第 6 卷◆第 5 期◆版本 1.0◆2025 年
文章类型：论文|刊号（ISSN）：2737-4580(P) / 2737-4599(O)

基于优化 PSO-SVR 算法的 $PM_{2.5}$ 浓度预测

DOI:10.12238/jpm.v6i5.8032

张亚博　南守珺　唐彦　杨云飞
新疆工程学院土木工程学院

[摘　要] 本研究针对支持向量回归（SVR）算法中核函数和惩罚参数选择导致的回归精度不足问题，提出采用粒子群优化算法（PSO）对 SVR 的核参数和惩罚参数进行优化选择，以提升 $PM_{2.5}$ 浓度预报的准确性。以北京市为研究区域，选取主要大气污染物、气象要素及 GNSS 天顶对流层延迟（ZTD）等作为预测变量，构建同期 $PM_{2.5}$ 浓度预测模型。实验结果显示，PSO-SVR 模型在回归精度方面显著优于 SVR 模型、随机森林模型和 BP 神经网络模型，能更有效地捕捉数据序列的潜在变化特征。在短期预测中，该模型不仅展现出优异的预测效果，同时保持了较好的泛化能力，体现出较强的适应性。

[关键词] $PM_{2.5}$ 浓度预测；支持向量回归；粒子群优化算法；优化算法

[中图分类号] P237；P951　　[文献标志码] A

PM2.5 concentration prediction based on optimized PSO-SVR algorithm

Zhang Yabo　Nan Shoujin　Tang Yan　Yang Yunfei
School of Civil Engineering，Xinjiang Institute of Technology

[Abstract] This study addresses the issue of insufficient regression accuracy caused by kernel function and penalty parameter selection in Support Vector Regression（SVR）algorithms. It proposes using Particle Swarm Optimization（PSO）to optimize the kernel parameters and penalty parameters of SVR，aiming to improve the accuracy of PM2.5 concentration forecasting. Taking Beijing as the research area，major atmospheric pollutants，meteorological elements，and GNSS zenith tropospheric delay（ZTD）are selected as prediction variables to construct a concurrent PM2.5 concentration prediction model. Experimental results show that the PSO-SVR model significantly outperforms the SVR model，random forest model，and BP neural network model in terms of regression accuracy，effectively capturing potential changes in data sequences. In short-term predictions，this model not only demonstrates excellent predictive performance but also maintains good generalization ability，showing strong adaptability.

[Key words] PM2.5 concentration prediction；support vector regression；particle swarm optimization algorithm；optimization algorithm

$PM_{2.5}$（空气动力学直径≤2.5μm 的细颗粒物）因其具有较　　强的吸附特性，能够携带多种有毒有害物质，并通过呼吸深入

人体肺泡区域。从而提高肺癌等呼吸系统疾病的患病风险及致死率[1]，因此，有必要分析 $PM_{2.5}$浓度的时间序列变化规律，并建立准确可靠的 $PM_{2.5}$预测模型。

传统的物理模型需要深厚的气象学知识以及假设条件，极其复杂且耗时。随着计算智能的发展，经验模型已成为 $PM_{2.5}$预测的主流方法。由于已建立的物理模型与特定区域高度相关，因此不适用于其他地区。统计模型基于历史数据快速推断未来时间序列，结构简单且建模成本低，在一定程度上缓解了物理模型的局限性。传统统计模型如隐半马尔可夫模型[2]和多元线性回归[3]虽然计算效率较高，然而这些方法中的统计特征是基于历史数据挖掘的，在 $PM_{2.5}$浓度的短期预报中效果不佳。Hu [4]等人开发了一种混合机器学习模型，利用小波分解方法、模拟退火和反向传播神经网络来预测 $PM_{2.5}$浓度。人工神经网络能够从过程数据中捕捉复杂的非线性动态[5]。近年来，随着卫星导航技术的进步，GNSS 衍生的天顶对流层延迟（ZTD）参数被证明与雾霾存在显著关联[6]，这为 $PM_{2.5}$预测提供了新的数据维度。

本研究在前人研究基础上，创新性地融合大气污染物、气象要素和 ZTD 三类特征，采用粒子群智能算法优化支持向量回归机的关键参数（包括核参数和惩罚参数），建立 PSO-SVR 组合预测模型。将北京市作为一个单点从时间维度进行 $PM_{2.5}$浓度预测，并分析其精度。

1 模型算法简介

1.1 支持向量回归模型

支持向量回归模型（SVR）[7]作为一种机器学习模型，适用于小样本、非线性及高维复杂数据的回归预测问题，具有出色的泛化能力。与传统 SVM 不同，SVR 引入了不敏感函数 ε 作为损失函数等，从而提升了模型的适应性。假设给定的样本集（x_i，y_i），i=1，2，…，n，$x_i \in R_n$，$y_i \in R$，对于线性函数：

$$f(x)=\omega x+k \tag{1}$$

考虑到函数过拟合，并基于结构最小化的目的引入松弛因子 ξ 和 ξ*来放宽条件，允许存在一定的误差，于是，线性回归函数转化为带约束的优化函数：

$$\min_{\omega,\xi,\xi^*} z=\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}(\xi_i+\xi_i^*) \tag{2}$$

$$s.t\begin{cases} y_i-x_i^{\mathrm{T}}\omega\le\varepsilon+\xi_i \\ x_i^{\mathrm{T}}\omega-y_i\le\varepsilon+\xi_i^*,\forall i \\ \xi_i\ge 0,\xi_i^*\ge 0 \end{cases} \tag{3}$$

式中，惩罚参数 C（其中 C ＞ 0）用来控制超出误差 ε 的惩罚程度。为了求解该优化问题，可以引入拉格朗日乘子，将原非线性回归问题转换为高维特征空间中的线性回归问题，最终得到的回归函数表达式为：

$$f(\mathrm{x})=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)b(x,x_i)+k \tag{4}$$

式中，k（x，x_i）为核函数，其作用是将原本适用于线性问题的 SVR 扩展到更高维的特征空间，从而有效处理非线性回归问题。研究表明，RBF 核函数因其良好的泛化能力和对高维数据的适应性，其数学表达式为：

$$k(x,x_i)=\exp[-\|x-x_i\|^2/2\delta^2] \tag{5}$$

其中，δ 作为径向基核函数的核宽参数。在支持向量回归模型中，预测性能对超参数的选择具有高度敏感性，其预测精度主要取决于以下两个关键参数：惩罚系数 C 和 RBF 核参数 σ，过大或过小的参数值都可能导致欠拟合或过拟合问题[8]。

1.2 粒子群优化算法

粒子群优化算法是由 Kennedy 等[9]基于人工生命和演化计算理论提出的一种进化计算技术，其中每个粒子代表一个候选解决方案，然后根据全局最佳位置和局部最佳位置对每个粒子进行更新来寻找最优解。这种基于群体智能的优化机制，按照式（6）使得粒子能够在解空间中高效地探索最优解。

$$\begin{cases} v_{ij}^{k+1}=\omega v_{ij}^k+c_1\times r_1\times(Pbest_{ij}-x_{ij}^k)+c_2\times r_2\times(Gbest_{ij}-x_{ij}^k) \\ x_{ij}^{k+1}=x_{ij}^k-v_{ij}^{k+1} \end{cases} \tag{6}$$

式中，Pbest 为个体经验，Gbest 群体经验，k 为迭代次数，搜索维度 j=1，2，…，D，惯性权重 ω∈[0，1]，学习因子 c_1，c_2∈[0，2]，当 c_1较小时会导致粒子缺乏认知能力，c_2较小会降低粒子间的信息共享能力，r_1，r_2为[0，1]之间的随机数。

2 实验数据

2.1 研究区域及数据

北京市地理区位特征显著，北部与辽东半岛接壤，南部毗邻山东半岛，东部面向渤海湾。气候类型属典型的暖温带半湿润半干旱季风气候，具有明显的季节差异性特征。

本研究选取 2017-2020 年连续四年的多源观测数据集作为研究基础，数据来源包括：

1）空气质量数据：涵盖 $PM_{2.5}$、PM_{10}、CO、SO_2、O_3、NO_2 等主要污染物的日均浓度监测数据，采集自环境专业知识服务系统；

2）气象观测数据：气压、风速、风向、温度、相对湿度

和降水量等要素，来源于欧洲中期天气预报中心的 ERA-5 再分析数据集；

3）GNSS 观测数据：天顶总延迟（ZTD）数据由中国地震局 GNSS 数据产品服务平台提供。

3 实验分析

3.1 相关性分析

本研究采用 Spearman 秩相关分析探讨各因素与 $PM_{2.5}$ 浓度之间的相关性。通过设定显著性水平 α =0.01 的条件下，筛选出与 $PM_{2.5}$ 浓度具有显著统计学相关性的变量作为潜在影响因素。

3.2 数据预处理

对 2017-2020 年北京市的主要污染物、气象因素和 ZTD 的历史数据进行缺失值剔除和相关性分析，为防止各变量因数值差异过大而导致对预测性能的影响，对数据按照式（7）进行标准化处理，将所有数据归一化到[0，1]区间。

$$x_i = \frac{y_i - y_{min}}{y_{max} - y_{min}} \qquad (7)$$

式中，x_i 为归一化后的数据，y_i 为原始数据，y_{max} 和 y_{min} 为原始数据的最大值和最小值。

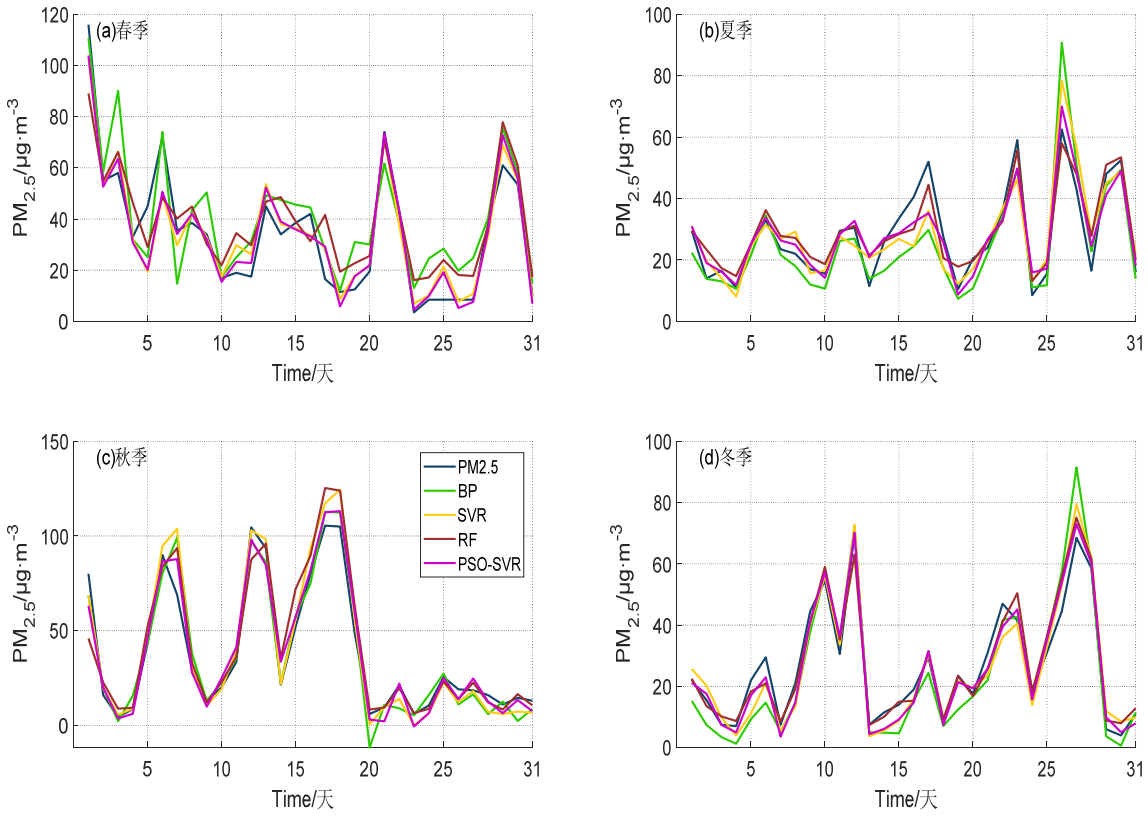

图 1　各季节各模型预测结果与误差绝对值

Fig.1　The prediction results and absolute values of errors of each model in different seasons

3.3 实验结果与分析

由于 $PM_{2.5}$ 浓度呈季节性变化[10]。选取北京市 2017-2020 年的数据进行实验，预测每个季节短期 $PM_{2.5}$ 浓度值，以春季为例，共有 369 天，其中前 338 天天进行训练，最后 31 天进行预测。通过 PSO 算法对 SVR 的核参数 σ 和惩罚系数 C 两个参数迭代寻优，经过 PSO 算法得到的最优参数 C=910.154， σ =0.127703。并与 BP 神经网络、随机森林和 SVR 的预测结果进行精度评估。并用相同方法将夏季、秋季和冬季的 $PM_{2.5}$ 浓度进行短期预测。各模型的预测结果和预测精度如图 1，表 1 为评价各模型的预测精度，采用 RMSE（μg/m^3）、MAE（μg/m^3）、MAPE（%）和 R^2 来评定模型精度。

根据图 1 和表 1 的结果分析，四种模型的预测结果与真实值均较为吻合，但 PSO-SVR 模型展现出显著的优越性。该模型不仅具有最佳的预测精度和稳定性，尤其对浓度拐点的捕捉能力突出，而且通过 PSO 算法有效优化了超参数，其回归性能明显优于传统 SVR 模型。随机森林模型虽能基本反映 $PM_{2.5}$ 浓度的变化趋势，但在极值点的预测上存在明显不足。相比之下，BP 神经网络模型的预测效果较差。

表 1　各季节各模型模型预测精度

Tab.1　The prediction accuracy of each model in different seasons

Season	Model	RMSE	MAE	MAPE	R^2	Season	Model	RMSE	MAE	MAPE	R^2
春	BP	12.48	10.00	58.97	0.74	夏	BP	8.85	6.29	21.91	0.64
	SVR	8.64	6.19	26.72	0.88		SVR	7.55	5.78	23.33	0.73
	RF	12.04	9.82	56.92	0.76		RF	5.20	4.17	21.60	0.87
	PSO-SVR	8.03	5.62	23.27	0.89		PSO-SVR	5.50	4.16	17.99	0.86
秋	BP	8.97	6.63	33.37	0.93	冬	BP	7.63	5.83	30.20	0.83
	SVR	9.19	6.24	24.77	0.93		SVR	5.38	4.47	26.62	0.92
	RF	11.37	7.77	22.72	0.89		RF	4.43	3.48	17.04	0.94
	PSO-SVR	7.12	5.68	21.35	0.96		PSO-SVR	4.14	3.53	18.47	0.95

针对 $PM_{2.5}$浓度呈现的季节性变化的特性，采用了分季节建模策略，通过建立不同季节的预测模型来提升预报准确性。实验数据分析显示，在北京市 $PM_{2.5}$浓度预测中，PSO-SVR 表现出较理想的预测精度和拟合效果，这充分证实了智能优化算法在模型参数调优方面的重要价值。实验结果表明，本文提出的 PSO-SVR 优化模型对于北京市的不同城市不同季节的 $PM_{2.5}$浓度均有较好的预测能力。

4 结论

本研究基于 2017-2020 年北京市大气污染物、气象要素及 ZTD 数据，采用 PSO-SVR、随机森林、BP 神经网络和 SVR 四种模型对各季节 $PM_{2.5}$浓度进行短期预报，获得以下主要结论：

利用 PSO 算法对 SVR 的两个关键参数进行优化建立 PSO-SVR 预测模型，并与 SVR、BP 神经网络、随机森林的预测精度对比分析，发现 PSO-SVR 模型在预测精度和稳定性方面均显著优于其他模型，能够有效捕捉 $PM_{2.5}$浓度的变化特征，尤其对浓度拐点的预测表现突出。由于区域间存在显著的空间关联性，未来在 $PM_{2.5}$浓度预测研究中，还需进一步考虑空间相关性的影响。

[参考文献]

[1]曹玉洁，王广鹤.大气细颗粒物与电子烟联合暴露毒性研究进展[J].环境与职业医学，2023，40（05）：595-600+608.

[2]Dong M，Yang D，Kuang Y et al. PM2.5 concentration prediction using hidden semi-Markov model-based times series datamining[J]. Expert Syst Appl 36：9046－9055.

[3]Elbayoumi M，Azam N，Faizah N et al. Multivariate methods for indoor PM10 and PM2.5 modelling in naturally ventilated schools buildings[J]. Atmos Environ 94：11－21.

[4]Hu S，Liu P，Qiao Y et al. PM2.5 concentration prediction based on WD-SA-LSTM-BP model：a case study of Nanjing city[J]. Environ Sci Pollut Res 29：70323－70339.

[5]Yuan XF，Huang B，Wang YL et al.Deep learning-based feature representation and its application for soft sensor modeling with variable-wise weighted SAE[J]. IEEE Trans Ind Informatics 14：3235－3243.

[6]王勇，闻德保，刘严萍，等.雾霾天气对 GPS 天顶对流层延迟与可降水量影响研究[J].大地测量与地球动力学，2014，34（02）：120-123+127.

[7]Cortes C，Vapnik V. Support-vector networks[J]. Machine Learning，1995，20（3）：273-297.

[8]彭令，牛瑞卿，吴婷. 时间序列分析与支持向量机的滑坡位移预测[J]. 浙江大学学报(工学版)，2013，47(09)：1672-1679.

[9]KENNEDY J，EBERHART R C. Particle swarm optimization[C]. International Conference on Networks，2002：1942-1948.

[10]谢劭峰，张亚博，黄良珂，等.顾及 PWV 的广西地区多尺度 PM2.5 浓度预测[J].桂林理工大学学报，2024，44（01）：90-95.

作者简介：张亚博（1998-），男，硕士研究生，研究方向：GNSS 气象学。