计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200600002

基于遗传算法与随机森林的 XGBoost 改进方法研究

王晓晖[1]　张　亮[1]　李俊清[1,2]　孙玉翠[1]　田　捷[1]　韩睿毅[1]
1 山东农业大学信息科学与工程学院　山东 泰安 271018
2 山东农业大学农业大数据研究中心　山东 泰安 271018
(wangxh1998@163.com)

摘　要　回归预测是机器学习中重要的研究方向之一,有着广阔的应用领域。为了进一步提升回归预测的精度,提出了基于遗传算法与随机森林的 XGBoost 改进方法(GA_XGBoost_RF)。首先利用遗传算法(Genetic Algorithm,GA)良好的搜索能力和灵活性,以交叉验证平均得分为目标函数值,对 XGBoost 算法和随机森林算法(Random Forest,RF)的参数进行调优,选出较好的参数集,分别建立 GA_XGBoost 和 GA_RF 模型。然后对 GA_XGBoost 和 GA_RF 进行变权组合,利用训练集的预测值与真实值的均方误差为目标函数,使用遗传算法确定模型的权重。在 UCI 数据集上进行了实验,结果表明,与 XGBoost,Random Forest,GA_XGBoost,GA_RF 算法相比,在大部分数据集上 GA_XGBoost_RF 方法的均方误差、绝对误差和拟合度均优于单一模型,其中在拟合度方面所提方法在不同数据集上提高了约 0.01%～2.1%,是一种有效的回归预测方法。

关键词:回归预测;XGBoost;组合预测;随机森林;遗传算法
中图法分类号　TP181

Study on XGBoost Improved Method Based on Genetic Algorithm and Random Forest

WANG Xiao-hui[1],ZHANG Liang[1],LI Jun-qing[1,2],SUN Yu-cui[1],TIAN Jie[1] and HAN Rui-yi[1]
1 School of Information Science and Engineering,Shandong Agricultural University,Taian,Shangdong 271018,China
2 Agricultural Big Data Research Center,Shandong Agricultural University,Taian,Shangdong 271018,China

Abstract　Regression prediction is one of the important research directions in machine learning and has a broad application field. In order to improve the accuracy of regression prediction,an improved XGBoost method (GA_XGBoost_RF) based on genetic algorithm and random forest is proposed. Firstly,with the good search ability and flexibility of Genetic Algorithm (GA),the XGBoost Algorithm and Random Forest Algorithm (RF) parameters are optimized with the average score of cross-validation as the objective function value,and the better parameter set is selected to establish GA_XGBoost and GA_RF models,respectively. Then the variable weight combination of GA_XGBoost and GA_RF is performed. The mean square error between the predicted value and the real value of the training set is used as the objective function,and the weight of the model is determined by genetic algorithm. On UCI data sets and the results show that the XGBoost and Random Forest,GA_XGBoost,GA_RF algorithm compared to GA_XGBoost_RF method in most of the data set is the fit of the mean square error (mse) and absolute error and are superior to single model,the proposed method on fitting on different data sets improves by about 0.01%～2.1%,is a kind of effective regression forecast method.

Keywords　Regression prediction,XGBoost,Combination prediction,Random forest,Genetic algorithm

1　引言

机器学习中,回归预测是重要的研究方向之一,回归预测在日常生活中有着广阔的应用领域,如房价预测[1]、疾病预测[2]、股票预测[3]等。XGBoost 算法由 Tianqi Chen 提出,是目前被数据科学家广泛使用的机器学习算法之一,在众多机器学习竞赛中获得了良好的成绩[4]。XGBoost 算法具有多种优点:快速处理,接受多种类型的输入数据,内置交叉验证,树剪枝,高度灵活,较其他增强模型能更好地控制过拟合[5]。XGBoost 算法已在众多预测领域中取得到了应用[6]。

目前,不少学者基于 XGboost 算法提出了相关的改进方法,并且有着较好的预测能力。Zhang 等提出了一种 GA_XGBoost 模型,利用遗传算法的寻优能力对 XGBoost 的参数组合进行了多次搜索得到近优解,结果显示 GA_XGBoost 模型与其他算法相比有着较好的预测效果,并且在运行时间上优于其他调参方法[2]。Chen 等提出了基于 LSTM 和 XGBoost 的组合预测方法,利用误差倒数法将 LSTM 和 XGBoost 进行加权组合,建立了 LSTM-XGBoost 组合预测模型,实验结果表明该方法的预测精度明显优于单一模型[6]。Li 等为提升 XGBoost 算法在不平衡数据集的预测精度,利用梯度分布调节策略对 XGBoost 算法进行了改进,通过计算每个样本的损失贡献密度,使算法向少数类和难以分类的样本进

基金项目:大数据驱动下流域水库群联合防洪调度研究(2019GSF111043)
This work was supported by the Joint Flood Control Operation of Reservoir Groups in River Basin Driven by Digdata (2019GSF111043).
通信作者:李俊清(a397858801@126.com)

行倾斜，达到了提升不平衡分类准确率的目的[7]。Yue 等提出了一种 XLC-Stacking 方法，使用 Stacking 方法集成 XGBoost，LightGBM，CatBoost 等 6 种异质分类算法，有效地提升了预测的精度[8]。Wang 等提出了一种 CNN-XGBoost 混合预测模型，该模型使用卷积神经网络对复杂的交通流数据进行特征提取，得到卷积神经网络输出的高维特征向量，并将其作为 XGBoost 模型的输入向量进行预测[9]。

部分改进方法也存在一定的不足之处。如 GA_XGBoost 算法在利用遗传算法进行参数寻优的过程中使用了测试集，影响了算法的泛化能力，并预测算法单一，未利用其他算法与 XGBoost 算法进行融合来提升算法的预测性能。LSTM-XGBoost 组合预测模型的组合方式过于简单，使用该方法所得到的权值并不一定是最优值，预测效果还有提升的空间。基于梯度分布调节策略的 XGBoost 优化算法在不平衡分类问题中优于其他常见算法，但是实验表明在回归问题中可能存在过拟合现象。

本文利用遗传算法调节 XGBoost 算法和 Random Forest 算法的超参数，采用交叉验证平均得分作为遗传算法的目标函数值，建立 GA_XGBoost 和 GA_RF 模型，再次利用遗传算法确定 XGBoost 算法和 Random Forest 算法的权值，最终建立 GA_XGBoost_RF 组合预测模型。本文采用 UCI 数据集验证所提方法的有效性，结果表明本文提出的改进方法与单一算法相比能够明显提升回归预测的精度。

2 相关理论介绍

2.1 XGBoost 算法

XGBoost(eXtreme Gradient Boosting)即极端梯度提升，是一种提升学习算法。该算法在梯度提升决策树(GDBT)算法的基础上对损失函数进行二阶泰勒展开并且添加了正则项，有效地避免了过拟合同时加快了收敛的速度。XGBoost 算法通过不断形成新的决策树来拟合之前预测的残差，使得预测值与真实值的残差不断减小，从而提高预测精度。XGBoost 算法可以表示成一种加法的形式，如式(1)所示：

$$\hat{y}_i=\sum_{k=1}^{K}f_k(x_i),f_k\in F \tag{1}$$

其中，$\hat{y}_i$ 表示模型的预测值；K 表示决策树的数目；f_k 表示第 k 个子模型；x_i 表示第 i 个输入样本；F 表示所有的决策树集合。XGBoost 的目标函数由损失函数和正则项两个部分组成：

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

$$\Omega(f_k)=\gamma T+\frac{1}{2}\lambda\parallel\omega\parallel^2 \tag{3}$$

其中，$\mathcal{L}(\varphi)^t$ 表示第 t 次迭代的目标函数；$\hat{y}_i^{(t-1)}$ 表示前 $t-1$ 次迭代的预测值；$\Omega(f_k)$ 表示第 t 次迭代的模型的正则项，起到了减少过拟合的作用；γ 和 λ 表示正则项系数，防止决策树过于复杂；T 表示该模型的叶结点个数。

对式(2)所示的目标函数使用泰勒公式展开可得：

$$\begin{aligned}\mathcal{L}(\phi)^t&\cong\sum_{i=1}^{n}[g_if_t(x_i)+\frac{1}{2}h_if_t^2(x_i)]+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2\\&\cong\sum_{j=1}^{T}[(\sum_{i\in I_j}g_i)\omega_j+\frac{1}{2}(\sum_{i\in I_j}h_i+\lambda)\omega_j^2]+\gamma T\end{aligned} \tag{4}$$

其中，g_i 表示样本 x_i 的一阶导数；h_i 表示样本 x_i 的二阶导数；ω_j 表示第 j 个叶子结点的输出值，I_j 表示第 j 个叶子结点值的样本子集。

由式(4)可以看出，目标函数是一个凸函数，对 ω_j 求导并且令导函数等于零，可求得使目标函数达到最小值的 ω_j，即：

$$\omega_j^*=-\frac{\sum_{i\in I_j}g_i}{\sum_{i\in I_j}h_i+\lambda} \tag{5}$$

$$\tilde{\mathcal{L}}(\phi)_{\min}=-\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in I_j}g_i)^2}{\sum_{i\in I_j}h_i+\lambda}+\gamma T \tag{6}$$

式(6)可以用来评价一颗树模型的好坏，其值越小，说明树模型越好。由此可以得出用于树进行结点分裂的得分公式：

$$Gain=\frac{1}{2}\left[\frac{(\sum_{i\in I_L}g_i)^2}{\sum_{i\in I_L}h_i+\lambda}+\frac{(\sum_{i\in I_R}g_i)^2}{\sum_{i\in I_R}h_i+\lambda}-\frac{(\sum_{i\in I}g_i)^2}{\sum_{i\in I}h_i+\lambda}\right]-\gamma \tag{7}$$

式(7)被用来计算树模型的分裂结点。

2.2 随机森林算法

随机森林(Random Forest，RF)算法是一种基于 Bagging 的集成学习方法。该算法结合了 Breimans 的 Bootstrap aggregating 和 Tin Kam Ho 的 random decision forests 方法[9]，从总体数据集中多次随机取一部分样本构成样本簇，在每一个新生成的样本簇上训练决策树，组合每一棵树的结果最终决定输出，从而解决了决策树泛化能力弱的缺点，提升了最终模型的预测效果。随机森林算法的具体步骤如算法 1 所示。

算法 1 随机森林算法

输入：$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$

输出：随机森林$\{Tree(x,\theta_1),Tree(x,\theta_2),\cdots,Tree(x,\theta_m)\}$

Step1 对样本集 D 进行 N 次 Bootstrap 抽样，形成一个新的样本集。

Step2 从众多特征中进行随机选取，形成特征子集。

Step3 在新样本集和特征子集上，找到最佳分割属性，建立决策树 $Tree(x,\theta_i)$。

Step4 重复 Step1－Step3，直到构建出 m 棵决策树。

Step5 组合 m 棵决策树的输出结果，建立随机森林模型。

对于数据量为 N 的样本集 D 而言，每个样本被抽到的概率为 $1/N$，则没有被抽到的概率为 $1-1/N$，对样本集 D 进行 N 次自助法抽样，则单个样本没有被抽到的概率为 $(1-1/N)^N$，当 N 足够大时，每个样本没有被抽到的概率为 $\lim_{N\to\infty}(1-1/N)^N\approx0.367$，说明数据集 D 中有 36.7%的样本没有被抽中，这部分样本称为袋外数据(Out Of Bag，OOB)，可用于对模型进行评价。

2.3 遗传算法

遗传算法(Genetic Algorithm，GA)，也称为进化算法。遗传算法借鉴了达尔文的种群进化理论和基因遗传进化理论，通过模拟种群的自然淘汰和个体基因的遗传变异过程来达到搜索问题最优解的方法。在遗传算法中种群的每个个体都是解空间上的一个可行解，通过模拟生物的进化过程，种群中的个体不断进行选择、交叉、变异，从而在解空间内自适应地搜索最优解[2]。

遗传算法是对种群中可行解进行交叉、变异操作，因此遗传算法的目标函数并不需要可导或者连续条件。遗传算法采用概率化的寻优方法，自动获取和指导优化的搜索空间，自适

应地调整搜索方向。遗传算法简单、通用,适于并行处理。该算法的具体步骤如算法 2 所示。

算法 2 遗传算法

输入:目标函数和约束条件

输出:历史最优解

Step1 选择一种合适的编码方案。

Step2 在解空间内随机生成一个种群作为问题的初代解。

Step3 利用适应度函数计算种群中每个个体的适应度。

Step4 根据适应度的高低选择参与繁衍的个体,将适应度高的个体保留,淘汰适应度低的个体。

Step5 对被保留个体的基因执行交叉操作,生成子代。

Step6 对生成的子代基因进行随机变异操作,增加基因的多样性。

Step7 淘汰种群中适应度较低的个体。

Step8 重复 Step3－Step7,直到满足结束条件为止。

Step9 选择所有子代中适应度最高的个体作为问题的最优解。

3 GA_XGBoost_RF 算法

本文利用遗传算法(GA)和随机森林(Random Forest, RF)算法与极端梯度提升(XGBoost)算法进行组合。XGBoost 算法是近年来兴起的一种高效集成学习方法[7],已在众多预测领域中取得了应用[6]。随机森林是一种基于 Bagging 的集成学习方法[16],XGBoost 算法属于 Boosting 集成学习方法[5],将两种不同的集成学习方法进行组合可以结合两种方法的优点。使用遗传算法对 XGBoost 进行调参在运行时间和调参效果上优于网格搜索和随机游走[2]。并且遗传算法对问题的可行解进行编码,通过适应度来选择判断基因优劣[19],对目标函数没有连续和可导的要求,因此简化了组合模型参数调优和权值调优的复杂程度。

利用遗传算法良好的全局搜索能力和灵活性[2],对 XGBoost 算法和 Random Foerst 算法的参数进行优化,然后利用遗传算法确定 GA_XGBoost 算法和 GA_RF 算法的权值,最终建立 GA_XGBoost_RF 组合预测模型。下面将分别介绍参数调优和权值调优的具体内容。GA_XGBoost_RF 算法的流程如图 1 所示。

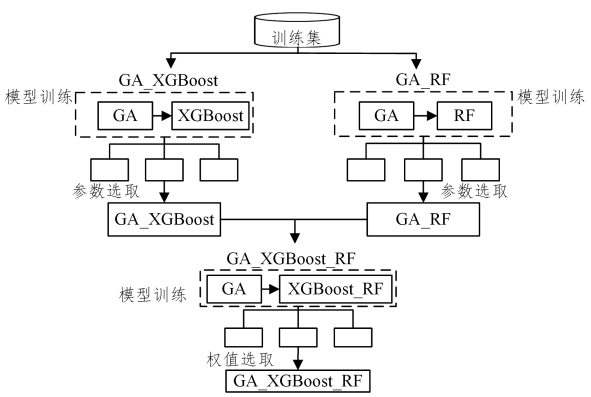

图 1 GA_XGBoost_RF 流程图

Fig. 1 Flowchart for GA XGBoost RF

3.1 参数优化

XGBoost 算法和 RF 算法的参数较多,参数选取直接影响算法的精度,合理的参数设置可以明显提升模型的预测精度。本文利用遗传算法的全局寻优能力对 XGBoost 和 RF 模型进行参数选择,使用交叉验证的平均得分作为适应度函数值,XGBoost 采用 5 折交叉验证,RF 因为有一定的随机性所以采用 7 折交叉验证,建立 GA_XGBoost 和 GA_RF 模型进行算法参数优化。GA_XGBoost(GA_RF)算法的伪代码如算法 3 所示。

算法 3 GA_XGBoost(GA_RF) 算法

Input: population size $\mathscr{P}$, iteration times $\mathscr{T}$, parameter number $\mathscr{N}$, number of outstanding individuals $\mathscr{M}$

output: Optimal combination of parameters

1. Initialize $(\theta_{i1},\theta_{i2},\cdots,\theta_{iN})$ //初始化算法参数$(\theta_{i1},\theta_{i2},\cdots,\theta_{iN})$
2. while termination condition is not met do
3. Cross validation on the XGBoost(RF) model //用 XGBoost(RF)进行交叉验证计算适应度
4. Calculate fitness value //根据适应度选出 M 最好算法参数组合
5. Select the optimal parameters of Mgroup according to fitness value //根据适应度选出 M 组最好算法参数组合
6. GA$(\mathscr{P},\mathscr{T},\mathscr{N},\mathscr{M},(\theta_{i1},\theta_{i2},\cdots,\theta_{iN}))$ //进行遗传、变异等运算
7. Produce new parameters //生成新的算法参数组合
8. end while

3.2 权值优化

对调参后的 GA_XGBoost 模型和 GA_RF 模型建立变权组合预测模型,其中最重要的是两个模型各自权值的确定,本文利用遗传算法来确定两个模型的权值。

首先利用 3.1 节中得到的参数组合在训练集上进行训练,构建 GA_XGBoost 和 GA_RF 模型。

其次使用遗传算法优化权值,设置权值参数的范围、遗传算法的迭代次数、初始种群的数量,然后随机生成 $\mathscr{P}$ 组初始值,若不满足停止条件,则将种群中的每个个体作为组合模型的权值对训练集进行预测,以真实值与预测值的均方误差为适应度函数(如式(8)所示),以权值之和等于 1 为约束条件(如式(9)所示),从种群中选取 $\boldsymbol{M}$ 个优秀个体,对选取的优秀个体进行交叉、变异,从而产生新的个体,循环这个过程直到满足条件时停止,从历代种群中选择最优值作为最终结果,得到组合模型的权值组合,建立 GA_XGBoost_RF 模型。GA_XGBoost_RF 算法如算法 4 所示。

$$fit=\sqrt{\frac{1}{m}\left(w_1\hat{y}_{\mathrm{XGB}}+w_2\hat{y}_{\mathrm{RF}}-y_{\mathrm{True}}\right)^2} \tag{8}$$

$$w_1+w_2=1 \tag{9}$$

其中,m 为训练集中的样本个数,w_1 和 w_2 分别为 XGBoost 和 RF 模型的权值,$\hat{y}_{\mathrm{XGB}}$ 和 $\hat{y}_{\mathrm{RF}}$ 分别为 XGBoost 和 RF 模型对训练集的预测值,y_{True} 为训练集的真实值。

算法 4 GA_XGBoost_RF 算法

Input: population size $\mathscr{P}$, iteration times $\mathscr{T}$, number of outstanding individuals $\mathscr{M}$

output: The optimal weights

1. The Train Set train GA_XGBoost model and GA_RF model //训练 GA_XGBoost 和 GA_RF 模型
2. Initialize (w_{i1},w_{i2}) //初始化权重
3. while termination condition is not met do
4. Predict Train Set on XGBoost_RF model //预测训练集
5. Calculate fitness value //计算适应度
6. Select the optimal parameters of $\mathscr{M}$ group according to fitness value //根据适应度选出 u 组最好权重组合
7. GA$(\mathscr{P},\mathscr{T},\mathscr{M},(w_{i1},w_{i2}))$ // 进行遗传、变异等运算

8. Produce new (w_{i1}, w_{i2}) //生成新的权重组合
9. end while

4 模型评估

4.1 实验数据集

为了验证提出方法的有效性，本文选取了多个 UCI 数据集。UCI 数据库是加州大学欧文分校（University of CaliforniaIrvine）提出的用于机器学习的数据库，UCI 数据集是一个常用的标准测试数据集[7]。另外，本文还选取了回归预测实验较为常用的波士顿房价数据集，该数据集由常用的机器学习第三方模块 sklearn 提供。数据集如表 1、表 2 所列。

表 1 UCI 数据集

Table 1 UCI dataset

Serial number	DateSet	Sample size	Feature size
01	Combined CyclePower Plant	9568	4
02	Computer Hardware	209	9
03	Parkinsons Telemonitoring	5875	20
04	Yacht Hydrodynamics	308	7
05	Residential Building	372	105

表 2 Boston house-prices 数据集

Table 2 Boston house-prices dataset

Serial number	DataSet	Sample size	Feature size
06	Bostonhouse-prices	506	13

4.2 实验评价指标

在回归预测问题中，常用的评价指标有均方误差（Mean Squared Error，MSE）、均方根误差（Root Mean Squared Error，RMSE）、平均绝对误差（Mean Absolute Error，MAE）、决定系数（Coefficient of Determination，R^2），评价指标的具体公式如下：

$$MSE=\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2 \tag{10}$$

$$RMSE=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2} \tag{11}$$

$$MAE=\frac{1}{m}\sum_{i=1}^{m}|y_i-\hat{y}_i| \tag{12}$$

$$R^2=1-\frac{\sum_{i=1}^{m}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{m}(y_i-\bar{y}_i)^2} \tag{13}$$

通过式（10）与式（11）可以看出，$RMSE=\sqrt{MSE}$，两个指标有一定的重复性，因此本文采用 MSE，MAE，R^2 为评价指标来验证模型的有效性。

4.3 实验分析

基于上述数据集和实验评价指标，设计回归预测实验，将本文提出的改进方法 GA_XGBoost_RF 与默认参数的 XGBoost 和 Random Foerest 算法、用遗传算法调参之后的 GA_XGBoost 算法和 GA_RF 算法进行比较，结果如表 3 所列。

表 3 GA_XGBoost_RF 方法回归预测比较实验结果

Table 3 GA XGBoost RF method is used to compare experimental results

数据集与评价指标		XGBoost	RF	GA_XGBoost	GA_RF	GA_XGBoost_RF
01	*MSE*	9.2415	10.5464	8.1321	9.7158	**8.1069**
	MAE	2.2799	2.4204	2.1054	2.3369	**2.1049**
	R^2	0.9678	0.9633	0.9717	0.9662	**0.9718**
02	*MSE*	1119.8	869.46	1003.9	856.91	**831.29**
	MAE	24.069	22.036	24.068	21.354	**21.144**
	R^2	0.9235	0.9406	0.9314	0.9414	**0.9432**
03	*MSE*	0.132%	0.121%	0.115%	0.120%	**0.114%**
	MAE	0.0266	0.0258	0.0253	0.0258	**0.0251**
	R^2	0.8477	0.8610	0.8673	0.8620	**0.8693**
04	*MSE*	0.4923	0.4522	0.4573	0.4073	**0.2304**
	MAE	0.2927	0.3752	0.3521	0.3583	0.2950
	R^2	0.9975	0.9977	0.9977	0.9980	**0.9988**
05	*MSE*	40939	53479	41243	55810	**40891**
	MAE	**103.49**	130.37	108.10	129.32	108.87
	R^2	0.9733	0.9651	0.9731	0.9636	**0.9733**
	MSE	748.39	676.30	692.69	628.71	**571.92**
	MAE	16.456	18.272	**14.794**	17.724	15.225
	R^2	0.9622	0.9658	0.9650	0.9682	**0.9711**
06	MSE	10.014	9.1475	8.7936	8.6095	**8.1909**
	MAE	2.1766	2.1160	2.0902	2.0533	**1.9642**
	R^2	0.8840	0.8940	0.8981	0.9002	**0.9051**

从表 5 可以看出，本文提出的 GA_XGBoost_RF 方法可以有效提高大部分数据集的预测精度，其中在 Combined Cycle Power Plant，Computer Hardware，Boston house-prices，Parkinsons Telemonitoring 4 个数据集中，GA_XGBoost_RF 方法与另外 4 种算法相比，3 个评价指标均有明显提升。在 Yacht Hydrodynamics 数据集中，GA_XGBoost_RF 方法在两项指标中的预测精度有所提高，其中明显提升了预测的均方误差，与其他算法相比提升了 17.69%～26.19%，并且其决定系数 R^2 相比其他算法有 0.08%～0.13%的提升，但未能降低平均绝对误差值，与其他算法中的最好值相差 0.23%。在 Residential Building 多输出数据集上，GA_XGBoost_RF 与其他 4 种方法相比，均方误差和 R^2 两项评价指标达到了最好值，在第一项预测中均方误差提升了 48～14 919，R^2 提升了 0～0.97%；在第二项预测中均方误差提升了 56～176.47，R^2 提升了 0.29%～0.89%，但在两项预测中均未能提升绝对误差，与其他算法中的最好值分别相差 5.38，0.431。实验证明，本文提出的 GA_XGBoost_RF 方法利用遗传算法将极端梯度提升算法和随机森林算法进行组合，与单一预测方法相比可以有效地提升回归预测的精度。

结束语 针对回归预测问题，本文选用 XGBoost 算法和 RF 算法的组合预测方式提高预测的精度，首先利用遗传算法的良好的寻优能力，确定 XGBoost、RF 算法的超参数，然后再

次利用遗传算法确定两种模型各自的权值。实验表明，该方法有效地提升了大部分数据集的预测能力，只有个别指标未达到最优值，原因是在遗传算法进行优化权重的过程中，以均方误差为目标函数，未兼顾到其他评价指标。在后续研究中，将继续开展改进工作，以更好地提升预测性能。

参 考 文 献

[1] YUAN B, LIU S, JIANG L X, et al. Housing rent prediction model based on random forest regression algorithm[J]. Computer Programming Skills & Maintenance, 2020(1): 23-25.

[2] ZHANG C F, WANG S, WU Y D, et al. Diabetes Risk Prediction Based on GA_Xgboost Model[J]. Computer Engineering, 2020(3): 315-320.

[3] WANG Y, GUO Y K. Application of Improved XGBoost Model in Stock Forecasting[J]. Computer Engineering and Applications, 2019(20): 202-207.

[4] 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. ACM, 2016: 785-794.

[5] CHEN H, WANG R T, XIAO C L, et al. Research on Intrusion Detection Model Based on DBN-XGBDT[J/OL]. Computer Engineering and Application. https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CAPJ&dbname=CAPJLAST&filename=JSGG20200107004&v=UVJbamaWiqN%25mmd2F9O2vyqQDdcTYYvCJ1fZFijf%25mmd2FWeamhJm61AxhCjVV6r5HZkDoH4xo.

[6] CHEN Z Y, LIU J B, LI C, et al. Ultra Short-term Power Load Forecasting Based on Combined LSTM and XGBoost Model[J]. Power System Technology, 2020(2): 1-8.

[7] LI H, ZHU Y. Improving Xgboost Based on Gradient Distribution Regulation Strategy[J]. Journal of Computer Applications, 2020(1): 1-6.

[8] YUE P, HOU L Y, YANG D L, et al. XLC-Stacking method for disease diagnosis based on XGBoost feature selection[J]. Computer Engineering and Applications, 2020(17): 136-141.

[9] WANG Q S, XIE X S, SHE H. Short-term Traffic Flow Prediction Based on CNN-XGBoost Hybrid Model[J]. Measurement & Control Technology, 2019(4): 37-40, 67.

[10] LI B, HAN R, HE Y G, et al. Application of Improved Random Forest Algorithm in Fault Diagnosis of Motor Bearings[J]. Proceedings of the CSEE, 2020(4): 1310-1319, 1422.

[11] DING D D, SUI L, CHEN S. Machine learning-dynamically coupled vehicle following models[J]. Journal of Transportation Systems Engineering and Information Technology, 2017(6): 33-39.

[12] YUE Y C, HUANG Y Z. A Method for Error Reciprocal Variable Weight Combined Forecasting[J]. Journal of University of Electronic Science and Technology of China, 2007(S1): 349-351.

[13] ZHOU Y S, CUI J Y, ZHOU L Y, et al. Study on the Evaluation of Personal Credit Risk Based on the Improved Random Forest Model[J]. Credit Reference, 2020(1): 25-30.

[14] SONG K, YAN F, DING T, et al. A steel property optimization model based on the XGBoost algorithm and improved PSO[J]. Computational Materials Science, 2020, 174(C).

[15] SHI X P, WONG Y D, LI Z F, et al. A feature learning approach based on XGBoost for driving assessment and risk prediction [J]. Accident Analysis and Prevention, 2019, 129(129).

[16] LIU Z X, WANG X. Flight Delay Prediction Based on Random Forest Regression[J]. Modern Computer, 2019(15): 20-24.

[17] XIE K, RONG Y T, HU F P, et al. Random Forest based on Data Ensembling[J/OL]. Computer Engineering. https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CAPJ&dbname=CAPJLAST&filename=JSJC20191206002&v=0pB3H536puZ4tfXwxmctFHXG08jgxGF4%25mmd2BPhds%25mmd2BTvGl4wpi4FuIthY5Id9ogKmt1A.

[18] SHI J Q, ZHANG J H. Load Forecasting Based on Multi-model by Stacking Ensemble Learning[J]. Proceedings of the CSEE, 2019(14): 4032-4042.

[19] LIU X Z Y, GAN L, XU J H, et al. Automatic Optimization of Parallel Parameters for Sunway TaihuLight SupercomputerApplication Program[J/OL]. Journal of Frontiers of Computer Science and Technology. https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CAPJ&dbname=CAPJLAST&filename=KXTS20200117000&v=jyKKAwjXo98Ft%25mmd2FhCSfCvhikiL1CADBYEajg0LyXpY1lp8Jk8Psm5yiUOe5IvYF23.

[20] LIU J, CHEN H H, ZHANG F F, et al. Multi-parameter identification of river water quality model based on animproved genetic algorithm[J]. Journal of Northeast Agricultural University, 2020(1): 73-82.

[21] XING Z W, HAN D H, LUO Q. Estimationof Flight Support Time Based on improved GA neural network[J]. Computer Engineering and Design, 2020(1): 107-114.

[22] LIN L C. Improved k-means algorithm based on genetic algorithm[J]. Electronic Technology & Software Engineering, 2020(1): 111-112.

[23] NIU W N, LI T, ZHANG X S, et al. Using XGBoost to Discover Infected Hosts Based on HTTP Traffic[J/OL]. https://schlr.cnki.net/Detail/index/WWMERGEJ02/SJHDD74B5ADB931A22462D32E1F64048A4BC.

[24] ZHONG Y, SHAO Y M, HU W W, et al. Short-term Traffic Flow Prediction Model Based on XGBoost[J]. Science Technology and Engineering, 2019(30): 337-342.

[25] XIE Y, XIANG Y, JI M Z, et al. An application and analysis of forecast housing rentalbased on xgboost and lightgbm algorithms [J]. Computer Applications and Software, 2019(9): 151-155, 191.

[26] WANG M H, LIANG X C. Personal Credit Evaluation Based on CPSO-XGBoost [J]. Computer Engineering and Design, 2019(7): 1891-1895.

[27] HE B, MA J, GAO H Y. A research on forecasting urban daily water-supply based on multi-granularityfeature and XGBoost integrated model[J]. Journal of Yangtze River Scientific Research Institute, 2020(5): 43-49.

[28] LUO X, QIAN Q, FU Y F. Improved Genetic Algorithm for Solving Flexible Job Shop Scheduling Problem[J]. Procedia Computer Science, 2020, 166(166).

[29] MIRALLES-PECHUÁN L, PONCE H, MARTÍNEZ-VILLASEÑOR L. A 2020 perspective on "A novel methodology for optimizing display advertising campaigns using genetic algorithms" [J]. Electronic Commerce Research and Applications, 2020, 40(40).

(下转第 463 页)