

模糊聚类支持向量机的区域空气 PM_{2.5} 浓度预报

李海琴¹, 杨 忠¹, 俞 杰², 史旭华¹

(1. 宁波大学 信息科学与工程学院, 浙江 宁波 315211; 2. 宁波市环境监测中心, 浙江 宁波 315012)

摘要: 在分析模糊 C 均值聚类算法与支持向量机回归的特点后, 将二者结合, 提出了模糊聚类支持向量机回归(FCM-SVR)算法, 对空气中颗粒物浓度 PM_{2.5} 进行预测. 该方法首先利用模糊 C 均值聚类算法把一个复杂的数据集分成多个群体, 再在每个群体上建立支持向量机回归(SVR)模型, 然后进行集成, 对区域空气的 PM_{2.5} 浓度进行预测. 预测结果分别与自组织竞争神经网络支持向量机回归(SOM-SVR)模型和单一的支持向量机回归(SVR)的结果进行比较. 结果表明, FCM-SVR 模型的预报准确率高于 SOM-SVR 模型和 SVR 模型.

关键词: 模糊 C 均值聚类; 支持向量机回归; PM_{2.5} 浓度预测

中图分类号: TP301.6

文献标志码: A

文章编号: 1001-5132 (2016) 04-0056-05

当今我国空气污染越来越严重, 严重影响了人们的日常生活, 危害了人体健康, 阻碍了经济可持续发展. 颗粒物浓度是灰霾发生的重要导火索, 其中细颗粒物浓度(PM_{2.5})会大大降低灰霾天气的能见度^[1], 故对空气中颗粒物浓度(PM_{2.5})的预报有着非常重要的意义.

已有很多学者对空气中颗粒物浓度(PM_{2.5})的预报展开了研究. 多元回归模型^[2]是最早应用于空气质量的统计预报模型. 随后人工神经网络^[3-6]因有着很强的容错性和非线性处理能力, 有学者将其应用于空气质量预报领域, 预报的准确率高于多元回归模型^[7]. 20 世纪 90 年代出现了支持向量机(Support Vector Machine, SVM), 支持向量机可有效地处理样本、剔除大量冗余样本, 且算法简单, 鲁棒性较好^[8]. 常涛^[9]使用支持向量机对乌鲁木齐的大气污染物浓度进行了预报; Xue 等^[10]利用支持向量机回归对 MODIS 数据进行了 PM₁₀ 浓度分析; Garcia 等^[11]利用支持向量机回归模型对西班牙北部奥维耶多城市的空气质量进行研究. 但上述模型或多或少存在因训练样本信息利用不充分而出现拟合现象, 导致预报的准确率不尽理想.

本文提出了模糊 C 均值聚类和支持向量机回

归模型(FCM-SVR)对 PM_{2.5} 浓度进行预测, 该方法先对训练样本进行聚类, 并在每一类上加上相应的气象数据, 然后分别训练 SVR 模型, 从而得到不同的 SVR 预测模型. FCM-SVR 模型充分利用了训练样本数据间的关系, 提高了对 PM_{2.5} 浓度预报的准确率.

1 聚类与支持向量机回归

1.1 SOM 聚类算法

自组织特征映射网络(Self-organizing Feature Map, SOM)也称 Kohonen 网络, 由 1981 年荷兰 Kohonen 教授提出, 该网络是一个无导师学习的神经网络^[12-13], 仅对输入变量进行学习和分类, 自动分析出输入数据之间的关系, 其模型如图 1 所示. 从图 1 可知, 该模型由 m 个神经元的输入层和 $a \times b$ 个神经元组成二维平面阵列的竞争层, 即输出层组成.

设 $X = (x_1, x_2, x_3, \dots, x_m)^T$ 为输入向量, $W_j = (\omega_{1j}, \omega_{2j}, \omega_{3j}, \dots, \omega_{mj})^T$ 为权值向量, 其中 j 表示第 j 个神经元在竞争层, 以计算各神经元的权值向量与输入向量的最小欧氏距离来选择权值向量与输入向量最为匹配的竞争层神经元, 即获得胜出神经元,

收稿日期: 2016-01-13.

宁波大学学报(理工版)网址: <http://journal.lg.nbu.edu.cn/>

基金项目: 浙江省自然科学基金(LY14F030004); 浙江省科技厅公益技术应用研究项目(2015C31017).

第一作者: 李海琴(1990-), 女, 安徽安庆人, 在读硕士研究生, 主要研究方向: 人工智能与模式识别. E-mail: lihaiqin543@163.com

*通信作者: 史旭华(1967-), 女, 浙江宁波人, 教授, 主要研究方向: 计算智能等. E-mail: shixuhua@nbu.edu.cn

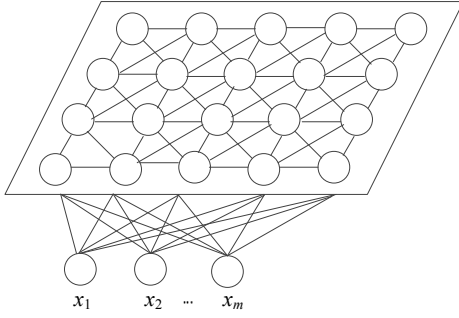


图1 SOM神经网络模型

记作 j^* , 其欧氏距离公式为:

$$d_j = \|X - W_j\| = \sqrt{\sum_{i=1}^m (x_i - \omega_{ij})^2}. \quad (1)$$

为了确定胜出神经元的邻接神经元区域的大小, 定义了领域函数, 设为高斯函数:

$$h_j = \exp(d_{jj^*}^2 / (2\sigma^2)), \quad (2)$$

式中: σ 为有效宽度, h_j 为领域, d_{jj^*} 为第 j 个神经元与胜出神经元 j^* 之间的欧式距离. 更新胜出神经元及其邻接神经元的权值向量:

$$W_j(t+1) = W_j(t) + \eta(t)h_j(t)(X - W_j(t)), \quad (3)$$

式中: t 为时间, η 为一个大于 0 小于 1 的常数, 随着时间变化逐渐下降到 0, $\eta(t) = 1/t$.

通过式(3)和对训练数据的反复运算, 对领域神经元的权值向量进行不断的调整更新, 最终计算输出. 输出公式为:

$$O_k = f(\min_j \|X - W_j\|), \quad (4)$$

式中: $f(*)$ 一般为 0~1 函数或者其他函数.

1.2 模糊 C 均值聚类算法

模糊 C 均值聚类算法(Fuzzy C-Mean, FCM)^[14-15]是从传统的硬均值方法改变而得, 用隶属度确定每个元素在某种程度上属于某个类别.

设 $X = \{X_1, X_2, \dots, X_n\}$ 是一组 n 个样本的训练数据, $x_k = \{x_{k1}, x_{k2}, \dots, x_{kr}\}$ 是第 k 个样本数据, 包含 r 个特征值. 给定类别个数为 c , 把 n 个样本数据 X 分为 c 个模糊类, 求出每个类的聚类中心, 相对应的类就用聚类中心来代表, 使模糊目标函数最小, 模糊聚类目标函数为:

$$J = \sum_{i=1}^n \sum_{j=1}^c (u_{ij})^{\bar{m}} \|X_i - v_j\|, \quad (5)$$

式中, u_{ij} 为个体 X_i 属于第 j 类的模糊隶属度, $\bar{m}$ 为模糊权重指数, v_j 为第 j 类的聚类中心. u_{ij} 和 v_j

的计算公式为:

$$u_{ij} = \begin{cases} \left(\sum_{k=1}^c \frac{\|X_i - v_j\|^{2/(m-1)}}{\|X_i - v_k\|^{2/(m-1)}} \right)^{-1}, & \|X_i - v_k\| \neq 0, \\ 1, & \|X_i - v_k\| = 0, \text{ 且 } k = j, \\ 0, & \|X_i - v_k\| = 0, \text{ 且 } k \neq j, \end{cases} \quad (6)$$

$$v_j = \sum_{i=1}^n u_{ij}^m X_i / \sum_{i=1}^n u_{ij}^m. \quad (7)$$

通过式(6)计算出模糊隶属度矩阵为:

$$u = \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1c} \\ u_{21} & u_{22} & \dots & u_{2c} \\ \vdots & \vdots & & \vdots \\ u_{n1} & u_{n2} & \dots & u_{nc} \end{bmatrix},$$

式(7)计算每类中心, 再通过式(5)计算模糊聚类目标值, 反复运算, 最终得出聚类结果.

1.3 支持向量机回归

支持向量回归机^[16](Support Vector Machine for Regression, SVR)是 Vapnik 等人在支持向量机分类的基础上发展的, 但其基本思想与支持向量机分类有所不同, 是寻找一个最优分类面使所有训练样本离该最优分类面的误差最小.

设有 s 个训练样本 $\{(x_1, y_1), (x_2, y_2), \dots, (x_s, y_s)\}$, 其中 (x_i, y_i) 是第 i 个训练样本, $x_i, y_i \in R, i = 1, 2, \dots, s$. 设回归函数为:

$$f(x) = w\Phi(x) + \bar{b}, \quad (8)$$

式中: $\Phi(x)$ 为非线性映射函数. 式(8)中的权系数向量 w 通过凸二次规划问题求得并引入松弛因子 $\xi_i \geq 0$ 和 $\xi_i^* \geq 0$ 得:

$$\min \frac{1}{2} \|w\|^2 + \bar{C} \sum_{i=1}^n (\xi_i + \xi_i^*). \quad (9)$$

约束条件为:

$$\begin{aligned} y_i - w\Phi(x) - \bar{b} &\leq \varepsilon + \xi_i, \quad i = 1, 2, \dots, n, \\ -y_i + w\Phi(x) + \bar{b} &\leq \varepsilon + \xi_i^*, \quad \xi_i \geq 0, \xi_i^* \geq 0, \end{aligned} \quad (10)$$

式中: $\bar{C}$ 是惩罚因子. 引入 Lagrange 函数将式(9)转化为如下对偶形式:

$$\begin{aligned} \max_{\partial_i, \partial_i^*} & \left[-\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (\partial_i - \partial_i^*)(\partial_j - \partial_j^*) K(x_i, x_j) - \right. \\ & \left. \sum_{i=1}^n (\partial_i + \partial_i^*) \varepsilon + \sum_{i=1}^n (\partial_i - \partial_i^*) y_i \right]. \end{aligned} \quad (11)$$

约束条件为:

$$\sum_{i=1}^n (\partial_i + \partial_i^*) = 0, \quad 0 \leq \partial_i \leq \bar{C}, \quad 0 \leq \partial_i^* \leq \bar{C}, \quad (12)$$

式中: $K(x_i, x_j) = \Phi(x_i)\Phi(x_j)$ 为核函数.

通过求解式(11)和(12), 得回归函数为:

$$f(x) = \sum_{i=1}^n (\partial_i - \partial_i^*) K(x_i, x) + b^*, \quad (13)$$

式中: ∂_i 和 ∂_i^* 为对偶问题的解, b^* 为阈值最优解.

2 基于模糊聚类 SVR 的 PM_{2.5} 浓度预测

2.1 聚类的样本选择

本研究结合宁波市某站点的各颗粒物浓度的历史小时数据和气象数据, 使用聚类和支持向量机回归结合算法预测 PM_{2.5} 的浓度. 将对 PM_{2.5} 浓度有影响的因素作为样本输入数据, 共 10 个因子, 分别为 $t-1$ 时刻的 SO₂、NO₂、PM₁₀、CO、O₃ 浓度, 以及风速、风向、气温、气压和湿度. 输出数据为 t 时刻的 PM_{2.5} 浓度.

聚类是对大量未知数据, 按数据的内在相似度将数据划分成多个子类. 聚类样本的选取对聚类的效果和预测结果都有很大的影响. 这里通过相关性分析, 从输入数据颗粒物浓度中选取与 PM_{2.5} 关系密切的数据作为聚类的样本数据, 其公式为:

$$R(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (14)$$

式中, $R(X, Y)$ 为相关系数. 当 $R(X, Y) > 0$ 时, 称变量 X 与 Y 之间是正相关, 其取值在 0~1 之间; 当 $R(X, Y) < 0$ 时, 称变量 X 与 Y 之间是负相关, 其取值在 -1~0 之间; 当 $R(X, Y) = 0$ 时, 称变量 X 与 Y 不相关. 当 R 的数值越大, 表示变量 X 与 Y 之间的关系密切程度越高.

2.2 基于模糊聚类 SVR 的 PM_{2.5} 浓度预测

2.2.1 SOM-SVR 模型预测

基于 SOM 聚类和支持向量机回归算法的基本思想是将训练样本用聚类算法进行聚类分析, 将其划分成多个子类, 然后分别在每个子类上建立支持向量机回归模型并对其进行训练, 用预测样本对其验证. 然后将训练好的 SOM 网络对预测样本进行归类, 再用相对应的 SVR 子模型对其进行

预测, 其预测模型如图 2 所示.

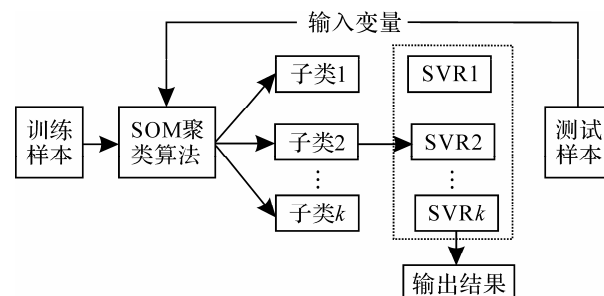


图2 SOM聚类和SVR的预测模型

2.2.2 FCM-SVR 模型预测

用 FCM 聚类算法和 SVR 相结合对 PM_{2.5} 浓度进行预测, 通过交叉验证思想将样本数据划分成训练数据和测试数据. 先通过相关性分析, 从输入变量的颗粒物浓度中选出聚类的样本数据, 并对其进行模糊聚类, 用每个训练样本的最大隶属度类别确定样本的类别, 将训练样本分成 c 个子类样本, 然后对每个子类样本数据进行归一化处理, 再分别对每个子类进行 SVR 训练, 得到 c 个 SVR 预测模型. 其归一化公式为:

$$\hat{X} = \frac{X - \min(X)}{\max(X) - \min(X)}, \quad (15)$$

其中: X 是训练样本, $\max(X)$ 、 $\min(X)$ 分别是训练样本的最大值和最小值.

用测试样本对建立好的模型进行预测验证, 同理先对预测样本进行模糊聚类将其归类, 再将其分别放入相对应的 SVR 预测模型中进行预测, 得出预测结果与实际值相比较, 验证该模型的好坏用相对误差来衡量, 其公式为:

$$\delta = \frac{1}{n_1} \sum_{i=1}^{n_1} |(y_i - y_{outi}) / y_i|, \quad (16)$$

式中: n_1 是测试样本数, y_i 对应于第 i 个数据的实际值, y_{outi} 对应于第 i 个数据的模型预测值.

FCM-SVR 预测模型算法步骤为: (1) 经过相关性分析从输入变量的颗粒物浓度中选取聚类的样本数据, 进行模糊聚类分成 c 类. (2) 每一类上加上相应的气象数据和 PM_{2.5} 的历史监测数据构成 SVR 预测模型的输入数据, 并对每一类的样本数据进行归一化处理后, 建立训练 SVR, 从而得到不同的 SVR 预测模型. (3) 最后用测试样本数据对建立好的模型进行预测并验证 SVR 模型的好坏.

2.2.3 参数选择

模糊C均值聚类算法 c 的大小对其聚类效果有一定影响,结合本文所使用的数据和经验将样本数据分成3类,即 $c=3$ 。核函数对SVR模型预测结果也有影响,本文选用高斯函数作为核函数,即:

$$K(x, x') = \exp(\|x - x'\|^2 / g^2), \quad (17)$$

式(11)中 $\bar{C}$ 和式(17)中 g^2 对SVR模型的预测精度有直接影响, $\bar{C}$ 过小会出现欠学习现象,过大会导致泛化能力下降^[17], g^2 过小会出现过学习现象,过大出现欠学习现象。据经验选 $\bar{C}=100$, $g^2=100$ 。

3 结果与分析

本研究预报数据来源于宁波市某站点2013年1月1日到6月6日的小时监测数据,其中包括颗粒物浓度数据和气象数据,共3500组。前3000组数据为训练样本,后500组数据为测试样本。

通过式(16)计算PM_{2.5}浓度与其他颗粒物浓度的相关性,确定聚类的样本数据,结果见表1。

表1 PM_{2.5}浓度与其他颗粒物浓度的相关性

相关性	SO ₂	NO ₂	PM ₁₀	CO	O ₃
PM _{2.5}	0.741 5	0.700 2	0.796 4	0.683 1	-0.371 2

由表1可知,SO₂、NO₂、PM₁₀和CO与PM_{2.5}浓度呈正相关,相关性较高;O₃与PM_{2.5}浓度呈负相关,相关性较低,故聚类的样本数据为SO₂、NO₂、PM₁₀和CO浓度。

使用模糊C均值聚类和支持向量机回归结合的预测模型对PM_{2.5}浓度进行预报,并分别与SOM聚类和支持向量机回归结合预测模型及单一支持向量机回归预测模型做比较,结果如图3~5所示。图中横坐标为2013年5月15日至6月6日的测试数据序列,共500组,纵坐标为PM_{2.5}浓度,曲线是实际值与预测值的仿真结果。

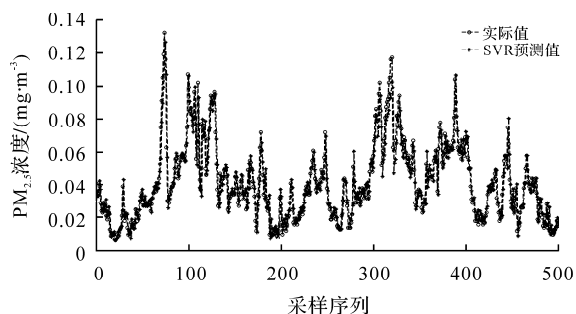


图3 单一SVR预报结果

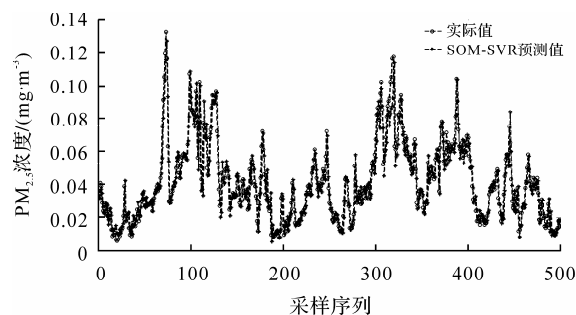


图4 SOM-SVR预报结果

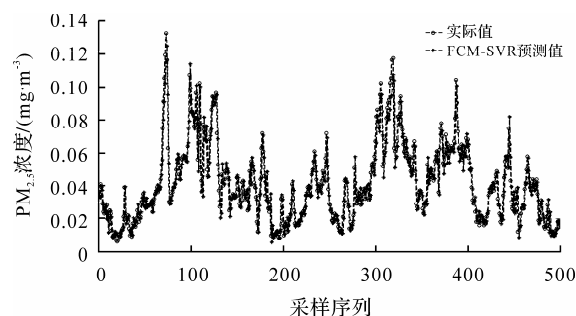


图5 FCM-SVR预报结果

为更好地反映出各预测模型的仿真效果,利用式(14)和(16)来衡量其预测效果,结果见表2。

从图3~5和表2可知,使用SOM-SVR模型的预报精度要高于单一SVR模型的预报,而FCM-SVR预报精度高于前两个模型。

表2 实际值与预测值的衡量指标

指标	SVR	SOM-SVR	FCM-SVR
相对误差/%	19.226 7	18.105 5	17.623 4
相关系数	0.917 3	0.917 9	0.922 7

4 结论

使用模糊C均值聚类算法和支持向量机回归相结合的方法对PM_{2.5}浓度进行预报,算法考虑了PM_{2.5}的各种影响因素,通过相关性计算选出聚类样本,对训练样本进行分类,然后在每一类上加入相对应的气象因素作为预测的输入变量,归一化数据后建立SVR训练,得到不同的SVR预报模型。仿真实验结果证明了该预报方法的有效性。

参考文献:

- [1] 杨洪斌, 邹旭东, 汪宏宇, 等. 大气环境中PM_{2.5}的研究进展与展望[J]. 气象与环境学报, 2012, 28(3): 77-82.
- [2] 张云海, 孙财涛, 杨洪斌. 基于逐步回归分析方法的PM₁₀浓度预报模型[J]. 环境科学与技术, 2009(5): 100-

- 102.
- [3] GEOFFREY C W. An enhanced $PM_{2.5}$ air quality forecast model based on nonlinear regression and back-trajectory concentrations[J]. *Atmospheric Environment*, 2010, 44(25): 3015-3023.
- [4] JEE H, CLEMENS M, GERWIN D, et al. Artificial neural network forecast for daily average PM_{10} hourly concentrations, in the greater area of Athens Greece[J]. *Atmospheric Environment*, 2006, 40(7):1216-1229.
- [5] IRANI R, NASIMI R. A short-term load forecasting model of natural gas based on optimized genetic algorithm and improved BP neural network[J]. *Expert Systems with Applications*, 2011, 38(8):9862-9866.
- [6] YUAN T. Prediction of free lime content in cement clinker based on RBF neural network[J]. *Journal of Wuhan University of Technology (Materials Science Edition)*, 2012(1):187-190.
- [7] 杨树平, 冯晖, 黄晓. B-P 神经网络在环境空气质量预报中的应用[J]. *云南环境科学*, 2003(S1):52-54.
- [8] 吴义国. 支持向量机研究及其在文本分类中的应用[D]. 广州: 广东工业大学, 2011.
- [9] 常涛. 支持向量机在乌鲁木齐大气污染预报中的应用研究[EB/OL]. [2015-11-16]. <http://www.doc88.com/p-6941561741753.html>.
- [10] XUE Y S, WU Y, YU L, et al. Support vector machine regression applied to MODIS data for PM_{10} concentration analysis[EB/OL]. [2015-11-16]. [http://xueshu.baidu.com/s?wd=paperuri:\(956bf799f9665484278f3c2fbc cf718b\)&filter=sc_long_sign&sc_ks_para=q%3DSupport+vector+machine+regression+applied+to+MODIS+data+for+PM10+concentration+analysis&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_us=15211614022895719430](http://xueshu.baidu.com/s?wd=paperuri:(956bf799f9665484278f3c2fbc cf718b)&filter=sc_long_sign&sc_ks_para=q%3DSupport+vector+machine+regression+applied+to+MODIS+data+for+PM10+concentration+analysis&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8&sc_us=15211614022895719430).
- [11] GARCÍA N P J, COMBARRO E F, DEL C D, et al. A SVM-based regression model to study the air quality at local scale in Oviedo urban area (Northern Spain): A case study[J]. *Applied Mathematics and Computation*, 2013, 21917:8923-8937.
- [12] RUIZ-AGUILAR J J, TURIAS I J, JIMÉNEZ-COME M J. A two-stage procedure for forecasting freight inspections at Border Inspection Posts using SOMs and support vector regression[J]. *International Journal of Production Research*, 2015, 537:2119-2130.
- [13] 孙强. 基于人工神经网络的汽车声品质评价与应用研究[D]. 长春: 吉林大学, 2010.
- [14] BALAFAR M A. Fuzzy C-mean based brain MRI segmentation algorithms[J]. *Artificial Intelligence Review*, 2014, 413:441-449.
- [15] 尤加春, 毛慧慧. 基于模糊 C 均值聚类—支持向量机的海底沉积物分类识别[J]. *海洋科学*, 2014(11):122-130.
- [16] 张学工. 关于统计学习理论的本质[J]. *自动化学报*, 2000(1):32-42.
- [17] 向昌盛, 袁哲明, 周子英. 混沌时间序列预测模型参数的联合优化[J]. *信息与控制*, 2011(5):673-679; 691.

Fuzzy Clustering Support Vector Machine for Predicting Regional $PM_{2.5}$ Concentration

LI Hai-qin¹, YANG Zhong¹, YU Jie², SHI Xu-hua^{1*}

(1. Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, China;

2. Ningbo Municipal Environmental Monitoring Center, Ningbo 315012, China).

Abstract: A fuzzy clustering support vector machine regression algorithm is proposed by analyzing and combining the characteristics of the fuzzy C-mean clustering algorithm and the support vector machine regression. The SVM is designed to forecast the particles density $PM_{2.5}$ in the air. Firstly, a complex data set is separated and inserted into multiple groups using fuzzy C-mean clustering algorithm. Then the SVM regression model in each group is established. The integrated fuzzy clustering SVM regression is applied to forecast the $PM_{2.5}$ in the local air. By comparing the predicted result with that of the self-organizing competitive neural network SVM regression model, as well as that of the single SVM regression model respectively, it is found that the predicted accuracy rate of the FCM-SVR is higher than that of the SOM-SVR model and SVR model.

Key words: fuzzy C-mean clustering; support vector machine regression; prediction of $PM_{2.5}$ concentration

(责任编辑 史小丽)