Operations Research and Fuzziology 运筹与模糊学, 2021, 11(1), 1-8
Published Online February 2021 in Hans. http://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2021.111001

Hans汉斯

GBRT-组合优化预测模型

——基于重庆市空气质量数据

谢景伊，李成凤

贵州大学数学与统计学院，贵州 贵阳
Email: 812047486@qq.com

收稿日期：2020年12月22日；录用日期：2021年1月22日；发布日期：2021年1月29日

摘　要

在处理多因素数据预测问题时，采用变量选择进行显著因素的筛选，更利于因变量的预测。本次研究的目的是提高重庆市AQI预测的精度，AQI是定量描述空气质量状况的指数，其数值越大说明空气污染状况越严重。对参与空气质量预测的主要6个指标，进行GBRT变量选择余下细颗粒物($PM_{2.5}$)、可吸入颗粒物(PM_{10})、二氧化氮(NO_2)、臭氧(O_3)这4个指标，采用遗传算法改进的最小二乘支持向量机进行组合预测，其预测误差从1.5329164降到0.1993641，表明该组合预测模型于空气质量预测有很好的应用前景。

关键词

空气质量，GBRT算法，遗传算法，最小二乘支持向量机算法，组合预测

GBRT-Combined Optimization Forecasting Model

—Based on Air Quality Data of Chongqing

Jingyi Xie, Chengfeng Li

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou
Email: 812047486@qq.com

Received: Dec. 22nd, 2020; accepted: Jan. 22nd, 2021; published: Jan. 29th, 2021

Abstract

AQI is an index that quantitatively describes the air quality. The larger the AQI is, the more serious the air pollution is. GBRT was used to select the variables of six main indexes involved in air qual-

文章引用: 谢景伊, 李成凤. GBRT-组合优化预测模型[J]. 运筹与模糊学, 2021, 11(1): 1-8.
DOI: 10.12677/orf.2021.111001

ity prediction, and the remaining four indexes were fine particulate matter ($PM_{2.5}$), inhalable particulate matter (PM_{10}), nitrogen dioxide (NO_2) and ozone (O_3). On this basis, the least square support vector machine improved by genetic algorithm is used for combination prediction, and the prediction error is reduced from 1.5329164 to 0.1993641, which indicates that the combination prediction model has a good application prospect in air quality prediction.

Keywords

Air Quality, GBRT Algorithms, Genetic Algorithms, LS-SVMR, Combination Prediction

1. 引言

随着中国乃至世界经济的高速发展，城市化进程也在急剧加快，人们对于生活品质的要求也越来越高。其中空气质量的高低直接或者间接的影响着人们的生活，因此，空气质量的预测可以及时地提醒相关部门调整，对进一步提高人们的生活品质有重要的意义。

在过去对空气质量的预测中，研究者们常用的预测方法主要：张叶娥[1]等采用 ARIMA 模型预测大同市空气质量，但该模型向前预测的时期越长，预测的误差将会越大，不适合于长期预测；宋雨辰等[2]采用 BP 神经网络和时间序列模型预测包头市空气质量，但 BP 神经网络需要大量训练且时间序列适于短时期；王先行、方彦等[3] [4]采用 GM-RBF 组合模型预测空气质量，灰色系统能减少 RBF 神经网络建模精度容易受数据随机性影响的问题，但灰色预测有小样本数据的局限性，组合预测确实提高了预测精度；卢彬、刘君[5] [6]采用因子分析与径向基神经网络结合进行空气质量预测，结果有效提高了收敛速度和预报准确度；胡邦辉等[7]采用最小二乘支持向量机模型云量预测，表明最小二乘支持向量机回归方法的预报效果要优于神经网络，预报准确率也不会因为训练样本的减少而降低，预测前景较好；付莲莲等[8]采用梯度提升回归模型的生猪价格预测，表明，梯度提升回归模型具有较高的预测精度；Shivang Agarwal 等[9]基于实时动态误差修正的人工神经网络预测高污染地区空气质量表明模型在多个评价指标上对所有污染物都有很好的评价效果；谷艳昌等[10]基于遗传算法优化支持向量机对大坝安全性态预测，表明 GA-SVM 模型渗压预测值与实测值最接近，预测精度较 SVM 模型和逐步回归模型提高了约 3 倍。

前面的研究者们对于空气质量的预测都是在原有的细颗粒物($PM_{2.5}$)、可吸入颗粒物(PM_{10})、二氧化硫(SO_2)、二氧化氮(NO_2)、臭氧(O_3)、一氧化碳(CO)这 6 个指标的基础上进行一系列的改进预测算法，却不曾想过对这 6 个指标进行精确化，以提升预测精度。本次采用梯度提升回归树(GBRT)进行变量选择来进一步提高预测精度，再结合遗传算法优化的最小二乘支持向量机模型进行预测，其预测误差显著降低。

2. 数据来源与研究方法

2.1. 数据来源

本次论文研究的数据来自于网络爬虫获取的重庆市 2020 年 1 月到 2020 年 8 月这 8 个月每天的空气质量指数(AQI)、空气中细颗粒物($PM_{2.5}$)、可吸入颗粒(PM_{10})、二氧化硫(SO_2)、一氧化碳(CO)、二氧化氮(CO_2)和臭氧(O_3)这 6 种主要的污染物浓度数据指标。

2.2. 研究方法

本次研究针对重庆市天时数据空气质量预测主要分以下三步进行：1) 先对 6 个影响空气质量主要污染物指标进行 GBRT 变量选择，筛选出更精确的主要污染物指标；2) 运用最小二乘支持向量机模型对变量选择前与变量选择后的数据进行预测，比较其预测精度；3) 对于最小二乘支持向量机模型运用遗传算法对参数进行优化，得到优化后的遗传算法最小二乘支持向量机模型，比较优化后的模型的精度，得到最优的组合预测模型。实际上就是一个串联式进化的预测过程，具体流程如下图 1。

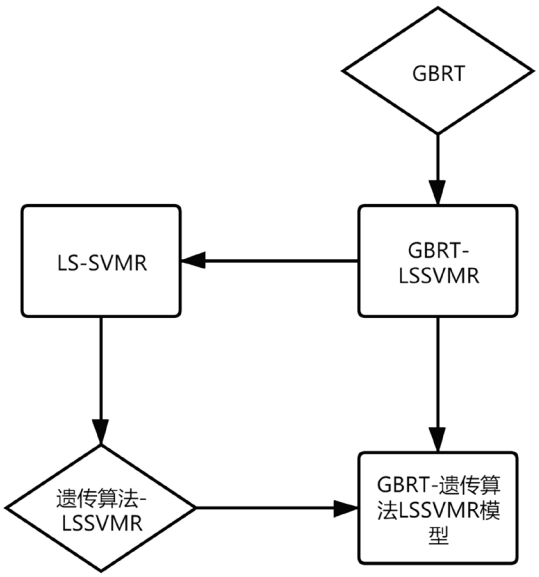

Figure 1. Modeling flow chart
图 1. 建模流程图

2.2.1. 梯度提升回归树(GBRT)

GBRT，Gradient Boosting Regression Tree，梯度提升回归树，由多棵树组成，所有树的结论累加得最后结果的迭代回归树算法，其泛化能力较强[11] [12]。具体算法如下：

输入：训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$。

输出：回归树 $f_M(x)$。

第一步：初始化

$$f_0(x)=\arg\min_c\sum_{i=1}^{N}L(y_i,c) \tag{1}$$

用来估计使损失函数最小化的常数值，这时它是只有一个根节点的树。

第二步：设 $m=1,2,\cdots,M$， $i=1,2,\cdots,N$，

$$r_{mi}=-\left[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)} \tag{2}$$

即计算损失函数的负梯度在当前模型的值，亦是残差的估计值。

由 r_{mi} 学习一颗回归树，得到第 m 棵树的叶节点区域 $R_{mj},j=1,2,\cdots,J$，以拟合残差的近似值，对每个 j 计算

$$c_{mj}=\arg\min_{c}\sum_{x_i\in R_{mj}}L\left(y_i,f_{m-1}\left(x_i\right)+c\right) \tag{3}$$

用线性搜索估计叶节点区域的值，使损失函数最小化，更新回归树

$$f_m\left(x\right)=f_{m-1}\left(x\right)+\sum_{x_i\in R_{mj}}c_{mj} \tag{4}$$

第三步：得到最终的回归树模型

$$\hat{f}\left(x\right)=f_M\left(x\right)=\sum_{m=1}^{M}\sum_{x_i\in R_{mj}}c_{mj} \tag{5}$$

为避免过拟合，引入一个参数 λ，$\lambda\subseteq\left[0.001,0.01\right]$，则可得到对应的回归树更新公式

$$f_m\left(x\right)=f_{m-1}\left(x\right)+\lambda\cdot\sum_{x_i\in R_{mj}}c_{mj} \tag{6}$$

2.2.2. 最小二乘支持向量机模型(LS-SVMR)

LS-SVMR，即最小二乘支持向量机模型，由 J. A. K. Suykens 提出。相较标准 SVM，它用等式约束代替不等式约束，解一组等式方程，避免了求解耗时的弊端，且求解速度相对加快[12] [13]。线性情况下的 LS-SVMR 的算法步骤如下：

输入：给定训练样本集 $\left\{\left(x_1,y_1\right),\left(x_2,y_2\right),\cdots,\left(x_n,y_n\right)\right\}$，$i=1,2,\cdots,n$。

输出：估计函数 $f\left(x_i\right)$。

第一步：首先定义线性回归方程：

$$f\left(x\right)=xw+b \tag{7}$$

其中，$x=\left(x_1,x_2,\cdots,x_n\right)'$，$I$ 为 n 维单位列向量。

第二步：找寻最优超平面，此时的优化目标函数为：

$$\begin{aligned}&\min_{w,b,\xi}\frac{1}{2}w'w+\frac{1}{2}\gamma\xi'\xi\\&\text{s.t. }y-xw-bI=\xi\end{aligned} \tag{8}$$

构造 Lagrange 函数：

$$L\left(w,b,\xi,\alpha\right)=\frac{1}{2}w'w+\frac{1}{2}\gamma\xi'\xi+\alpha'\left(y-xw-bI-\xi\right) \tag{9}$$

再对上面的 Lagrange 函数分别求 w,b,ξ,α 的偏导，并令其为 0，即可得到最小二乘支持向量机的估计函数：

$$f\left(x_i\right)=\sum_{j=1}^{n}\alpha_j\left(x_ix_j'\right)+b \tag{10}$$

同理针对非线性情况其估计函数为：

$$f\left(x_i\right)=\sum_{j=1}^{n}\alpha_jk\left(x_i,x_j\right)+b \tag{11}$$

2.2.3. 遗传算法

遗传算法是常见的随机优化搜索方法，它有 J. Holland 提出，遗传算法是模拟自然界遗传选择与生物进化计算的模型。遗传算法将多个个体集合解进行编码、选择、交叉、变异后，逐代进化，从子代找出求解问题的全局最优解[12]。遗传算法的步骤如下图 2。

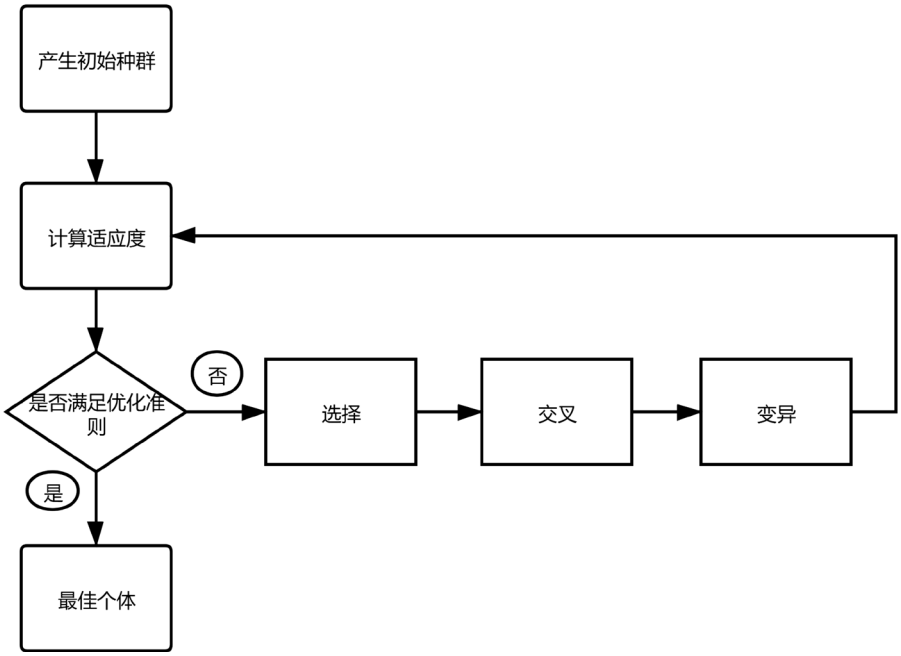

Figure 2. Genetic algorithm steps
图 2. 遗传算法步骤

3. 模型的建立

3.1. GBRT 变量选择

对于本次研究重庆市空气质量 AQI 指数，其有 6 个主影响因子，分别是空气中细颗粒物($PM_{2.5}$)、可吸入颗粒(PM_{10})、二氧化硫(SO_2)、一氧化碳(CO)、二氧化氮(NO_2)和臭氧(O_3)。这些因素之间可能存在相关性、冗余性等，为提升建模效率和更好的特征表示，对数据进行降维处理会对预测精度有显著提升。应用 GBRT 进行 6 个特征变量的筛选，由图示，变量在 6 个指标中的影响作用基本可忽略不计，且由图 4 也可以看出与 y 无显著线性相关性，故特征选择进行后面的建模，如下图 3 和图 4。

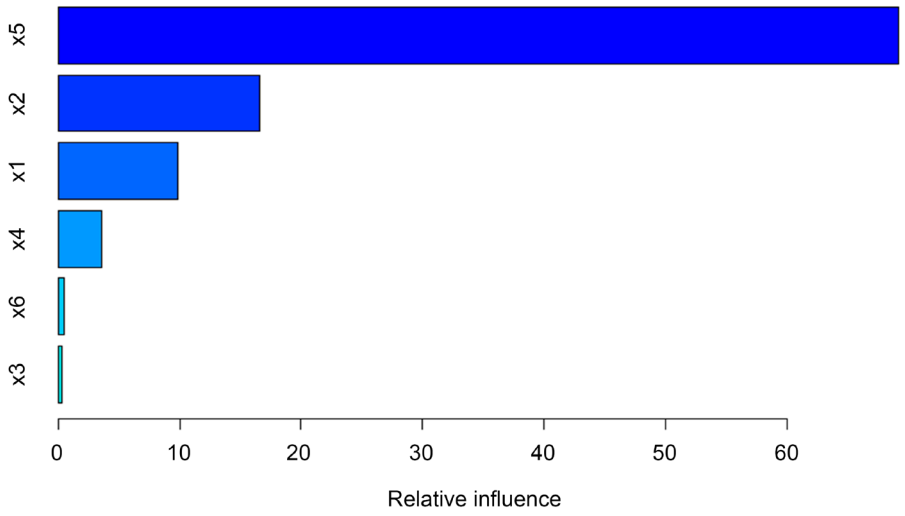

Figure 3. GBRT importance ranking
图 3. GBRT 重要性排序

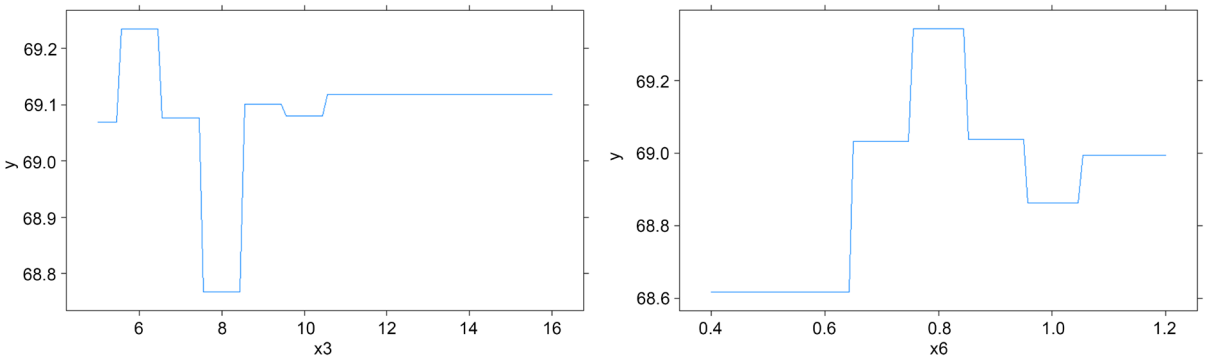

Figure 4. Marginal graph of x_3, x_6
图 4. x_3, x_6 边际图

3.2. 建立 GBRT-最小支持向量机模型

考虑到预测所涉及数据具有维数低和样本量少的特征。应用最小二乘支持向量机模型相对于支持向量机模型能够对数据预测有较好的准确性以及预测精度的提升，对比于神经网络又可以克服训练时间长，训练结果存在随机性以及学习的不足。在前文 GBRT 变量选择的基础上，再进行 LS-SVMR 预测，对比预测误差见下表 1，GBRT 改进后的 LS-SVMR 预测效果更好。

Table 1. Comparison of prediction errors between least square support vector machine and GBRT
表 1. 最小二乘支持向量机与 GBRT 组合预测误差对比

名称	最小二乘支持向量机	GBRT-最小二乘支持向量机
残差平方和/n	1.5329164	1.02862

3.3. 建立 GBRT-遗传算法优化最小二乘支持向量机模型

对于最小二乘支持向量机模型本身而言，其中的关键性参数一般是经验取值，不一定能得到最佳预测值。于是本小节采用了遗传算法优化参数，得到最适合重庆市空气质量数据的参数。遗传算法的目标就是找到合适参数使得在测试集上的误差 error 最小。遗产算法确定最优参数如下表 2。

Table 2. Optimization parameters of genetic algorithm
表 2. 遗传算法优化参数

名称	ξ	γ
GBRT 前	2.618618	199.884949
GBRT 后	1.294949	174.471560

最后将优化后的参数代入到之前的最小二乘支持向量机模型中去，可以看出在本身不用遗传算法优化的前提下，其误差也有显著的降低。最后在前面的基础上构造 GBRT-遗传算法优化 LS-SVMR 模型，可以看出其误差降低的更多，如下表 3。

Table 3. Comparison of model errors
表 3. 模型误差对比

名称	LS-SVMR	GBRT-LSSVMR	遗传算法 LS-SVMR	GBRT 遗传算法优化的 LS-SVMR
残差平方和/n	1.5329164	1.02860	0.8417785	0.1993641

由上表可以看出，由循环式串联进化算法，在基于 GBRT 变量选择的情况下，再在遗传算法优化后的最小二乘支持向量机里进行预测，其预测误差由最初的 1.5329164 降到了仅仅只有 0.1993641，其误差降低 87%，所以该 GBRT 遗传算法优化的最小支持二乘向量机模型是更适合重庆市空气质量指数 AQI 的预测，其预测精度有显著的提升。

3.4. 最终预测模型预测图

根据上文最后得到的 GBRT 结合遗传算法优化的最小二乘支持向量机模型，产生重庆市空气质量 AQI 值预测模型。

本文结合预测模型，对重庆市 2020 年 1 月到 2020 年 8 月每天的空气质量 AQI 进行预测，并结合实际的数据进行比对，结果如下图 5 所示蓝色(y)与红色(ypred)基本重合：

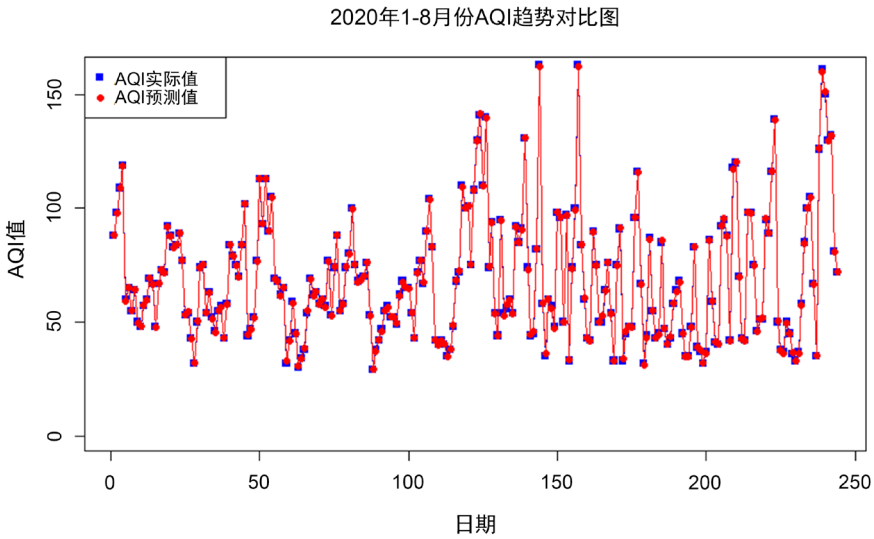

Figure 5. Comparison between true value and predicted value
图 5. 真实值与预测值对比

4. 结语

本文采用串联式循环进化算法模型，进行算法的改进提升预测的精度，减少预测的误差，得到了最终的 GBRT 遗传算法优化的最小二乘支持向量机模型对重庆市空气质量 AQI 进行预测，其预测误差显著降低，因此该模型应用于空气质量预测是更优的。

参考文献

[1] 张叶娥, 高云. 基于 ARIMA 模型的大同市空气质量预测研究[J]. 软件, 2019, 40(12): 85-89.

[2] 宋宇辰, 甄莎. BP 神经网络和时间序列模型在包头市空气质量预测中的应用[J]. 干旱区资源与环境, 2013, 27(7): 65-70.

[3] 王先行, 吴若怡, 邬雯雅, 张程博, 应婷婷, 周鑫隆. 基于 GM-RBF 组合模型的空气质量预测研究[J]. 宁波工程学院学报, 2018, 30(2): 46-52+96.

[4] 方彦. 基于灰色 RBF 神经网络的空气质量预测[J]. 中国科技信息, 2018(22): 100-102.

[5] 卢彬, 马行, 穆春阳, 张鄂. 基于 PCA-BN 的银川市空气质量预测[J]. 安全与环境工程, 2020, 27(5): 70-76.

[6] 刘君. 基于因子分析与径向基神经网络的空气质量预测研究[J]. 科技视界, 2020(21): 156-157.

[7] 胡邦辉, 刘丹军, 王学忠, 高传智. 最小二乘支持向量机在云量预报中的应用[J]. 气象科学, 2011, 31(2):

187-193.

[8] 付莲莲, 伍健. 基于梯度提升回归模型的生猪价格预测[J]. 计算机仿真, 2020, 37(1): 347-350.

[9] Agarwal, S., Sharma, S., Suresh, R., *et al.* (2020) Air Quality Forecasting Using Artificial Neural Networks with Real Time Dynamic Error Correction in Highly Polluted Regions. *Science of the Total Environment*, **735**, 139454. https://doi.org/10.1016/j.scitotenv.2020.139454

[10] 谷艳昌, 吴云星, 黄海兵, 庞琼. 基于遗传算法优化支持向量机的大坝安全性态预测模型[J]. 河海大学学报(自然科学版), 2020, 48(5): 419-425.

[11] 曲文龙, 陈笑屹, 李一漪, 汪慎文. 一种深度梯度提升回归预测模型[J]. 计算机应用与软件, 2020, 37(9): 194-201.

[12] 游皓麟. R 语言预测实战[M]. 北京: 电子工业出版社, 2016.

[13] Friedman. J.H. (2002) Stochastic Gradient Boosting. *Computational Statistics & Data Analysis*, **38**, 367. https://doi.org/10.1016/S0167-9473(01)00065-2