基于优化 LSSVM 算法的 PM2.5 浓度预测

张亚博 南守琎 唐 彦 杨云飞
新疆工程学院土木工程学院，新疆 乌鲁木齐 830023

[摘要]针对最小二乘支持向量机（LSSVM）算法中选取的核函数和正规化参数回归精度较低，该文结合粒子群优化算法（PSO）来选取最优的核参数和正规化参数，以此提高 LSSVM 模型对 $PM_{2.5}$ 质量浓度的预测精度。以 2018 年南宁市为例，选取空气主要污染物、气象因素和 GNSS 天顶对流层延迟（zenith tropospheric delay，ZTD）作为变量对同期的 $PM_{2.5}$ 浓度进行预测，并采用平均影响值（Mean Impact Value，MIV）筛选主要影响变量，实验结果显示，变量筛选后的模型对未来 48h 的 $PM_{2.5}$ 有较高预测精度，相对于 LSSVM、PSO-LSSVM 和 BP 神经网络具有更高的回归精度，表明该模型能够真实反映数据序列的内在规律，表现出了对短期预测具有较好的预测性能，具有较强的普适性。

[关键词]$PM_{2.5}$ 浓度预测；最小二乘-支持向量机；粒子群优化算法；平均影响值；优化算法
DOI：10.33142/ec.v7i9.13374 中图分类号：U491.14 文献标识码：A

$PM_{2.5}$ Concentration Prediction Based on Optimized LSSVM Algorithm

ZHANG Yabo, NAN Shoujin, TANG Yan, YANG Yunfei
School of Civil Engineering, Xinjiang Institute of Engineering, Urumqi, Xinjiang, 830023, China

Abstract: In response to the low regression accuracy of the kernel function and regularization parameters selected in the least squares support vector machine (LSSVM) algorithm, this paper combines particle swarm optimization (PSO) algorithm to select the optimal kernel parameters and regularization parameters, in order to improve the prediction accuracy of the LSSVM model for $PM_{2.5}$ mass concentration. Taking Nanning City in 2018 as an example, the main air pollutants, meteorological factors, and GNSS zenith tropospheric delay (ZTD) were selected as variables to predict the $PM_{2.5}$ concentration during the same period. The mean impact value (MIV) was used to screen the main influencing variables. The experimental results showed that the model after variable screening had high prediction accuracy for $PM_{2.5}$ in the next 48 hours, and had higher regression accuracy compared to LSSVM, PSO-LSSVM, and BP neural network, which indicates that the model can truly reflect the inherent rules of the data sequence, and has good predictive performance for short-term prediction, with strong universality.

Keywords: $PM_{2.5}$ concentration prediction; LSSVM; PSO; MIV; optimization algorithm

$PM_{2.5}$指空气动力学当量直径小于等于2.5微米的颗粒物，可以附着有害物质直接进入并粘附在人的下呼吸道和肺叶，对人体造成危害，并且有研究表明 $PM_{2.5}$ 与肺癌有直接关系[1]，因此了解 $PM_{2.5}$ 浓度的变化规律，并精准预报，能够及时提醒人们做好防护措施，对环境治理也起到重要意义。

目前，国内外的 $PM_{2.5}$ 浓度预测模型主要分为物理模型和经验模型两类。对于物理模型，如 CMAQ[2] 和 WRFChem-MADRID[3]，基于污染物的生成与传输机理，追踪和模拟污染物的变动，并提出污染物排放和空气污染之间的直接联系，然而这些物理化学模型依赖于先验知识。经验模型基于历史 $PM_{2.5}$ 数据与多个自变量的关系，挖掘出潜在的映射规律，主要包括回归模型和机器学习算法，近年来，许多机器学习模型由于能够较好的处理非线性关系而得到广泛应用，Chelani 等[4]使用支持向量机预测每日最大 O_3 浓度，与神经网络相比，具有更好的精度。朱亚杰等[5]利用空气质量监测数据和气象数据，构建支持向量机预测模型，对 $PM_{2.5}$ 浓度值进行预测，结果显示 3 日内日均值预报和 24h 内的逐小时预报都具有较高精度。随着全球导航卫星系统（Global Navigation Satellite System，GNSS）的建设发展，学者们开展了大量 GNSS 天顶对流层延迟（zenith tropospheric delay，ZTD）与雾霾监测与雾霾预报方面的研究，王勇等[6]采用北京市 GPS 连续观测网数据分析雾霾与 ZTD 的相关性，发现雾霾与 ZTD 具有良好的一致性特征。姚宜斌等[7]从时频空间分布角度进行研究，基于小波相干算法构建雾霾与 ZTD 相关性分析方法，发现 ZTD 与雾霾在一定的时域上具有很强的相关性。郭骐嘉等[8]融合 GNSS 气象参数和空气污染物建立 $PM_{2.5}$ 随机森林预测模型，在时效性 6h 范围内取得较好的预测效果。以上研究表明雾霾与 ZTD 存在较强的相关关系。因此，本文顾及大气污染物、气象参数和 ZTD 这三类影响因素，利用粒子群优化算法（Particle swarm optimization）选取最佳的 LSSVM 模型参数，构建 PSO-LSSVM 模型。为进一步分析变量对预报精度的影响，采用平均影响值（MIV）算法筛选出主要影响变量作为输入变量，构建 MIV-PSO-LSSVM 模型，将南宁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 最小二乘支持向量机

最小二乘支持向量机（Least Squares Support Vector Machines，LSSVM）是 Suykens 和 Vandewalle[9]在支持向量机（Support Vector Machines，SVM）的基础上提出的改进方法，根据结构风险不等式原则，LSSVM 将不等式约束条件改为等式约束条件，将耗时的二次规划求解问题转换成线性方程组的求解，减少了 SVM 训练时间长的缺点。对于给定的样本集（x_i, y_i），i=1，2，…，n，$x_i \in R^n$，$y_i \in R$，回归模型为[10]：

$$y(x) = w^{\mathrm{T}} \varphi(x) + c \tag{1}$$

式中，w 为权值系数；Φ(x)为非线性映射函数；c 为偏置向量。

根据结构风险最小化原理和约束条件，回归问题可以优化为：

$$\min: \frac{1}{2}\|w\|^2 + \frac{\lambda}{2}\sum_{i=1}^{n}\xi_i^2$$
$$s.t. y_i = w^{\mathrm{T}} \varphi(x_i) + c + \xi_i, (i=1,2,\cdots, n) \tag{2}$$

考虑到式（2）的等式约束，建立 Lagrange 方程，将优化问题转化为：

$$L = \frac{1}{2}\|w\|^2 + \frac{\lambda}{2}\sum_{i=1}^{n}\xi_i^2 - \sum_{i=1}^{n}\beta_i(w\varphi(x_i)+c+\xi_i - y_i) \tag{3}$$

由 KKT 优化条件求解，最终 LSSVM 的预测函数形式如下：

$$y(x) = \sum_{i=1}^{n}\beta_i k(x, x_i) + c \tag{4}$$

式中，λ 为正则化参数；ξ_i 为松弛变量；β_i 为拉格朗日因子，k(x, x_i)为核函数，核函数可以把应用在线性问题上的 LSSVM 扩展到更高维的特征空间。常见的核函数有线性、多项式以及径向基核函数，有研究表明，径向基核函数具有较强的泛化能力[11]，因此本文选择径向基函数，表达式为：

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

式中，δ 为径向基函数的核宽，用于控制函数的径向作用范围。

1.2 粒子群优化算法

粒子群优化算法是由 Kennedy 等[12]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。

由于 LSSVM 中核参数 σ 和正则参数 γ 对模型的回归精度有影响较大，因此确定一个合理的参数范围就显得至关重要，本文运用 PSO 算法对 LSSVM 模型的超参数寻优对 $PM_{2.5}$ 浓度预测总体思路如图 1 所示。

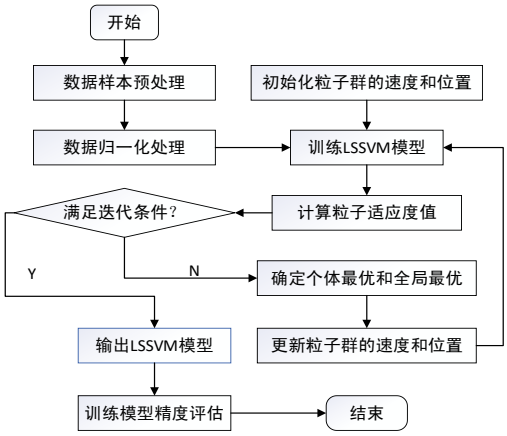

图 1 $PM_{2.5}$ 浓度预测流程

2 实验数据

2.1 研究区域

南宁位于北回归线南侧，属湿润的亚热带季风气候，阳光充足，雨量充沛，霜少无雪，气候温和，夏长冬短，年平均气温在 21.6℃左右，年均降雨量达 1304.2mm，年平均相对湿度为 79%，气候特点是炎热潮湿。相对而言，一般是夏季潮湿，而冬季稍显干燥，干湿季节分明，春秋两季气候温和，雨季主要集中在 7～9 月。

本文使用的数据包括空气主要污染物，包括 $PM_{2.5}$、PM_{10}、CO、SO_2、O_3 和 NO_2，气象因素主要包括气压（PRE）、风速（WS）、风向（WD）、温度（TEM）、相对湿度（RH）、降雨量（PRCP）和 ZTD，其中，空气污染物数据来源于广西生态环境数据中心；气象数据来源于国家气象科学数据中心；ZTD 数据来源于中国地震局 GNSS 数据产品服务平台，采用陆态网南宁站的逐小时数据。

2.2 相关性分析

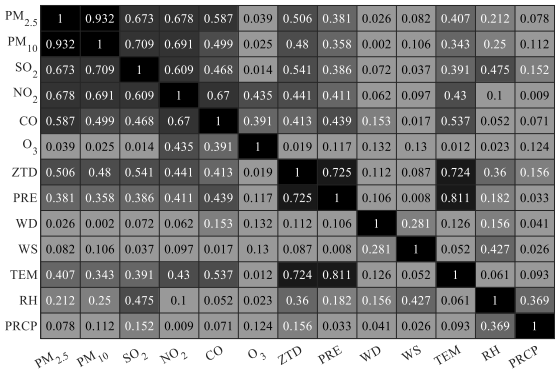

图 2 $PM_{2.5}$ 质量浓度与空气污染物、气象因素和 ZTD 的相关性系数

首先为了确定 2018 年全年各变量与 $PM_{2.5}$ 之间的相关性，采用 Spearman 秩相关系数进行相关性分析，相关系数在（0,1）区间变化，相关系数越接近 1，则表示变量之间的相关性越强，反之越接近 0 则表示变量之间无相关性。其中，相关系数在 1～0.8 之间为极强相关，相关系数在 0.8～0.6 之间为强相关，相关系数在 0.6～0.4 之间

为中等相关，相关系数在 0.4～0 之间为弱相关。从图 3 可以看出，$PM_{2.5}$ 与 PM_{10} 与为极强相关，主要是由于 $PM_{2.5}$ 与 PM_{10} 可以在一定条件下互相转化，且 PM_{10} 和 $PM_{2.5}$ 都属于颗粒物；$PM_{2.5}$ 与 SO_2、NO_2、CO 这些因素基本为强相关，$PM_{2.5}$ 与 O_3 无相关性；$PM_{2.5}$ 与 ZTD 为中等相关；对于气象因素，$PM_{2.5}$ 与温度为中等相关，与其他因素均呈弱相关。

3 实验分析

选取南宁市 2018-02 月的数据进行实验，训练模型输入的数据长度相同，分辨率为 1h，由于天气情况较难预测，可能短期之内经历多种天气状况，多变的天气对预测结果也会造成影响，因此进行 $PM_{2.5}$ 浓度值短期预测。2 月份有效数据为 626h，将前 578h 的数据作为训练集的输入，将最后 48h 的 $PM_{2.5}$ 浓度作为预测输出。通过相关性分析，选取与 $PM_{2.5}$ 浓度值显著性相关的变量，包括 PM_{10}、SO_2、NO_2 等 9 个变量。此外，并结合 MIV 算法筛选出主要影响变量，由于 LSSVM 模型训练每次选取核参数 σ 和正则参数 γ 不同，则 MIV 算法筛选变量所占的权重也不一样，因此，运行 10 次 MIV-LSSVM 算法筛选出主要影响变量，计算各变量的平均 MIV 绝对值，表 1 为对南宁市 2 月份 $PM_{2.5}$ 浓度预测模型各输入变量的 MIV 绝对值和 MIV 累计百分比。

表 1　2 月输入变量 MIV 值

排序	变量	MIV 绝对值	MIV 累计百分比	排序	变量	MIV 绝对值	MIV 累计百分比
1	PM_{10}	8.51	61.5%	6	风速	0.37	94.2%
2	相对湿度	1.93	75.4%	7	NO_2	0.32	96.5%
3	O_3	1.15	83.8%	8	风向	0.27	98.5%
4	ZTD	0.66	88.6%	9	SO_2	0.28	100%
5	CO	0.40	91.5%				

由表 1 的 MIV 绝对值可以看出，输入变量 PM_{10} 的 MIV 绝对值最大，说明该变量对模型预测影响最大，由于变量的 MIV 累计百分比不得小于 85%[14]，因此选取前 4 个变量作为输入变量。

通过 PSO 算法对 LSSVM 的核参数 σ 和正则参数 γ 两个参数迭代寻优，首先对 PSO 算法参数进行设置，取 D=2，σ 和 γ 在[0,1000]区间，种群数 M 为 40，迭代次数 k 为 100，取 c_1=1.5，c_2=1.7，惯性权重系数 ω=0.5。经过 PSO 算法得到的最优参数，模型参数寻优结果见表 2，将由 Spearman 秩相关系数筛选的输入变量所构建的预测模型命名为 PSO-LSSVM 模型，将 MIV 算法筛选的输入变量构建的预测模型命名为 MIV-PSO-LSSVM 模型。

表 2　2 月各模型超参数选取

模型参数	LSSVM	PSO-LSSVM	MIV-PSO-LSSVM
核参数 σ	5.47	4.19	4.19
正则参数 γ	99.17	55.06	55.06

模型预测结果如图 3 所示，并与 LSSVM 和 BP 神经网络的预测结果进行精度评估，表 3 为各模型的预测精度统计。为评价各模型的预测精度，采用均方根误差（RMSE）、平均绝对误差（MAE）以及平均绝对百分比误差（MAPE）来评定模型精度。

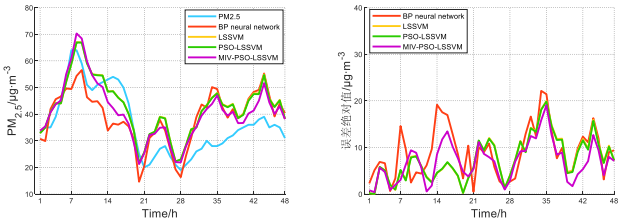

图 3　2 月各模型预测结果与误差绝对值

表 3　2 月各模型预测精度

Model	RMSE（μg/m³）	MAE（μg/m³）	MAPE（%）
BP neural network	10.45	8.93	26.16
LSSVM	8.92	7.85	24.11
PSO-LSSVM	8.81	7.49	23.86
MIV-PSO-LSSVM	8.14	6.96	21.38

结合图 3 和表 3 可知，四种模型与真实值变化有较好的一致性，PSO-LSSVM 模型与 LSSVM 模型相比，模型的回归精度提升了一些，说明 PSO 算法达到了优化超参数的作用。对于 BP 神经网络，也能够大致预测 $PM_{2.5}$ 质量浓度的走势，但在一些极值点上预测效果明显较差。MIV-PSO-LSSVM 的预测效果优于其他模型，稳定性最高，特别是对拐点处有很好的预测能力，说明 MIV 筛选的主要输入变量所构建的模型更能反映真实值，综合性能也优于其他模型。

为了进一步验证各模型对 $PM_{2.5}$ 质量浓度的预测精度，又由于 $PM_{2.5}$ 浓度值具有季节性变化规律，因此用相同方法对每个季节最后一个月进行分析，预测结果如表 4 所示。

表 4　不同月份模型预测精度

Month	Model	RMSE（μg/m³）	MAE（μg/m³）	MAPE（%）
	BP neural network	4.38	3.33	15.21
5 月	LSSVM	4.23	3.01	13.27
	PSO-LSSVM	3.02	2.26	9.96
	MIV-PSO-LSSVM	2.68	2.01	8.94
	BP neural network	3.44	2.57	23.29
8 月	LSSVM	3.69	2.68	24.15
	PSO-LSSVM	3.14	2.51	22.69
	MIV-PSO-LSSVM	2.63	2.16	20.18
	BP neural network	8.24	6.96	11.59
11 月	LSSVM	7.83	6.28	10.41
	PSO-LSSVM	6.99	5.55	9.33
	MIV-PSO-LSSVM	6.73	5.69	8.78

结合上表可以看出，4 种模型都能在一定程度上预测未来 48 小时的 $PM_{2.5}$ 质量浓度值，MIV-PSO-LSSVM 预测精度均高于其他模型，说明模型预测的误差离散型较小。

总体来看，4 种模型对 $PM_{2.5}$ 浓度的预测，在夏秋两季的精度要优于春冬两季，主要原因是颗粒物的产生还受到污染源排放的影响，即工业生产、机动车尾气排放和秸秆的燃烧等，此外，由于春季和秋季温度降低，烧煤取暖也会造成大气环境中颗粒物浓度增加，造成模型预测精度较低。

4 结论

本文结合 2018 年广西南宁市的大气污染物、气象因素和 ZTD 数据，运用 BP 神经网络、LSSVM、PSO-LSSVM 和 MIV-PSO-LSSVM 模型分别对 2018 年每个季节最后一个月（即 2 月、5 月、8 月和 11 月份）最后 2 天的 $PM_{2.5}$ 浓度值进行短期预测，得出以下结论：

（1）LSSVM 模型将复杂的非线性问题转化为求解线性方程组问题，对空气污染物有较好的预测精度，可以从历史数据中挖掘潜在规律，但是由于模型参数较难确定合理的范围，因此通过 PSO 算法对 LSSVM 两个超参数进行优化，此外，还通过 MIV 算法筛选出主要影响变量，将输入变量降低到 5 个左右，降低模型训练复杂程度，减少冗余数据。对 $PM_{2.5}$ 浓度值预测有较大影响因素主要包括 PM_{10}、相对湿度、CO、ZTD 和 O_3 等

（2）通过对 4 个模型预测结果精度评估，发现 MIV-PSO-LSSVM 预测精度高于其他模型，泛化能力更好，较大幅度的提高了预测精度和稳定性，能够有效地预测不同季节 $PM_{2.5}$ 浓度值的变化，特别是拐点的变化，使得模型构建更灵活和客观。

[参考文献]

[1]程春英,尹学博.雾霾之 $PM_{2.5}$ 的来源、成分、形成及危害[J].大学化学,2014,29(5):1-6.

[2]Mathur R,Yu S,Kang D.Assessment of the wintertime performance of developmental particulate matter forecasts with the Eta-Community Multiscale Air Quality modeling system[J].Journal of Geophysical Research Atmospheres,2008,113(2):89.

[3]Chuang M T, Zhang Y,Kang D.Application of WRF/Chem-MADRID for real-time air quality forecasting over the Southeastern United States[J].Atmospheric Environment,2011,45(34):6241-6250.

[4]Chelani A B.Prediction of daily maximum ground ozone concentration using support vector machine[J].Environmental Monitoring and Assessment,2010,162(4):169-176.

[5]朱亚杰,李琦,侯俊雄.基于支持向量回归的 PM2.5 浓度实时预报[J].测绘科学,2016,41(1):12-17.

[6]王勇,闻德保,刘严萍.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(2):120-123.

[7]姚宜斌,罗亦泳,张静影.基于小波相干的雾霾与 GNSS 对流层延迟相关性分析[J].武汉大学学报(信息科学版),2018,43(12):2131-2138.

[8]郭骐嘉,姚宜斌,周永江,等.融合 GNSS 气象参数的 PM2.5 随机森林预测模型[J].测绘科学,2021,46(4):37-42.

[9]SUYKENS J A K,Vandewalle J.Least squares support vector machine classifiers[J].Neural Processing Letters,1999,9(8):293-300.

[10]曹净,丁文云,赵党书.基于 LSSVM-ARMA 模型的基坑变形时间序列预测[J].岩土力学,2014,35(2):579-586.

[11]YAN W W,SHAO H H.Application of support vector machines and least squares support vector machines to heart disaase diagnses[J].Control and Decision,2003,18(3):356-360.

[12]KENNEDY J,EBERHART R.Particle Swarm Optimization[J].Proceedings of the IEEE International Conference on Neural Networks,1995,4(8):1942-1948.

[13]Dombi G W,Nandi P,Saxe J M.Prediction of rib fracture injury outcome by an artificial neural network [J].Journal of Trauma,1995,39(5):915.

[14]LI Z D,Han S N,Jiang J L.Antitumor compound identification from Zanthoxylum bungeanum essential oil based on composition-activity relationship[J].Chemical Research in Chinese Universities,2013,29(6):1065-1071.

作者简介：张亚博（1998.3—），毕业院校：桂林理工大学，所学专业：测绘工程，当前就职单位：新疆工程学院，职务：专任教师，职称级别：助教，研究方向：GNSS 气象学。