研究与开发

基于随机森林回归分析的 PM2.5 浓度预测模型

杜续[1]，冯景瑜[1,2]，吕少卿[1,2]，石薇[1]

（1. 西安邮电大学通信与信息工程学院，陕西 西安 710121；
2. 西安邮电大学陕西省信息通信网络及安全重点实验室，陕西 西安 710121）

摘　要：针对神经网络算法在当前 PM2.5 浓度预测领域存在的易过拟合、网络结构复杂、学习效率低等问题，引入 RFR（random forest regression，随机森林回归）算法，分析气象条件、大气污染物浓度和季节所包含的 22 项特征因素，通过调整参数的最优组合，设计出一种新的 PM2.5 浓度预测模型——RFRP 模型。同时，收集了西安市 2013—2016 年的历史气象数据，进行模型的有效性实验分析。实验结果表明，RFRP 模型不仅能有效预测 PM2.5 浓度，还能在不影响预测精度的同时，较好地提升模型的运行效率，其平均运行时间为 0.281 s，约为 BP-NN（back propagation neural network，BP 神经网络）预测模型的 5.88%。

关键词：PM2.5 浓度预测；随机森林回归分析；BP 神经网络

中图分类号：TP391　　**文献标识码：**A

doi: 10.11959/j.issn.1000−0801.2017211

PM2.5 concentration prediction model based on random forest regression analysis

DU Xu[1], FENG Jingyu[1,2], LV Shaoqing[1,2], SHI Wei[1]

1. Institute of Communication Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, China
2. Shaanxi Key Laboratory of Information Communication Network and Security, Xi'an University of Posts and Telecommunications, Xi'an 710121, China

Abstract: The random foreat regression algorithm was introduced to solve the shortcomings of neural network in predicting the PM2.5 concentration, such as over-fitting, complex network structure, low learning efficiency. A novel PM2.5 concentration prediction model named RFRP was designed by analyzing the 22 characteristic factors including the meteorological conditions, the concentration of air pollutants and the season. The historical meteorological data of Xi'an in 2013—2016 were collected to verify the effectiveness of the model. The experimental results show that the proposed model can not only predict the PM2.5 concentration effectively, but also improve the operating efficiency of the model without affecting the prediction accuracy. The average run time of the proposed model is 0.281 s, which is about 5.58% of the neural network prediction model.

Key words: PM2.5 concentration prediction, random forest regression analysis, back propagation neural network

收稿日期：2017−03−30；**修回日期：**2017−06−29

基金项目：国家自然科学基金资助项目（No.61301091）；陕西省工业公关计划基金资助项目（No.2016GY-113）；陕西省教育厅专项科研计划基金资助项目（No.15JK1671）；西安邮电大学“西邮新星”团队支持计划基金资助项目（No.2015-01）

Foundation Items: The National Natural Science Foundation of China (No.61301091), The Industrial Science and Technology Project of Shaanxi Province (No.2016GY-113), The Natural Science Foundation of Education Department of Shaanxi Province (No.15JK1671), The New Star Team of Xi'an University of Posts & Telecommunications (No.2015-01)

1　引言

2013 年至今，以 PM2.5[1]（可入肺颗粒物）为首要污染物的雾霾问题在我国愈发严峻，不仅严重威胁着人们的日常生活与健康以及地球的气候系统，还造成了我国经济利益的巨大损失。美国环境流行病学教授 Schwart 研究发现 PM2.5 的日平均浓度每增加 10 μg/m^3，当日心肺疾病的病死率会提高 1.5%[2]。中国社会科学院的相关专家研究证明，随着 GDP 的不断增加，PM2.5 浓度持续呈上升状态；能源消耗结构中煤炭消耗所占比每升高 1%，PM2.5 浓度会增加 0.064%或 0.052%[3]。PM2.5 已经成为我国的环境公害。因此，科学、准确地预测 PM2.5 浓度，提前做好预防措施，对于降低 PM2.5 对国家和人民造成的危害有着十分重要的现实意义。

针对 PM2.5 浓度预测问题，国内外学者做了大量研究工作，提出了一系列模型。在预测方法方面，主要有线性回归[4]、时间序列[5]、灰色模型[6]、支持向量机[7]、贝叶斯[8]等传统方法以及近期以神经网络（neural network，NN）算法[9,10]为主导的人工智能方法。传统方法具有结构简单易识别、模型解释能力强等特点，但 PM2.5 的形成属于一个复杂的物理变化过程，具有明显的时空分异和非线性特征，因此传统方法很难反映实际情况。目前，神经网络算法具有较强的非线性和自我学习能力，已广泛应用于空气污染预测领域，主要是通过建立单个预测模型来预测 PM2.5 浓度。比如，参考文献[11]采用模糊神经网络，将外部气象条件作为特征，具有较好的实时性，但未考虑其他大气污染物对 PM2.5 浓度的影响；参考文献[12]采用模糊神经网络，只考虑了大气污染物浓度而忽略了气象条件；参考文献[13]对气象条件和大气污染物浓度进行逐步回归，预测结果比较合理，但回归模型较为简单，导致精确度不高。但是，神经网络算法对于高维特征的预测问题具有收敛速度慢、易造成过拟合的问题，且需要大量调整参数以确定网络结构。

鉴于此，本文利用随机森林回归（random forest regression，RFR）算法可以同时处理连续、离散属性，运行效率高，具有较强的顽健性、抗噪声、防止过拟合等优点，从预测方法和特征选取两个方面寻求创新，采用风力、降水、温度等气象条件，CO、NO、SO_2 等大气污染物浓度和季节以及前一天的 PM2.5 浓度等 22 项特征作为输入变量，结合集成思想，设计出一种新的 PM2.5 浓度预测模型——RFRP 模型。

2　RFR 算法

RFR 算法是由 Leo Breiman 和 Adele Cutler[14]于 2001 年共同提出的一种基于决策树的集成学习算法。RFR 算法是机器学习算法中精确度较高的算法，能克服单个预测或分类模型的缺点，被广泛应用于经济、医学、生物等领域，但在环境领域应用较少。

2.1　原理

定义 1[14]　随机森林回归算法是由一组回归决策子树 $\{h(x,\theta_t),t=1,2,\cdots,T\}$ 构成的组合模型。其中 θ_t 是服从独立同分布的随机变量，x 表示自变量，T 表示决策树的个数。利用集成学习的思想取各决策子树 $\{h(X,\theta_t)\}$ 的均值作为回归预测结果：

$$\bar{h}(x)=\frac{1}{T}\sum_{t=1}^{T}\{h(x,\theta_t)\} \quad (1)$$

其中，$h(x,\theta_t)$ 为基于 x 和 θ 的输出。

为了克服决策树模型精度不高、易出现过拟合的问题，RFR 引入了 bagging（套袋）[15]和随机子空间[16]的思想。

（1）bagging 思想

从原始样本中有放回地随机抽取多个训练样

本，且每个训练样本量等于原始样本量，并对每个训练样本分别构建回归决策子树 T，最后取各棵树的平均值作为最终预测结果。

假设 S 为原始样本，N 为 S 中的样本数，则 S 中每个样本没有被抽取的概率为 $\left(1-\frac{1}{N}\right)^N$。

定理 1 当 $N\to\infty$ 时：

$$\left(1-\frac{1}{N}\right)^N \approx \frac{1}{\mathrm{e}} \approx 0.368 \tag{2}$$

式（2）表明，每次约有 36.8%的样本未被抽取，称为 OOB（out-of-bag，袋外数据）。bagging 思想不仅可以使用随机化建立更多的回归决策子树，同时还保证了子树与子树之间的独立性。

（2）随机子空间思想

在构建回归决策子树的过程中，每个分裂节点都从总的特征空间中随机抽取特征子空间 F 作为节点的候选特征集，并从中选取最优特征进行分裂。该方法既保证了树与树之间的节点以及每棵树节点之间特征子集都不同，又保证了树的独立性和多样性，进而提高了 RFR 节点分裂的随机性。

在 RFR 中，回归决策子树 T 和特征子空间中的特征个数 f 决定着模型最终的预测性能。

2.2 泛化误差

泛化误差反映了模型对训练集以外的数据的预测能力，是判断模型好坏的重要指标。

定义 2 假设从独立同分布的随机向量 (X,Y) 中抽取训练集，形成的训练集之间各自独立。则输出 $h(X)$ 的均方泛化误差为 $E_{X,Y}\left(Y-h(X)\right)^2$。

当回归决策树足够多时，根据强大数定律，且 $h_t(X)=h(X,\theta_t)$，可得到如下定理。

定理 2[14] 当 $t\to\infty$ 时，均方泛化误差收敛于：

$$E_{X,Y}\left(Y-\bar{h}(X,\theta_t)\right)^2 \to E_{X,Y}\left(Y-E_\theta h(X,\theta)\right)^2 = PE_{\mathrm{b}}^* \tag{3}$$

其中，θ_t 为第 t 棵回归决策树子的随机变量，E_θ 为数学期望。式（3）右边即 RFR 的泛化误差 PE_{b}^*。

定理 2 表明，RFR 随着回归决策子树 t 的增加会逐步收敛，不会出现过拟合的问题，但 PE_{b}^* 最终会趋于一个稳定值。

定义 3 每一棵回归决策树 $h(X,\theta)$ 的平均泛化误差定义为：

$$PE_{\mathrm{c}}^* = E_\theta E_{X,Y}\left(Y-h(X,\theta)\right)^2 \tag{4}$$

定理 3[14] 假设对于所有的随机变量 θ，回归决策子树都是无偏的，即 $EY=E_X h(X,\theta)$，则有：

$$PE_{\mathrm{c}}^* \leqslant \bar{\rho} PE_{\mathrm{b}}^* \tag{5}$$

其中，$\bar{\rho}$ 为残差 $Y-h(X,\theta)$ 和 $Y-h(X,\theta')$ 的相关系数，θ 和 θ' 相互独立。

2.3 算法流程

步骤 1 利用 bagging 思想，随机产生样本子集。

步骤 2 利用随机子空间思想，随机抽取 f 个特征，进行节点分裂，构建单棵回归决策子树。

步骤 3 重复步骤 1、步骤 2，构建 T 棵回归决策子树，每棵树自由生长，不进行剪枝，形成森林。

步骤 4 T 棵决策子树的预测值取平均，作为最终结果。

3 基于 RFR 算法的 PM2.5 浓度预测模型设计

基于 RFR 算法建立 PM2.5 浓度预测模型（以下简称 RFRP 模型）的整体设计思想是：先确定与 PM2.5 浓度相关的特征因素，并收集整理数据集，应用 RFR 算法建立模型，然后通过调整参数的最优组合，不断优化模型。其设计流程如图 1 所示。

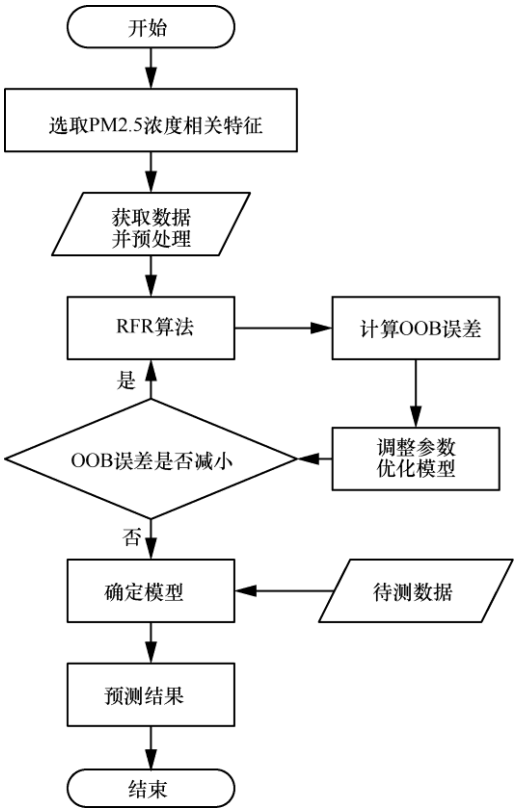

图 1　RFRP 模型的设计流程

根据图 1，可将其主要设计步骤总结如下。

步骤 1　选取相关特征 $f_1, f_2, \cdots, f_n$，收集数据并进行预处理，得到建模数据。

步骤 2　将建模数据作为 RFR 算法的输入，训练得到 RFRP 模型。

步骤 3　调整参数，优化模型。

步骤 4　应用优化后的模型对新的数据进行预测。

3.1　特征选取与数据预处理

特征选取涉及影响 PM2.5 浓度的不同要素，选择出相关的、信息量大的、有差异的、独立的影响特征，是建立预测模型中关键的一步。特征选取是将原始数据转化为特征，以便更好地表示模型处理的实际问题，提升对于未知数据的准确性。模型可以通过优质特征描述的数据的固有结构进行很好的学习，即使不是最优的模型，优质的特征也可以得到较好的效果。通过特征选择，也可以提高模型的运行效率，使模型泛化能力更强，减少过拟合。

经过充分的调研和实验分析，选取气象条件、大气污染物浓度和季节所包含的22项因素为特征分析对象。另外，考虑到气象条件和大气污染物浓度存在的时延，前日的气象条件和大气污染物会对当日的 PM2.5 浓度产生影响，因此，需要将前日的数据作为一项重要的指标。

对天气后报网（http://www.tianqihoubao.com）和天气网（http://www.tianqi.com）上的公开数据进行爬取，整理出西安市 2013 年 10 月 29 日—2016 年 12 月 28 日的历史数据，共 1 156 条。其中，随机选取其中的 75%作为训练集建立模型，剩余的 25%用来测试模型的准确率。表 1 列出了选取出的具体特征因素。

表 1　特征因素选取

特征类型	具体特征	符号	取值
气象条件	当（前）天风力	Wind_pow_1(2)	[1,2,3,4,5]
	当（前）天风向	Wind_dir_1(2)	[1,2,3,4,5,6,7,8,9]
	当（前）天最高温	Tem_high_1(2)	real
	当（前）天最低温	Tem_low_1(2)	real
	当（前）天天气	Weather_1(2)	[1,2,3,4]
大气污染物	当（前）天 O_3	O_3_1(2)	real
	当（前）天 NO_2	NO_2_1(2)	real
	当（前）天 CO	CO_1(2)	real
	当（前）天 SO_2	SO_2_1(2)	real
	当（前）天 PM10	PM_{10}_1(2)	real
	前天 PM2.5	$PM_{2.5}$_2	real
季节	季节	Mon	[1,2,3,4]

如表 1 所示，选取气象条件、大气污染物浓度和季节 3 个方面共 22 个相关特征因素。其中温度和各大气污染物浓度属于数值型特征用“real”表示，其余均属于非数值型。在处理非数据值型特征时，本文对其进行了量化：将非数值型特征转化为离散的数值型特征，并放入“[]”，以此表示取值范围。如“Wind_pow_1(2)”的取值为“[1,2,3,4,5]”，分别代表 5 种风力类型：微风、1~2 级、3 级、3~4 级、4~5 级；“Weather_1(2)”的取值为“[1,2,3,4,]”，代表 4 种天气类型：晴、多云、霾、降水。

对原始数据进行预处理，形成一个 22×N 的矩阵：

$$A=\begin{bmatrix} a_{1f_1} & a_{1f_2} & \mathrm{L} & a_{1f_{22}} \\ a_{2f_1} & a_{2f_2} & \mathrm{L} & a_{2f_{22}} \\ \mathrm{M} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ a_{Nf_1} & a_{Nf_2} & \mathrm{L} & a_{Nf_{22}} \end{bmatrix} \quad (6)$$

其中，每一行代表同一时间各个相关特征的实测值；每一列代表同一相关特征的样本数量，用 N 表示；$f_1, f_2, \mathrm{L}, f_{22}$ 代表选取的风力、风向等 22 项相关特征。

由于 RFR 算法对数据的单位和量纲并不敏感以及两大随机特性，所以不需要对整理好的数据进行归一化处理和特征选择。在 BP-NN 算法中，不同属性单位的数据不仅影响神经网络的输出精度，还降低了程序运行的收敛速度，所以需要将数据进行归一化处理，保证数据值在同一区间，避免不同特征数量级差别较大所造成的影响。与 BP-NN 算法相比，RFR 算法大大简化了数据处理过程，提高了数据处理效率。

3.2 RFRP 模型的实现

RFRP 模型的实现流程如图 2 所示。

在 PM2.5 预测模型的实现过程中，RFR 算法作为一种集成学习算法，应用并行方式建立许多小而薄弱的 PM2.5 浓度预测模型，各子模型独立地学习并预测出 PM2.5 浓度值，最后再将所有预测值取平均作为最终的预测结果。因此，应用基于 RFR 算法建立的模型，其预测结果优于任何一个单预测模型做出的预测。

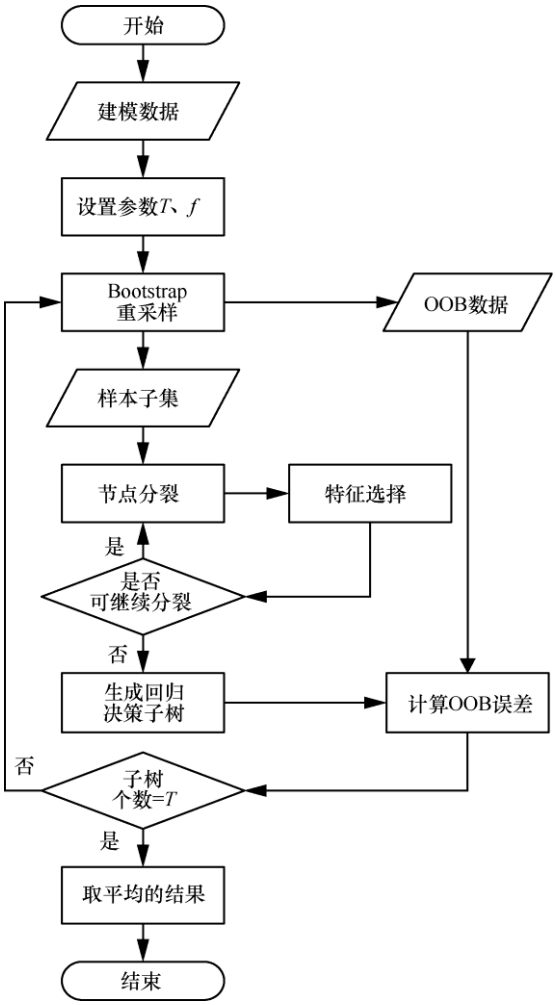

图 2　基于 RFR 算法的 PM2.5 浓度预测模型的实现流程

原始数据经过预处理后，得到矩阵 A，直接作为 RFR 算法的输入。利用 Bootstrap 重采样随机生成样本子集，同时产生 OOB 数据。在树的每个节点随机选择 f 个特征进行分裂（f<F），建立多个 PM2.5 预测模型，并利用 OOB 数据计算 OOB 误差。最后将多个模型的预测值取平均作为模型输出。实现流程的伪代码算法如下。

输入　训练集 S；测试集 U；参数：回归决策子树 T、特征总数 F、特征选择个数 f

输出　PM2.5 浓度预测值

Function RandomForestRegression(*S*,*F*)　//定义随机森林回归函数

for *i* from 1 to *T* do:

$S^{(i)}$ ← randomly split *S*　//对 *S* 进行 Bootstrap 重采样随机生成样本子集

h_i ← Subtree($S^{(i)}$,*F*)　//调用 Subtree 函数

end for

return h_i

end Function

Function Subtree(*S*,*F*)　//定义建立子树函数

at each node:　//在每个节点操作

f ← randomly select features from *F*　//从 *F* 中随机选取 *f* 个特征（*f*<*F*）

split on the best feature in *f*　//从 *f* 中选择最优的特征进行节点分裂

calculate OOB error　//计算 OOB 误差

return the subtree

end Function

3.3　RFRP 模型的性能度量指标

采取通用的模型精度和效率作为度量指标，进行 RFRP 模型的参数调整和性能分析。

（1）模型精度

包括 MRE（mean relative error，平均相对误差）和确定性系数（R^2）。其中，MRE 越小，R^2 越大，说明模型精度越高。

$$MRE=\frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{Q}_i-Q_i}{Q_i}\right| \tag{7}$$

$$R^2=1-\frac{\sum_{i=1}^{N}\left(Q_i-\hat{Q}_i\right)^2}{\sum_{i=1}^{N}\left(Q_i-\overline{Q_c}\right)^2} \tag{8}$$

式（7）、式（8）中，Q_i 为真实值；$\hat{Q}_i$ 为预测值；$\overline{Q_c}$ 为真实值的均值；*N* 为样本数。

（2）模型效率

指模型的训练时间。训练时间越小，说明模型预测效率越高。

3.4　参数优化

影响 RFRP 模型预测能力的参数主要有两个：构建回归决策子树的棵数 *t* 和随机选择特征分裂时特征选择的个数 *f*。考虑到实际情况中 RFR 算法的泛化误差不能直接计算求得，参考文献[17]中指出，可以在测试集上使用交叉验证的方式估计泛化误差，但这样会导致巨大的计算量，降低算法的运行效率；Breiman[15]提出采用 OOB 误差估计的方法，只需要增加少量计算即可，并指出 OOB 误差近似于交叉验证的结果。OOB 估计也可以用来估计单棵回归决策子树，计算森林中每棵子树的 OOB 估计的均值即可得到 RFRP 模型的泛化误差。

因此，采用 OOB 估计和模型运行时间为指标，选择最优的参数组合 *t* 和 *f*。考虑到两个参数均对 OOB 估计产生影响，且组合参数计算量较大（*t*×*f*），故在固定一个参数的前提下对另一个参数进行调整。由于 RFR 算法的两大随机特性，模型每次抽取样本和特征都不同，为准确衡量模型的性能，每次实验重复进行 20 次，取 20 次结果的均值作为最终实验结果。

首先选取 $f=\frac{F}{2}=11$，观察 OOB 误差和模型训练时间随 *t* 的变化情况，如图 3 所示。由图 3 可知，随着 *t* 增加，OOB 误差表现出不同的变化趋势：当 *t*<20 时，急剧下降；当 20<*t*<60 时，缓慢下降；当 *t*>60 时，接近收敛。

同时，随着 *t* 值的增加，模型的训练时间也在直线上升。综合误差和效率两个方面考虑，选取最优值 *t*=60。OOB 误差和模型训练时间随 *t* 的变化情况如图 3 所示。

其次固定 *t*=60，观察 OOB 误差和模型运行时间随 *f* 的变化情况，如图 4 所示。由图 4 可知，随着 *f* 增加，模型的训练时间总体呈上升趋势；而 OOB 误差先迅速下降，之后下降缓慢。在 *f*>18 后，模型 OOB 误差变化幅度很小，但其训练时间上升幅度较大。因此，选择 *f*=18 为最优。综合上述实验，最终选定模型的最优参数组合为 *t*=60，*f*=18。

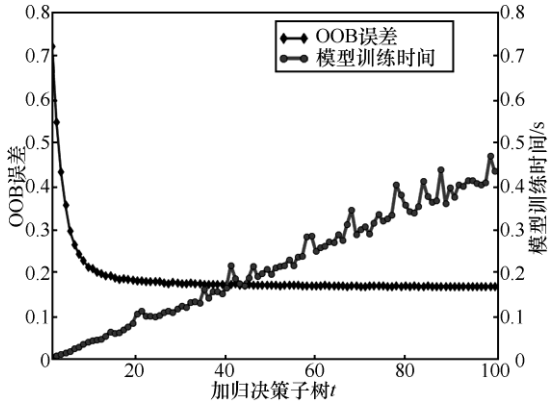

图 3　OOB 误差和模型训练时间随回归决策子数 t 的变化情况

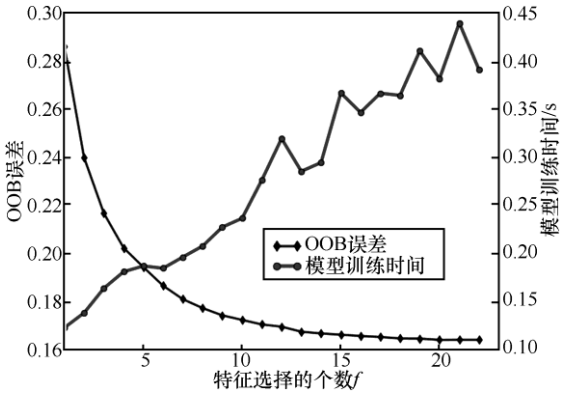

图 4　OOB 误差和模型训练时间随特征选择个数 f 的变化情况

4　实验结果分析

本文在 Python 环境下进行实验。实验设备配置：操作系统 Ubuntu16.04 LTS，处理器 Intel Core i7-4790 CPU @ 3.60 GHz×8，内存 16 GB，硬盘 1 TB。

在上述实验环境配置下，采用构造回归决策子树为 60、特征子空间为 18 的最优参数组合对训练集进行训练，建立 PM2.5 浓度预测模型。该模型反映了 PM2.5 浓度与各影响因素之间复杂的非线性关系。运用该模型对测试集进行预测，并计算模型的精度和效率，结果见表 2。

表 2　RFRP 模型性能分析

数据集	MRE	R^2	训练时间/s
训练集	0.165	0.942	0.283
测试集	0.159	0.934	0.281

由表 2 可以看出，无论是训练集还是测试集，RFRP 模型的确定性系数 R^2 均达到 90%以上，说明模型具备良好的学习能力与泛化能力；就误差而言，训练集和测试集上的误差分别为 0.165 和 0.159，在可接受的范围内。此外，RFRP 模型在测试集上的各项指标均未远远超过训练集上的各项指标，说明模型没有出现过拟合，泛化性能较好，能够有效地预测 PM2.5 的浓度。测试集上的预测结果如图 5 所示。

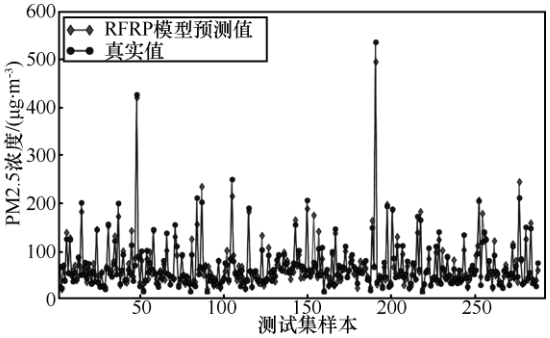

图 5　RFRP 模型在测试集上的预测结果

5　算法对比分析

为进一步验证 RFRP 模型在 PM2.5 浓度预测问题上的优劣，综合评价模型性能，本文应用参考文献[18]描述的 BP-NN 算法预测模型进行理论分析和实验对比，并采用相同数据集建立 PM2.5 浓度预测模型对测试集上的数据进行预测。

5.1　原理方法对比

随机森林是以单棵决策树作为基本分类预测器的一个集成学习模型，结合了集成学习理论和随机子空间的两大随机思想，分别用于随机选取训练样本和随机选取分裂属性值。同时克服了决策树易过拟合的缺点，对于噪声和异常值不敏感，具有较好的顽健性，是一种有效的分类预测方法，具有较高的精度和较强的泛化能力。参考文献[14]针对随机森林的优势给出了数学理论上的推理证明。随机森林的收敛定理

（convergence theorem），即式（3）证明了随机森林不会出现过拟合问题；泛化误差界（generalization error bound），即式（5）给出了随机森林预测的一个理论上界；袋外估计（out-of-bag estimation），即式（2）提出了一种利用袋外数据估计泛化误差界的 OOB 误差估计方法，该方法效率较高，其结果近似于需要大量计算的 K 折交叉验证。

与随机森林相比，BP-NN 也适用于求解内部机制较复杂的非线性问题。BP-NN 是一种按误差逆传播算法训练的多层前馈网络，基本思想是根据梯度下降法，使得网络的实际输出和期望输出均方差得到最小，拓扑结构包括输入层（input layer）、隐含层（hidden layer）、输出层（output layer）[19]，分为信息正向传播和误差反向传播。通过实际输出与期望输出之间的误差来调整各层连接权值和各节点之间的阈值，调整参数的过程是周而复始的，一直到满足终止条件为止，这个过程也是 BP-NN 的学习训练过程。

图 6 是本文设计的 3 层 BP-NN 预测模型的网络结构。

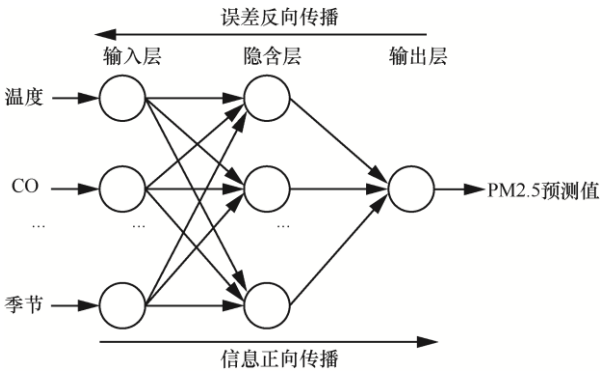

图 6　本文设计的 BP-NN 预测模型结构

BP-NN 具有大规模并行、自学习、自适应、高容错、泛化能力强等优点，但随着应用领域的扩大，BP-NN 也暴露出了一些缺点和不足。

（1）收敛速度慢[20]

BP-NN 本质为梯度下降法，优化的目标函数较复杂，会出现“锯齿形”现象。

（2）局部最小化[21]

从数学角度分析，BP-NN 属于一种局部搜索的优化方法。

（3）网络结构复杂难定[21]

关于网络结构的选择，并没有权威的理论指导，往往根据经验选择。而网络结构的选择会直接影响网络是否收敛，出现过拟合、容错性下降等问题。

（4）数据需要进行归一化处理[22,23]

BP-NN 的隐含层一般采用的是 Sigmoid 函数，根据 Sigmoid 函数特点，需要将输入向量进行归一化后才能作为网络的输入量。目的在于提高收敛速度和灵敏性及有效避开函数的饱和区。

5.2　预测结果对比

与 RFR 算法不同，BP-NN 算法需要对数据进行归一化处理。本文采用隐含层为 Logistic 激活函数的 BP-NN，网络结构为 22-10-1，即输入层节点数为 22，隐含层节点数为 10，输出层节点数为 1。此外，本实验将与 LR（linear regression，线性回归）、SVM（support vector machine，支持向量机）两个经典算法进行比较。4 个模型预测结果对比见表 3。

表 3　RFRP 模型和 BP-NN 模型预测结果对比

模型	MRE	R^2	训练时间/s
RFRP	0.159	0.934	0.281
BP-NN	0.161	0.971	4.766
LR	0.322	0.827	0.108
SVM	0.216	0.893	0.132

由表 3 可知，在模型精度方面，RFRP 和 BP-NN 表现明显优于 LR 和 SVM。而 RFRP 和 BP-NN 两个模型相差不多：RFRP 的 MRE 误差为 0.159，相比于 BP-NN 的 0.161，降低了 0.002，而确定性系数还略低于 BP-NN 的 0.971。但在模型的运行效率方面，RFRP、LR、SVM 的模型训练时间均远少于 BP-NN 的 4.766 s。

综合比较预测结果，RFRP 和 BP-NN 两个模型在 PM2.5 浓度预测问题上的效果优于 LR、

SVM，均具有较高的准确率，能够较好地解决复杂的非线性问题，BP-NN 模型的拟合度甚至还优于 RFRP 模型。在运行效率方面，RFRP 模型的优势就明显了，其运行时间为 0.281 s，只有 BP-NN 模型的 5.88%。

相比而言，RFR 算法结构简单易于理解、实现模型简单、参数易调节，且不需要对数据进行归一化处理和交叉验证。而 BP-NN 存在着计算量大、学习效率低，易过拟合、网络结构参数难确定的缺点。最主要的是 RFR 算法的两大随机特性解决了过拟合问题，同时提升了训练速度。所以，RFRP 在综合性能上具有一定的优势。

最终的实验结果表明，RFRP 模型与 BP-NN 模型相比，在不降低预测精度的同时，有效地提高了 PM2.5 的预测效率。

6 结束语

随机森林作为一种高效的机器学习算法，已经广泛应用到多个领域，然而在 PM2.5 浓度预测中的应用研究却很少。本文通过对基于 RFR 算法的 PM2.5 浓度预测模型的研究和实现，发现该模型具有较高的预测精度，且为 PM2.5 浓度预测问题提供了可参考的特征。实验证明，基于 RFR 算法设计出的 PM2.5 浓度预测模型，相比于 BP-NN 模型，不仅能保证较高的预测精度，还能大幅度地提高模型的预测效率。

参考文献：

[1] MA H, SHEN H, LIANG Z, et al. Passenger's exposure to PM2.5, PM10, and CO_2 in typical underground subway platforms in shanghai[J]. Lecture Notes in Electrical Engineering, 2014(261): 237-245.

[2] SCHWARTZ J, DOCKERY D W, NEAS L S. Is daily mortality associated specifically with fine particles? [J]. Journal of the Air and Waste Management Association, 1996(46): 927.

[3] 马丽梅, 张晓. 中国雾霾污染的空间效应及经济、能源结构影响[J]. 中国工业经济, 2014(4): 19-31.
MA L M, ZHANG X. The spatial effect of China's haze pollution and the impact from economic change and energy structure[J]. China Induxtrial Economics, 2014(4): 19-31.

[4] 付倩娆. 基于多元线性回归的雾霾预测方法研究[J]. 计算机科学, 2016, 43(6A): 526-528.
FU Q R. Research on haze prediction based on multivariate linear regression[J]. Computer Science, 2016, 43(6A): 526-528.

[5] CHELANI A B, DEVOTTA S. Prediction of ambient carbon monoxide concentration using nonlinear time series analysis technique[J]. Transportation Research Part D: Transport and Environment, 2007, 12(8): 596-600.

[6] 毛毳, 孙宇, 冯葭, 等. 空气中 PM2.5 浓度的灰色预测与关联因素分析[J]. 宁夏大学学报(自然科学版）, 2014, 35(3): 283-288.
MAO C, SUN Y, FENG C, et al. Grey forecast and correlation factors analysis of PM2.5 in the air[J]. Journal of Ningxia University(Natural Science Edition), 2014, 35(3): 283-288.

[7] ZHANG C J, DAI L J, MA L M. Rolling forecasting model for PM2.5 concentration based on support vector machine and particle swarm optimization[C]//International Symposium on Optoelectronic Technology and Application, May 9-11, 2016, Beijing, China. [S.l.:s.n.], 2016.

[8] BALACHANDRAN S, CHANG H, PACHON J, et al. Bayesian-based ensemble source apportionment of PM2.5[J]. Environment Science & Technology, 2013, 47(23): 13511-13518.

[9] GRIVAS G, CHALOULAKOU A. Artificial neural network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece[J]. Atmospheric Environment, 2006, 40(7): 1216-1229.

[10] 马天成, 刘大铭, 李雪洁, 等. 基于改进型 PSO 的模糊神经网络 PM2.5 浓度预测[J]. 计算机工程与设计, 2014, 35(9): 3258-3262.
MA T C. LIU D M, LI X J, et al. Improved particle swarm optimization based fuzzy neural network for PM2.5 concentration prediction[J]. Computer Engineering and Design, 2014, 35(9): 3258-3262.

[11] LIN C J, CHEN C H, LIN C T. A hybrid of cooperative particle swarm optimization and cultural algorithm for neural fuzzy networks and its prediction applications[J]. IEEE Transactions on Systems Man & Cybernetics Part C, 2009, 39(1): 55-68.

[12] 杨云, 付彦丽. 基于 T-S 模型模糊神经网络的 PM2.5 质量浓度预测[J]. 陕西科技大学学报(自然科学版), 2015, 33(6): 162-166.
YANG Y, FU Y L. The prediction of mass concentration of PM2.5 based on T-S fuzzy neural network[J]. Journal of Shaanxi University of Science & Technology(Natural Science Edition), 2015, 33(6): 162-166.

[13] MCKEEN S, CHUNG S H, WILCZAK J, et al. Evaluation of several PM2.5 forecast models using data collected during the ICARTT/NEAQS 2004 field study[J]. Journal of Geophysical Research, 2007, 112(D10): 541-553.

[14] BREIMAN L, CUTLER A. Random forests[J]. Machine

Learning, 2001, 45(1): 5-32.
[15] BREIMAN L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123-140.
[16] HO T K. The random subspace method for constructing decision forests[J]. IEEE Transactions on Pattern Analysisand Machine Intelligence, 1998, 20(8): 832-844.
[17] WOLPERT D H, MACREADY W G. An efficient method to estimate bagging's generalization error[J]. Machine Learning, 1999, 35(1): 41-45.
[18] CHEN B, WANG X P, YU L X, et al. Prediction of PM2.5 concentration in a agricultural park based on artificial neural network[J]. Advance Journal of Food Science and Technology, 2016, 11(4): 274-280.
[19] 吕昌国. 基于 BP 算法的网格资源调度研究[D]. 哈尔滨: 哈尔滨理工大学, 2007.
LV C G. Grid resources scheduling research based on BP algorithm[D]. Harbin: HUST, 2007.
[20] 鲍立威, 何敏, 沈平. 关于 BP 模型的缺陷的讨论[J]. 模式识别与人工智能, 1995, 8(1): 1-5.
BAO L W, HE M, SHEN P. Argument on the shortcoming of BP-model[J]. Pattern Recognition and Artificial Intelligence, 1995, 8(1): 1-5.
[21] BP 神经网络优缺点的讨论 [EB/OL]. (2008-12-01)[2017-03-30]. http://www.paper.edu.cn/releasepaper/content/ 200812-27.
BP neural network to discuss the advantages and disadvantages [EB/OL]. (2008-12-01)[2017-03-30]. http://www.paper.edu.cn/releas- epaper/ content/200812-27.
[22] 赵会敏, 雒江涛, 杨军超, 等. 集成 BP 神经网络预测模型的研究与应用[J]. 电信科学, 2016, 32(2): 60-67.
ZHAO H M, LUO J T, YANG J C, et al. Research and application of prediction model based on ensemble BP neural network[J]. Telecommunications Science, 2016, 32(2): 60-67.
[23] 张国玲. 基于情感神经网络的风电功率预测[J]. 电信科学, 2017, 33(3): 168-172.
ZHANG G L. An emotional neural network based approach for wind power prediction[J]. Telecommunications Science, 2017, 33(3): 168-172.

[作者简介]

杜续（1989−），男，西安邮电大学硕士生，主要研究方向为大数据分析与数据挖掘。

冯景瑜（1984−），男，博士，西安邮电大学副教授，主要研究方向为无线通信安全、认知无线网络等。

吕少卿（1987−），男，博士，西安邮电大学讲师，主要研究方向为大数据分析与网络安全。

石薇（1980−），女，西安邮电大学讲师，主要研究方向为大数据分析与通信网络规划。