文章编号:1007-5321(2019)06-0162-08 **DOI**:10.13190/j.jbupt.2019-153

$PM_{2.5}$ 浓度预测与影响因素分析

彭 岩, 赵梓如, 吴婷娴, 王 洁
(首都师范大学 管理学院, 北京 100056)

摘要: 针对 $PM_{2.5}$ 浓度的非线性和不确定性,提出了一种基于集成树-梯度提升决策树(EnsembleTrees-GBDT)的 $PM_{2.5}$ 预测模型. 该模型首先在集成树框架下进行特征选择,即选取 $PM_{2.5}$ 浓度主要影响因素,使用算术均值聚合法计算出各项特征对 $PM_{2.5}$ 浓度增加的影响程度,并以影响程度由强到弱的次序排序;其次使用网格搜索对 GBDT 算法进行参数优化,选取树的深度等参数的最优值;最后构建完整的 $PM_{2.5}$ 浓度集成预测模型. 使用北京市 2015—2016 年的污染物浓度和气象条件观测值 2 个数据集,对模型进行了预测仿真实验. 对比实验结果表明,所提出的 EnsembleTrees-GBDT 预测模型相比于决策树、随机森林、支持向量机等模型,具有更低的平均绝对误差和均方根误差,同时具有更好的泛化能力,能够更准确地预测 $PM_{2.5}$ 浓度,并实现对 $PM_{2.5}$ 浓度影响因素的有效分析.

关 键 词: $PM_{2.5}$ 预测模型; 集成特征选择; 梯度提升决策树; 影响因素分析
中图分类号: TP391 **文献标志码**: A

Prediction of $PM_{2.5}$ Concentration Based on Ensemble Learning

PENG Yan, ZHAO Zi-ru, WU Ting-xian, WANG Jie
(School of Management, Capital Normal University, Beijing 100056, China)

Abstract: The increase of PM2.5 is a cause of haze. Effectively predicting PM2.5 concentration and analyzing its influence factors play an important role in air quality forecasting and controlling. Considering nonlinearity and uncertainty of PM2.5 concentration, a PM2.5 concentration prediction model which firstly selects features using integrated trees was presented based on ensemble trees-gradient boosting decision tree(GBDT). With standard arithmetic mean aggregation method, the article calculates the influence degree of each feature on the increment of PM2.5 concentration, and provides the impact ranking from strong to weak. The grid-search to select the optimal parameters of the GBDT algorithm was used , such as the depth of the tree. Two datasets, the pollutant concentration data and meteorological observation data of Beijing from 2015 to 2016, are used in the prediction model proposed. Compared with standard models such as decision tree, random forest and support vector machine, the ensemble trees-GBDT model is found to be lower mean absolute errors, lower root mean square errors and better generalization ability.
Key words: $PM_{2.5}$ prediction model; integrated feature selection; gradient boosting decision tree; analysis of influencing factors

收稿日期: 2019-07-22
基金项目: 全国教育科学规划项目-教育部重点课题(DLA190426)
作者简介: 彭 岩(1967—), 女, 教授.
通信作者: 王 洁(1978—), 女, 副教授, E-mail:wangjie@cnu.edu.can.

$PM_{2.5}$是指大气中直径小于或等于 2.5 μm 的颗粒物，也称为细颗粒物，目前已经成为大气首要污染物[1]. $PM_{2.5}$粒径小，扩散面积大，易附带有毒、有害物质，是霾发生的主要因素之一[2]. 目前，人们在$PM_{2.5}$浓度的预测上已经做了大量的工作，预测方面主要包括确定性模型和统计模型 2 种[3]. 确定性模型需要历史气象数据、化学初始条件和边界条件数据等信息来模拟污染物复杂的形成过程. 此类模型预测实现需要较长的系统运行时间，同时这些数据通常难以精确获取，影响模型精度[4]. 随着机器学习的发展，神经网络、随机森林(RF, random forest)等模型已成功应用于 $PM_{2.5}$浓度预测. 任才溶等[5]提出利用 K-Means 算法对原始气象数据聚类，然后利用欠采样方法对数据进行平衡采样，最后利用泛化能力好的 RF 构建预测模型. 黄婕等[6]通过 Stacking 集成策略对递归神经网络和卷积神经网络进行融合，结合了递归神经网络的时序记忆优势和卷积神经网络的特征表达能力，并充分利用时间轴上的关联信息对 $PM_{2.5}$的小时浓度进行预测. 在特征选择方面，张俐等[7]通过最大相关系数的方式改进快速过滤特征选择算法. 崔鸿雁等[8]归纳了用于特征选择的相关性度量方法、稀疏选择方集成方法、神经网络方法和主成分分析方法，为研究者提供参考.

单一模型通常具有一定的局限性，例如，支持向量机(SVM, support vector machine)模型对缺失数据敏感；决策树(DT, decision tree)模型的结果不稳定，在数据中一个很小的变化可能导致生成一个完全不同的树；人工神经网络方法需要调整的参数过多等. 为了解决这个问题，笔者提出了一种新的集成学习模型——基于集成树-梯度提升决策树(EnsembleTrees-GBDT, EnsembleTrees-gradient boosting decision tree)模型进行 $PM_{2.5}$的浓度预测与影响因素分析：一方面使得特征选择相关性更高、更准确；另一方面降低了模型构建中的复杂性和模型过拟合的风险. 模型集成 Bagging 的 DT、RF 和极端随机树(ET, extra tree)3 种算法，通过增加单个树的差异性以提高泛化性能和预测精度. 由于梯度提升决策树(GBDT, gradient boosting decision tree)算法既可以处理离散值，又可以处理连续值，并且能够很好地适应异常值的鲁棒性，使得该算法适合用于解决$PM_{2.5}$干扰项多和异常情况多的情况. 同时，通过网格搜索(GS, grid search)对 GBDT 的基学习器个数、树深度等参数进行优化，提高了模型的运行效率.

1　相关研究

集成学习[8]是将多个弱学习器按照一定的规则结合起来，得出性能表现优于单个弱学习器的模型. 通过多年的发展，出现了很多新的思想和模型，并对多种集成学习模型进行融合. 在 Bagging、Boosting、AdaBoost 的基础上，发展出了 RF、GBDT 等算法.

1.1　RF 算法

RF 是基于 Bagging 算法的集成模型，一般采用分类与回归树(CART, classification and regression tree)作为基础模型，它包含多棵随机产生的 DT[9]. 由于各 DT 构建过程的随机性，RF 被证明不会过拟合[10]，故每棵树都尽可能地生长而不需要剪枝.

1.2　ET 算法

Geurts 等[11]提出了 ET 方法，根据经典的自上而下的方法，ET 构建了一系列“自由生长”的回归树集合. ET 方法是完全随机地得到分叉值，从而进行对回归树的分叉[12].

1.3　GBDT 算法

GBDT 属于 Boosting 算法，结合了回归树与提升树的思想. GBDT 与 RF 算法类似，但属于不同的 DT 模型组合方式. GBDT 输出为每棵 DT 输出结果的累加，利用梯度提升和回归 DT 的组合方式，每次建立新的 DT 模型都是在之前模型损失函数梯度的下降方向，使得决策模型不断改进[13].

GBDT 首先使用最速下降的近似方法来计算残差的近似值，即

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

GBDT 的算法如下.

输入：训练数据集 $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$，当损失函数为均方差时，有

$$L(y, f(x)) = (y - f(x))^2 \tag{2}$$

步骤 1　初始化：

$$f_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i, \gamma) \tag{3}$$

步骤 2　对 $m = 1,2,\cdots,M$ 进行迭代.

步骤 3　对于每一个样本(x_i, y_i)，计算残差：

$$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial, f(x_i)}\right] \tag{4}$$

步骤 4　利用$\{(x_i, R_{mi}), i = 1,2,\cdots,N\}$拟合一

棵 CART 回归树,得出第 m 棵回归树 T_m,其叶节点划分的区域为 $R_{mj}, j=1,2,\cdots,J$.

步骤 5 对于回归树 T_m 的每一个叶节点,计算其输出值:

$$c_{mj}=\underset{c}{\operatorname{argmin}}\sum_{x_i\in R_{mj}}L(y_i,f_{m-1}(x_i)+c) \tag{5}$$

步骤 6 更新回归树:

$$f_m(x)=f_{m-1}(x)+\sum_{j=1}^{J}c_{mj}I(x\in R_{mj}) \tag{6}$$

步骤 7 得到最终提升回归树:

$$\hat{f}(x)=f_M(x)=\sum_{m=1}^{M}\sum_{j=1}^{J}c_{mj}I(x\in R_{mj}) \tag{7}$$

输出:梯度提升树 $\hat{f}(x)$.

2 EnsembleTrees-GBDT 预测模型构建

相对于 DT,RF、GBDT 既具有继承了 DT 的可解释性强,又能够很好地处理特征间的相关关系等优点,提高了泛化能力. 笔者使用了 DT、RF、ET 等异质学习器以及 GBDT 算法等同质学习器.

所提出的 $PM_{2.5}$ 预测模型框架如图 1 所示. 在进行原始数据清洗后,使用 Z-score 标准化处理解决数据集中各变量单位不同、大小差异大等问题,以确保数据的可靠性与可用性. 模型在 EnsembleTrees 框架下进行特征选择,采取算术均值聚合得出最终特征影响程度强弱排序,将其作为模型的输入变量;使用 GS 对 GBDT 算法的基学习器个数、GBDT 树深度、特征个数进行优化,选出最优参数构建预测模型.

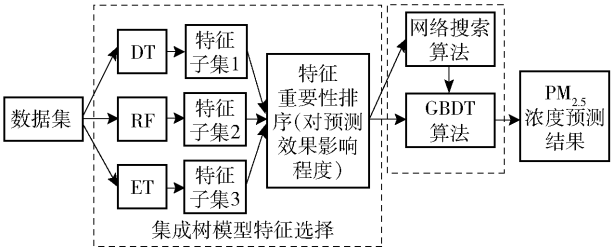

图 1 基于 EnsembleTrees-GBDT 的 $PM_{2.5}$ 预测模型框架

2.1 数据预处理

2.1.1 数据来源

实验数据来源于国家人口与健康科学数据共享服务平台(http://www.ncmi.cn/):一部分是环保部发布的 2015-01-01—2016-12-31 全国省会城市空气污染指数资料,从中提取出了北京市可吸入颗粒物(PM_{10})、二氧化硫(SO_2)、二氧化氮(NO_2)、臭氧(O_3)4 种污染物数据;另一部分是气象条件数据,其来源于 2015-01-01—2016-12-31 全国 700 个气象站的地面气象资料,从中提取出了北京市的 15 项气象条件参数,如表 1 所示. 其中 4 种污染物浓度与气象条件数据构成了 19 维的特征数据,按日作为研究对象,即构成一个 19×732 的数据矩阵,表示为

$$\boldsymbol{X}=[x_{11},x_{12},\cdots,x_{ij}],i=1,2,\cdots,731,\quad j=1,2,\cdots,19 \tag{8}$$

同时将 $PM_{2.5}$ 的单日浓度作为预测目标,表示为

$$\boldsymbol{Y}=[y_1,y_2,\cdots,y_i],i=1,2,\cdots,731 \tag{9}$$

表 1 气象参数列表

变量名	注释	单位
X_1	PM_{10}	μg/m³
X_2	SO_2	μg/m³
X_3	NO_2	μg/m³
X_4	O_3	μg/m³
X_5	平均气压	0.1 hPa
X_6	最高气压	0.1 hPa
X_7	最低气压	0.1 hPa
X_8	平均气温	0.1 ℃
X_9	最高气温	0.1 ℃
X_{10}	最低气温	0.1 ℃
X_{11}	相对湿度	1%
X_{12}	最小相对湿度	1%
X_{13}	24 h 降水量	0.1 mm
X_{14}	平均风速	0.1 m/s
X_{15}	最大风速	0.1 m/s
X_{16}	最大风速风向	-
X_{17}	极大风速	0.1 m/s
X_{18}	极大风速风向	-
X_{19}	日照时数	0.1 h

2.1.2 数据标准化

变量采用不同单位来衡量,在用于建模之前,需进行无量纲标准化处理. 根据气象因素等变量接近正态分布,选取了 Z-Score 算法进行标准化处理. 将某个具体的观测值表示为 x_i,μ 为观测值所在数据组的均值,σ 为该组数据的标准差(见式(10)),对原始数据进行标准化,以消除不同数据水平和变量单位对结果的影响.

$$Z(x_i)=\frac{x_i-\mu}{\sigma} \tag{10}$$

2.2 基于 EnsembleTrees 的特征选择

特征选择的目的是对所有特征进行影响程度强弱排序,即选择出对预测 $PM_{2.5}$ 浓度影响程度较大的

特征. 分析特征与自变量之间的相关性,消除冗余特征和不相关特征. 笔者分别使用 DT、RF、ET 得出特征强弱排序,再通过算术均值聚合得出最终特征强弱排序. 差异性是集成学习性能优于其他算法的前提,对 DT、RF、ET 这 3 种特征选择方法的结果采用算术均值聚合,既可以保证学习器的差异性,也可以降低模型的整体误差.

特征选择的详细过程描述如下:

步骤 1　计算基于 DT 的特征排序.

对于 DT 来说,假设最初总共有 K 类,样本属于第 k 类的概率为 p_k,则该概率分布的 Gini 值为

$$\mathrm{Gini}(p)=\sum_{k=1}^{K}p_k(1-p_k)=1-\sum_{k=1}^{K}p_k^2 \quad (11)$$

Gini 的选择标准为:每个子节点都达到最高的纯度,此时 Gini 值最小,纯度最高,不确定度最小.

用平方误差最小化原则寻找树的划分点,选择最优切分点,求解

$$\min_{(j,s)}\left[\min_{c1}\sum_{x_1\in D_1(j,s)}(y_i-c_1)^2+\min_{c2}\sum_{x_1\in D_2(j,s)}(y_i-c_2)^2\right] \quad (12)$$

遍历每个特征的每个分割点时,使用特征 $A=a$ 将 D 划分为 D_1和 D_2两部分,分别表示满足 $A=a$ 的样本集合和不满足该条件的样本集合. 式(12)中,c_1、c_2是 D_1、D_2的样本均值,j 为当前的样本特征,s 为划分点.

确定划分点后,计算其 Gini,寻找 Gini 系数最小特征的分割点,使得划分前和划分后的 Gini 系数差值最大. 差值越大,则说明当前的特征对浓度的影响越大.

遍历所有特征,得出所有特征对浓度影响的强弱程度,并对其按强弱程度降序排序,最重要为 1,最不重要为 19,放入集合 F 中,记为 F_{DT}.

步骤 2　计算基于 RF 的特征排序.

对于 RF 的 CART 而言,Gini 表示为

$$\mathrm{Gini}(p)=2p(1-p) \quad (13)$$

在特征 $A=a$ 的条件下,D 的 Gini 指数为

$$\mathrm{Gini}(D,A)=\frac{[D_1]}{|D|}\mathrm{Gini}(D_1)+\frac{[D_2]}{|D|}\mathrm{Gini}(D_2) \quad (14)$$

其中:集合 D 的不确定性用 $\mathrm{Gini}(D)$表示,$\mathrm{Gini}(A,D)$则表示在经过 $A=a$ 分割后,集合 D 的不确定性.

RF 中的每棵 CART 需要通过不断遍历这棵树的特征子集所有可能的分割点来寻找 Gini 系数最小特征的分割点,使得划分后的结果很纯,划分前和划分后的 Gini 系数差值最大. 与 DT 的差异在于它不是在分割节点时搜索最重要的特征,而是在随机特征子集中搜索最佳特征.

遍历所有特征,得出所有特征对浓度的影响程度,并对其按影响程度大小降序排序,放入集合 F 中,记为 F_{RF}.

步骤 3　计算基于 ET 的特征排序.

对于 ET 而言,Gini 指数的计算方式与 RF 一致. 但在寻找最佳分割点时,需先将节点的样本分成两组,对每个随机选择的特征进行随机的分割,然后选择出最佳分割.

遍历所有特征,得出所有特征对浓度的影响程度大小,并对其按影响程度大小降序排序,放入集合 F 中,记为 F_{ET}.

步骤 4　计算基于 DT + RF + ET 的 Ensemble-Trees 模型的特征影响强度值 F_{T}.

按照算术均值聚合法,计算特征 X_i的平均影响程度,F_{T}值越小,影响越强:

$$F_{\mathrm{T}}=(F_{\mathrm{DT}}+F_{\mathrm{RT}}+F_{\mathrm{ET}})/3 \quad (15)$$

根据 F_{T}值进行升序排序. 各算法特征排序结果如表 2 所示.

表 2　各算法特征排序结果

特征	DT 特征排序	RF 特征排序	ET 特征排序	所提算法特征排序
PM_{10}	1	1	1	1
NO_2	2	2	2	2
RH_ave	3	9	4	3
SO_2	8	11	3	4
sun	5	13	5	5
O_3	9	7	8	6
wind_max	7	14	7	7
temp_min	15	5	11	8
wind_ex	6	12	13	9
pre_min	16	4	16	10
wind_direction_ex	14	8	14	11
temp_max	12	16	9	12
pre_max	11	10	17	13
RH_min	4	3	6	14
temp_ave	17	6	18	15
pre_ave	19	17	10	16
wind_direction_max	13	18	15	17
precipitation	18	19	19	18
wind_ave	10	15	12	19

2.3 $PM_{2.5}$预测模型构建

由于GBDT算法能够很好地适应异常值的鲁棒性,并且GBDT可以同时处理离散值和连续值,使得GBDT算法适应于$PM_{2.5}$数据非线性和干扰项多的情况,因此使用GBDT算法构建$PM_{2.5}$浓度预测模型.

通过train_test_split函数对经过特征选择得到的特征子集F_m按照训练集70%、测试集30%的比例将其划分为训练集F_{train}和测试集T_{test}.

GBDT模型在Boosting迭代框架下进行M次迭代,如式(16)所示. 其中,$i=0$时,$f(0)$为初始预测值;$i=1,2,\cdots,M$时,$f_i(x)$为第i次迭代的函数增量. 初始预测值和各函数增量之和即为预测值:

$$f(x)=\sum_{i=0}^{M} f_i(x) \tag{16}$$

2.3.1 GS算法

GS算法将待搜索参数在一定的空间范围内划分成网格,通过遍历网格中所有的点来寻找最优参数. 这种方法在寻优区间足够大且步距足够小的情况下可以找出全局最优解,但须对所有参数进行排列组合,缺点是模型搜索的时间复杂度高[14].

2.3.2 GBDT模型调参

根据GS算法的思想,首先需要设置将要选择的参数组合区间. 基于GBDT算法结合GS算法,进行参数优化,不断地对模型进行训练,通过评价函数对每个参数组合得到的结果进行评价,最终得到最优参数组合. 该方法能够克服交叉验证的缺点,最后将最优参数组合代入GBDT算法,从而使预测性能得到提升.

笔者利用Python平台建立模型,GBDT模型需要设置基学习器个数(n-estimators)、最大特征数(max-features)以及树的深度(max-depth)3个重要参数. 为防止过拟合,学习率设为0.03,对学习率设置较小的值对模型可达到正则化的效果. 本研究选用的是huber损失函数来计算预测结果的残差,huber损失函数对异常值的鲁棒性非常强. 基于此,对基学习器个数、分类回归树的深度以及最大特征数进行GS,最大深度设置为3~10,最大迭代次数设为50~500.

由GS结果得出,最大深度为6,最大迭代次数为200,最大特征数为9. 通过集成特征选择方法选取了PM_{10}、SO_2、NO_2、O_3、temp_min、RH_ave、wind_max、wind_ex、sun共9项影响程度高的特征.

2.3.3 部分依赖图

部分依赖(PD, partial dependence)图显示了1个或2个特征对机器学习模型的预测结果的边际效应[15]. PD图可以显示目标与特征之间的关系是线性还是非线性的,是单调还是复杂的. 例如,应用于线性回归模型时,PD图总是显示线性关系. 回归的PD函数定义为

$$f_{x_s}(x_s)=E_{x_c}[f(x_s,x_c)] \tag{17}$$

在模型f中,x_s是PD图中绘制的特征,x_c是学习模型中的其他特征. 通常,集合S中只有1个或2个特征. S中的特征是对预测结果有影响的特征. 特征向量$\boldsymbol{x}_s$和$\boldsymbol{x}_c$组合构成总特征空间X. PD性通过边缘化集合C中的特征分布在学习模型的输出而起作用,因此该函数显示目标集合S中的特征与预测结果之间的关系. 通过边缘化其他特征,得到的函数仅依赖于S中的特征,包括该特征与其他特征的交互.

PD图展示了特征S的给定值对预测的平均边际效应. $PM_{2.5}$浓度影响因素的PD图如图2、图3所示.

图2(a)、(b)、(c)、(d)分别为PM_{10}浓度、NO_2浓度、SO_2浓度以及wind_direction_ex4个重要变量的PD图,图2(e)为控制其他变量后的PM_{10}与NO_2的PD图,可以看出,不同变量之间存在差异效应. 由于这些变量的连续性,PM_{10}的PD函数近似阶梯函数,其他变量的PD函数近似线性函数. 这4个变量的增大分别使$PM_{2.5}$浓度有不同程度的上升. 同时,PM_{10}浓度也随NO_2浓度增大而上升.

图3显示了因变量与2个最重要变量(PM_{10}浓度与NO_2浓度)的联合PD图. 因为它们在单个图中显示对2个变量的联合PD性,所以此图以三维呈现,显示了所涉及的变量之间不同的相互作用:PM_{10}浓度一定,NO_2依赖度高于0.5时,$PM_{2.5}$浓度有明显上升;NO_2浓度一定时,$PM_{2.5}$浓度随PM_{10}浓度上升而上升.

3 对比实验

为了更好地验证所提出的集成模型Ensemble-Trees-GBDT的有效性,设置了3组对比实验:

1) 所采用的特征选择方法与未使用该方法的结果对比;

2) 所采用的GS优化的GBDT与使用默认参数的GBDT的对比;

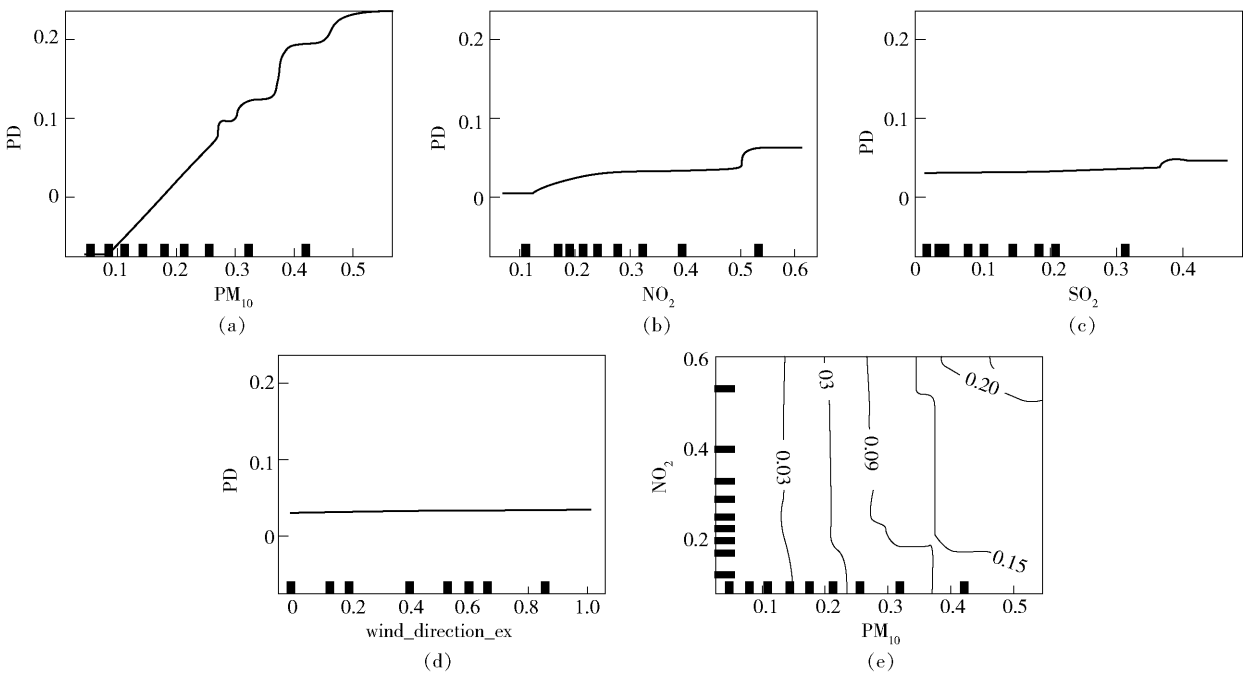

图 2　PM_{10}、NO_2、SO_2、wind_direction_ex PD 图

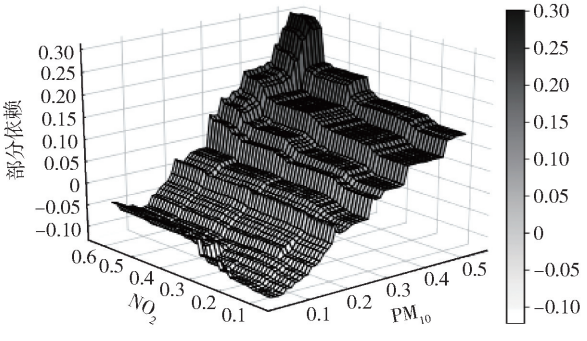

图 3　PM_{10}与 NO_2联合 PD 图

3）所采用的联合模型 EnsembleTrees-GBDT 与使用传统 DT、RF 以及 SVM 等单一算法对比. 对预测结果通过引用平均绝对误差(MAE, mean absolute errors)和均方根误差(RMSE, root mean square errors)来进行评估.

3.1　EnsembleTrees 模型特征选择方法与未使用该方法的对比分析

比较结果如表 3 所示. 可以看出,使用 EnsembleTrees 模型特征选择算法与未使用该方法相比,RF、DT、SVM 和 GS-GBDT 的预测精度均有提升,更加接近真实值.

3.2　模型参数优化对比分析

为了验证 GS 优化模型的有效性,将经过 GS 优化的GBDT与使用默认参数设置的GBDT进行对比,预测结果如图 4 所示. 由图 4 可以看出,经过 GS 优化的预测结果更加接近真实值.

表 3　使用与未使用 EnsembleTrees 模型特征选择结果对比

预测算法	MAE（Ensemble Trees 特征选择算法）	RMSE（Ensemble Trees 特征选择算法）	MAE（未进行 Ensemble Trees 特征选择）	RMSE（未进行 Ensemble Trees 特征选择）
DT	0.037 0	0.060 6	0.043 2	0.071 4
RF	0.032 9	0.051 5	0.039 7	0.063 2
SVM	0.042 7	0.057 9	0.049 2	0.064 7
所提 GBDT 算法	0.016 4	0.033 4	0.020 2	0.037 4

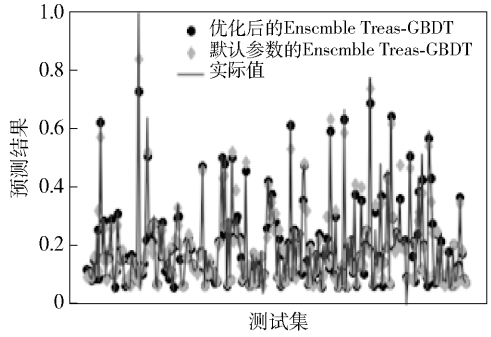

图 4　模型预测结果对比

表 4 所示为经过参数优化的 GBDT 与使用默认参数 GBDT 预测结果的 MAE 和 RMSE 对比. 由表 4 可以看出,参数优化前后的 MAE 和 RMSE 都很低,

说明预测结果误差较小. 相对于直接使用 EnsembleTrees-GBDT,GS 优化之后,MAE、RMSE 分别由 0.031 1、0.044 8 降到了 0.016 4、0.033 4,模型精度更高. 结果表明,GS 优化后的 EnsembleTrees-GBDT 模型能够有效预测 $PM_{2.5}$的浓度.

表 4 模型结果对比

预测算法	MAE	RMSE
默认参数的 EnsembleTrees-GBDT	0.031 1	0.044 8
优化后的 EnsembleTrees-GBDT	0.016 4	0.033 4

3.3 GS 优化的 EnsembleTrees-GBDT 与 DT 的对比分析

为了更好地验证所提出 GS 优化的集成模型 EnsembleTrees-GBDT 的有效性,将其与 DT 进行对比. 预测结果如图 5 所示,可以看出,相比于 DT,GS 优化的 EnsembleTrees-GBDT 的预测结果更加靠近 $PM_{2.5}$浓度的真实值.

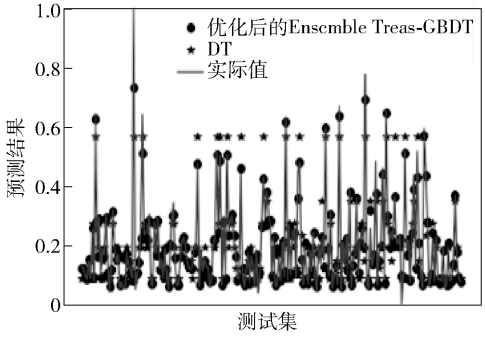

图 5 优化后的 EnsembleTrees-GBDT 与 DT 预测结果对比

由表 5 可以看出,相对于使用 DT,GS 优化的 EnsembleTrees-GBDT 模型的 MAE、RMSE 更低,误差更小,分别由 0.037 0、0.060 6 降到了 0.016 4、0.033 4. 结果表明,优化后的 EnsembleTrees-GBDT 模型预测精度更高.

表 5 模型结果对比

预测算法	MAE	RMSE
DT	0.037 0	0.060 6
优化后的 EnsembleTrees-GBDT	0.016 4	0.033 4

4 结束语

目前,集成学习的理论和算法研究是机器学习领域的一个热点,越来越多的集成学习方法被广泛应用于预测中. 与使用单一的机器学习算法进行 $PM_{2.5}$预测的方法相比,笔者使用的 EnsembleTrees-GBDT 集成模型能够有效地分析出各影响因素对 $PM_{2.5}$浓度影响的大小,同时使用 GS 对基学习器个数、GBDT 树深度、特征个数进行优化,能够有效提高模型性能与预测精度. 对比实验表明,相比于单一模型,所提出的 EnsembleTrees-GBDT 集成模型预测误差降低、精度提升. 综合实验结果表明,所提出的 EnsembleTrees-GBDT 集成模型可以从影响因素分析和浓度预测两方面对 $PM_{2.5}$污染进行分析和预测,研究工作可以为北京市 $PM_{2.5}$污染物的防治工作提供科学的决策依据.

参考文献:

[1] 张青, 饶灿. 典型区域城市 $PM_{2.5}$与 PM_{10}比值相关性研究[J]. 绿色科技, 2019(12): 129-130.
Zhang Qing, Rao Can. Correlation analysis between $PM_{2.5}$ and PM_{10} ratio in typical regional cities[J]. Journal of Green Science and Technology, 2019(12): 129-130.

[2] 刘晓红, 王慧. 基于中欧对比视角的货运机动车尾气排放 $PM_{2.5}$分析研究[J]. 环境科学学报, 2019, 39(8): 2830-2838.
Liu Xiaohong, Wang Hui. An analysis of vehicle-related $PM_{2.5}$ emissions: the perspective from China and Europe[J]. Acta Scientiae Circumstantiae, 2019, 39(8): 2830-2838.

[3] 李建新, 刘小生, 刘静, 等. 基于 MRMR-HK-SVM 模型的 $PM_{2.5}$浓度预测[J]. 中国环境科学, 2019, 39(6): 2304-2310.
Li Jianxin, Liu Xiaosheng, Liu Jing, et al. Prediction of $PM_{2.5}$ concentration based on MRMR-HK-SVM model[J]. China Environmental Science, 2019, 39(6): 2304-2310.

[4] 王平, 张红, 秦作栋, 等. 基于 wavelet-SVM 的 PM_{10}浓度时序数据预测[J]. 环境科学, 2017, 38(8): 3153-3161.
Wang Ping, Zhang Hong, Qin Zuodong, et al. PM_{10} concentration forecasting model based on wavelet-SVM[J]. Environmental Science, 2017, 38(8): 3153-3161.

[5] 任才溶, 谢刚. 基于随机森林和气象参数的 $PM_{2.5}$浓度等级预测[J]. 计算机工程与应用, 2019, 55(2): 213-220.
Ren Cairong, Xie Gang. Prediction of $PM_{2.5}$ concentration level based on random forest and meteorological parameters[J]. Computer Engineering and Applications, 2019, 55(2): 213-220.

[6] 黄婕，张丰，杜震洪，等. 基于 RNN-CNN 集成深度学习模型的 $PM_{2.5}$小时浓度预测[J]. 浙江大学学报(理学版)，2019，46(3)：370-379.
Huang Jie，Zhang Feng，Du Zhenhong，et al. Hourly concentration prediction of $PM_{2.5}$ based on RNN-CNN ensemble deep learning model[J]. Journal of Zhejiang University(Science Edition)，2019，46(3)：370-379.

[7] 张俐，袁玉宇，王枞. 基于最大相关信息系数的 FCBF 特征选择算法[J]. 北京邮电大学学报，2018，41(4)：86-90.
Zhang Li，Yuan Yuyu，Wang Cong. FCBF feature selection algorithm based on maximum information coefficient [J]. Journal of Beijing University of Posts and Telecommunications，2018，41(4)：86-90.

[8] 崔鸿雁，徐帅，张利锋，等. 机器学习中的特征选择方法研究及展望[J]. 北京邮电大学学报，2018，41(1)：1-12.
Cui Hongyan，Xu Shuai，Zhang Lifeng，et al. The key-techniques and future vision of feature selection in machine learning[J]. Journal of Beijing University of Posts and Telecommunications，2018，41(1)：1-12.

[9] Dietterich T G. Machine learning research：four current directions[J]. AI Magazine，1997，18(4)：97-136.

[10] 刘云翔，陈斌，周子宜. 一种基于随机森林的改进特征筛选算法[J]. 现代电子技术，2019，42(12)：117-121.
Liu Yunxiang，Chen Bin，Zhou Ziyi. An improved feature selection algorithm based on random forest[J]. Modern Electronics Technique，2019，42(12)：117-121.

[11] Geurts P，Ernst D，Wehenkel L. Extremely randomized trees[J]. Machine Learning，2006，63(1)：3-42.

[12] 黄丛吾，陈报章，马超群，等. 基于极端随机树方法的 WRF-CMAQ-MOS 模型研究[J]. 气象学报，2018，76(5)：779-789.
Huang Congwu，Chen Baozhang，Ma Chaoqun，et al. WRF-CMAQ-MOS studies based on extremely randomized trees[J]. Acta Meteorologica Sinica，2018，76(5)：779-789.

[13] 刘金硕，刘必为，张密，等. 基于 GBDT 的电力计量设备故障预测[J]. 计算机科学，2019，46(S1)：392-396.
Liu Jinshuo，Liu Biwei，Zhang Mi，et al. Fault prediction of power metering equipment based on GBDT[J]. Computer Science，2019，46(S1)：392-396.

[14] 雷雪梅，谢依彤. 用于高血压菜谱识别的基于遗传算法的改进 XGBoost 模型[J]. 计算机科学，2018，45(增刊 1)：476-481.
Lei Xuemei，Xie Yitong. Improved XGBoost model based on genetic algorithm for hypertension recipe recognition[J]. Computer Science，2018，45(S1)：476-481.

[15] Friedman J H. Greedy function approximation：a gradient boosting machine[J]. The Annals of Statistics，2001，29(5)：1189-1232.