提升，MAE 下降 3.70%，RMSE 下降 6.21%；与改进 XGBoost 模型相比，MAE 上升 169.17%，RMSE 下降 96.54%，拟合优度下降 12.68%。

表 3　不同模型预测结果对比

Tab. 3　Comparison of prediction results of different models

预测模型	模型预测评价指标		
	R^2	MAE	RMSE
改进单一模型			
改进 XGBoost 模型	0.984 014	0.003 000	0.003 930
改进 SVR 模型	0.849 622	0.008 348	0.000 145
GXS 组合模型			
基于修正误差 MAE 预测	0.998 601	0.002 509	0.000 342
基于修正误差 RMSE 预测	0.859 194	0.008 075	0.000 136

通过分析组合模型的预测表现，与王芳[7]18-42利用网格搜索算法优化参数后的预测结果进行比较，发现基于修正误差 MAE 赋权的 GXS 组合模型对沪深 300 指数开盘价预测效果更优，拟合优度提升 0.02%，RMSE 下降 98.41%。在组合模型中，基于修正误差 MAE 计算出的权重ω_1为 0.735 6，ω_2为 0.264 4，对改进 XGBoost 模型、改进 SVR 模型都进行训练学习，使组合模型在预测中的能力达到最优。通过表 3 看到，考虑到预测时间成本，改进 XGBoost 模型的预测能力比在修正误差 RMSE 赋权下的组合模型预测能力更强。实验结果表明，权重设置直接影响组合模型的学习能力，对多模型进行组合时，避免选择权重存在严重倾向的组合模型，综合纳入了各个模型的组合模型学习能力更强。

四、结论与展望

本文使用网格搜索算法分别对单一模型：XGBoost 模型、SVR 模型进行参数寻优，改进后单一模型的预测能力都得到明显提升，使用误差倒数法对改进单一模型的修正预测误差 MAE、RMSE 进行赋权，实证分析表明，权重存在严重偏向的 GXS 组合模型预测表现欠佳，综合考虑参数优化前后及组合模型的预测表现，发现基于修正误差 MAE 赋权的 GXS 组合模型对沪深 300 指数开盘价预测表现最优。本文在指标选取方面，仅选取沪深 300 指数的基本面指标：每日收盘价、每日最高价、每日最低价、每日成交量和每日成交额对沪深 300 指数的每日开盘价进行回归预测。在未来的研究中，可以从更加全面的角度进行综合考虑，结合技术指标、财务指标和社会舆论关注度等方面，深入研究各指标对沪深 300 指数开盘价的影响。

注释：

①数据来源：聚宽平台 https://www.joinquant.com/default/research/index?target=self&url=/default/research.

参考文献：

[1] THAKUR M，KUMAR D. A hybrid financial trading support system using multi-category classifier and random forset [J]. Applied Soft Computing，2018，67(1)：337-349.

[2] KRAUSS C，DO X A，HUCK N. Deep neural networks，gradient-boosted trees，random forests：statistical arbitrage on

the S&P 500 [J]. European Journal of Operational Research, 2017,259(2):689-702.
[3] 谢琪,程耕国,徐旭. 基于神经网络集成学习股票预测模型的研究 [J]. 计算机工程与应用,2019,55(8):238-243.
[4] 黄卿,谢合亮. 机器学习方法在股指期货预测中的应用研究——基于 BP 神经网络、SVM 和 XGBoost 的比较分析 [J]. 数学的实践与认识,2018,48(8):297-307.
[5] 王芊. 基于机器学习算法的股票收益率方向预测及分析 [D]. 合肥:中国科学技术大学,2019.
[6] 王燕,郭元凯. 改进的 XGBoost 模型在股票预测中的应用 [J]. 计算机工程与应用,2019,55(20):202-207.
[7] 王芳. 基于支持向量机的沪深 300 指数回归预测 [D]. 济南:山东大学,2015.
[8] 郑明,李娌芝,官心果,等. 基于模糊信息粒化和支持向量机的股票价格回归预测 [J]. 云南民族大学学报(自然科学版),2018,27(6):517-524.
[9] 赛英,张凤廷,张涛. 基于支持向量机的中国股指期货回归预测研究 [J]. 中国管理科学,2013,21(3):35-39.
[10] 魏勤,张宇霖. 基于 SVM 神经网络的沪深 300 股指期货的实证研究 [J]. 产业与科技论坛,2012,10(11):123-126.
[11] VAPNIK V. 统计学习理论的本质 [M]. 张学工,译. 北京:清华大学出版社,2000.

Research on the Prediction Method of CSI 300 Index Based on Inverted Error Combination Optimization Algorithm

XIE Chengyan, FANG Hongbin, GUO Mengjie, YANG Mengzhuo

(School of Economics, Anhui University, Hefei 230601, China)

Abstract: Aiming at the problem of multivariate prediction of stock index opening price, in order to improve the prediction accuracy, a GXS combination model based on inversion error is proposed to make regression prediction of daily opening price of CSI 300 index. In this paper, GridSearchCV algorithm and 10-fold cross validation are used to optimize the parameters of the extreme gradient boost tree (XGBoost) model and the support vector regression (SVR) model based on the radial basis RBF kernel function, and the corrected prediction error is weighted by the error inverse method to build the GXS composite model. The results of empirical analysis show that the GXS combination model based on the modified error MAE weighting has the best prediction effect on the opening price of CSI 300 index.

Key words: CSI 300 index; XGBoost model; the SVR model; inverted error method; regression prediction

(责任编辑: 杨成平)