

DOI:10.13232/j.cnki.jnju.2020.01.016

基于社区划分的空气质量指数(AQI)预测算法

袁 燕, 陈伯伦*, 朱国畅, 花 勇, 于永涛

(淮阴工学院计算机与软件工程学院, 淮安, 223003)

摘 要: AQI (Air Quality Index) 是判定空气质量好坏的重要指标, 做好 AQI 的预测, 对大气污染的治理有积极的推进作用, 但目前预测 AQI 的算法通常没有综合考虑气象因素和周边城市对预测性能的影响. 将气象因素和周边城市的污染物因素作为算法设计的基础, 提出一种基于社区划分的空气质量指数预测的算法. 首先根据气象特征计算城市之间的相似度, 接着对各城市间的相似度矩阵进行社区划分; 然后将属于同一社区的城市污染物时序信息作为预测目标城市空气质量指数的依据, 并考虑目标城市的周边城市对其的影响; 最后使用非线性回归的方法进行预测建模. 通过对江苏省内 20 座城市的大气污染数据和气象数据的采集与分析, 证明该算法不但预测精度有所提高, 而且与传统的时间序列预测模型相比, 降低了时间复杂度.

关键词: 空气质量指数(AQI)预测, 气象因素, 时序信息, 社区划分

中图分类号: TP301.6

文献标识码: A

Prediction of Air Quality Index (AQI) based on community division

Yuan Yan, Chen Bolun*, Zhu Guochang, Hua Yong, Yu Yongtao

(College of Computer and Software Engineering, Huaiyin Institute of Technology, Huai'an, 223003, China)

Abstract: AQI (Air Quality Index) is an important indicator to judge the air quality. Effectively predicting the AQI has positive impact on the control of air pollution. However, the existing AQI prediction methods scarcely consider the weather factors and the influence on the prediction performance of the surrounding cities. In this paper, we propose a community division based AQI prediction method by considering the weather factors and the pollutant factors of the surrounding cities. Firstly, the similarity between cities is computed according to the weather factors. Then, community division is performed on the similarity matrices of each pair of cities. Next, by considering the impact of the surrounding cities of the target city, the city pollutant time series information belonging to the same community is treated as the basis for predicting the AQI of the target city. Finally, nonlinear regression is conducted for predictive modelling. Through the collection and analysis on the air pollution data and weather data of 20 cities in Jiangsu Province, it demonstrates that the proposed method improves greatly in prediction accuracy and performs computationally effectively compared with the traditional time series based prediction models.

Key words: Air Quality Index (AQI), weather factors, time series information, community division

近年来,随着中国社会和经济的快速发展,人民的生活水平逐渐提高,但随之而来的生态环境

问题却影响着人们的生活质量. 由于不合理的开发和利用,空气质量不断恶化,我国的生态环境面

基金项目: 国家自然科学基金(61602202), 国家重点研发项目(2018YFB1004904), 江苏省自然科学基金(BK20160428), 江苏省六大人才高峰项目(XYDXX-034)

收稿日期: 2019-08-13

* 通讯联系人, E-mail: chenbolun1986@163.com

临严重威胁. 大气污染问题已经严重影响社会的可持续发展,对广大居民的生命健康也造成极大的威胁^[1-3]. 空气质量指数(Air Quality Index, AQI)^[4-6]是国际上普遍采用的判定空气质量的重要指标,AQI越高,空气污染越严重. AQI的预测可以及时向政府提供大气环境质量的变化趋势,也可用于对大气污染的控制和管理. 因此,AQI的预测研究十分重要.

AQI的预测已成为研究热点,越来越多的预测方法被提出,预测准确度也在不断地提高. 例如,Carbajal-Hernández et al^[7]设计了一种新的空气质量评估计算模型来评估可能对城市地区敏感人群造成伤害的有毒化合物,在建模过程中使用igma运算符统计空气质量参数,确定其对空气质量的负面影响. Corani and Scanagatta^[8]设计了一种基于贝叶斯网络的多标签分类器对类变量的依赖性进行建模,使用贝叶斯网络评估空气污染物克服某个阈值的概率,提高了预测的准确度. Singh et al^[9]进行了线性和非线性建模,其建立的五种建模分别是偏最小二乘回归、多元多项式回归、多层感知器网络、径向基函数网络和广义回归神经网络,使用统计标准参数比较五种不同模型的泛化和预测能力. 李博群等^[10]通过改进模糊时间序列模型对AQI进行预测,实验表明该思想比传统的时间序列ARIMA(Autoregressive Integrated Moving Average Model)模型的准确率更高. 高帅等^[11]提出一种改进的思维进化算法,借鉴遗传算法和粒子群算法的优点,优化了思维进化算法的随机性,并通过优化神经网络的初始权值和阈值对AQI进行预测. 刘洪通等^[12]提出一种基于Storm的空气质量指数的实时预测模型S-OKNN(Storm k -Nearest Neighbor),通过对KNN(k -Nearest Neighbor)算法进行分布式拓展,并且利用Storm的实时流数据计算特点,实现对AQI的实时预测. 张春露和白艳萍^[13]提出一种基于TensorFlow的LSTM(Long Short-Term Memory)模型,可以利用时序数据中远距离依赖信息的能力,对AQI进行预测. 白鹤鸣等^[14]提出一种基于BP神经网络算法的新模型,利用近十年的北京市地面气象观测资料和空气污染指数数据,构建不同季节的空气污染指数预测模型,对北京市

空气污染指数进行了预测. Li et al^[15]提出一种新的基于时空深度学习的空气质量预测方法,主要考虑空间和时间的相关性,通过贪婪的分层方式进行训练,与传统的时间序列预测模型相比,降低了时间复杂度. Xi et al^[16]通过机器学习来改进预测空气污染的方法,充分利用WRF-Chem(Weather Research and Forecasting-Chemistry)模型对污染物、化学成分的预测,设计综合评价框架,提高预测性能.

虽然预测AQI的算法越来越多,但大部分算法没有综合考虑目标城市的周边城市的气象因素和污染物因素对计算的影响. 由于气象因素对AQI的影响是不可或缺的^[17-19],因此本文把气象因素考虑进预测模型中,提出一种基于社区划分的空气质量指数预测算法CK-UNR(Cosine K-means-Unlinear Regression). 该算法选取气象因素中的气压、温度、湿度和两分钟平均风速四个因素作为研究对象,通过社区划分综合考虑气象因素和污染物因素对AQI计算的影响. 实验证明,该算法不但降低了时间复杂度,预测性能也得到了提高.

1 相关理论

1.1 空气质量指数(AQI) AQI^[20-21]是评价每日空气质量的指标. 表1为空气质量分指数及对应的污染物项目浓度限值,根据表1及AQI计算公式分别计算细颗粒物(PM_{2.5})、可吸入颗粒物(PM₁₀)、二氧化硫(SO₂)、二氧化氮(NO₂)、臭氧(O₃)、一氧化碳(CO)等各项污染物的空气质量分指数,最后选取最大值作为AQI的最终值.

污染物因素A的空气质量分指数如式(1)所示:

$$IAQI_A = \frac{IAQI_{Hi} - IAQI_{Lo}}{BP_{Hi} - BP_{Lo}} (C_A - BP_{Lo}) + IAQI_{Lo} \quad (1)$$

其中, $IAQI_A$ 表示污染物因素A的空气质量分指数, C_A 表示污染物因素A的质量浓度, BP_{Hi} 表示表1中与 C_A 相近的污染物浓度限值的高位值, BP_{Lo} 表示表1中与 C_A 相近的污染物浓度限值的低位值, $IAQI_{Hi}$ 表示表1中与 BP_{Hi} 对应的空气质量分指数, $IAQI_{Lo}$ 表示表1中与 BP_{Lo} 对应的空气

表 1 空气质量分指数及对应的污染物项目浓度限值

Table 1 Air quality sub-index and corresponding pollutant concentration limit

IAQI	Contaminant project concentration limit					
	SO ₂ (μg·m ⁻¹)	NO ₂ (μg·m ⁻¹)	PM ₁₀ (μg·m ⁻¹)	CO (μg·m ⁻¹)	O ₃ (μg·m ⁻¹)	PM _{2.5} (μg·m ⁻¹)
0	0	0	0	0	0	0
50	50	40	50	2	100	35
100	150	80	150	4	160	75
150	475	180	250	15	215	115
200	800	280	350	24	265	150
300	1600	565	420	36	800	250
400	2100	750	500	48		350
500	2620	940	600	60		500

质量分指数. 最终 AQI 的计算过程如式(2)所示:

$$AQI = \max\{IAQI_A\} \quad (2)$$

1.2 符号介绍 见表 2.

1.3 模块度 本文通过社区划分来进行模型的设计, 通常通过模块度来衡量社区结构的好坏^[22-23]. 模块度是网络或图形结构的一种衡量标准, 旨在衡量网络划分为模块的强度. 高模块性的网络在社区内的节点之间有紧密连接, 但在不同社区中的节点间有稀疏连接. 模块度的值越大表明网络中的社区结构越明显. 实际网络中, 模块度的值在 0.3 到 0.7 之间的社区结构较好.

假设有 n 个城市, 城市与城市间互相连接, 因此 n 个城市构成了一个网络. 通过社区划分, 把网络划分成不同的社区. 定义一个 q 行 p 列的矩阵 $C, q, p \in n, C_{mr} = 1$ 表示第 m 个城市属于第 r 个社区, 则模块度 Q 的矩阵形式如式(3)所示:

$$Q = \frac{1}{2w} \text{Tr}(C^T DC) \quad (3)$$

D 的公式为:

$$D_{mn} = A_{mn} - \frac{k_m k_n}{2w} \quad (4)$$

其中, w 表示整个网络的边数, Tr 指秩, A_{mn} 是该网络的邻接矩阵, k_m 和 k_n 表示城市 m 和城市 n 的度.

2 算法描述

2.1 算法的基本思想 通过一系列的研究发现, AQI 除了与污染物的排放量相关, 还与气象因素中的气压、风速、温度等因素相关. 因此本文将气

表 2 符号介绍

Table 2 Symbol introduction

符号	描述
$IAQI_A$	污染物因素 A 的空气质量分指数
C_A	污染物因素 A 的质量浓度值
BP_{Hi}	表 1 中与 C_A 相近的污染物浓度限值的高位值
BP_{Lo}	表 1 中与 C_A 相近的污染物浓度限值的低位值
$IAQI_{Hi}$	表 1 中与 BP_{Hi} 对应的空气质量分指数
$IAQI_{Lo}$	表 1 中与 BP_{Lo} 对应的空气质量分指数
u_i	样本中编号为 i 的城市
d_t	样本的天数 t
x_t	第 t 天的气象因素 X 的值
y_t	第 t 天的污染因素 Y 的值
$B'_i(X)$	江苏省第 i 个城市第 t 天第 X 个气象因素的值
$A'_i(Y)$	江苏省第 i 个城市第 t 天第 Y 个污染物因素的值
AQI'_i	预测城市 i 第 t 天 AQI 的值
$\text{sim}(B'_m, B'_n)$	任意两个城市 m 和 n 间气象因素的相似度矩阵
k	表示第 k 种气象因素
B'_i	表示气象因素数据矩阵
A'_i	表示污染物因素数据矩阵
$A_h(Y)$	表示目标城市 h 关于污染物因素 Y 的矩阵
$SA(m, n)$	城市 m 和 n 间的相似度矩阵

象因素考虑到模型的设计中,提出CK-UNR算法.选取四种气象因素作为研究对象,分别是气压、两分钟平均风速、温度、相对湿度.通过节点相似性算法计算得出各城市间关于气象因素的相似度矩阵,然后对相似度矩阵进行社区划分,将与目标城市相似性较高的城市化为一个社区,最后通过非线性回归算法预测空气质量指数的值.该算法不仅考虑气象因素对目标城市空气质量指数的影响,还综合考虑周边城市对目标城市的影响.

为了进行算法模型的设计,针对原始网络,可以使用如下的数据结构进行样本数据的存储.

定义 $B'_i(X) = \{u_i, d_i, x_i\}$ 为城市的气象因素样本数据网络,其中 u_i 为样本中编号为 i 的城市, d_i 为样本的天数 t , x_i 为第 t 天的气象因素 X 的值.假设第 i 个城市 B_i 的气象因素数据包括 n 天 m 个气象因素,则 B_i 的气象因素构成一个 $n \times m$ 的矩阵 $B_i = [x_i]$.图1为气象因素数据进行建模的示意图.

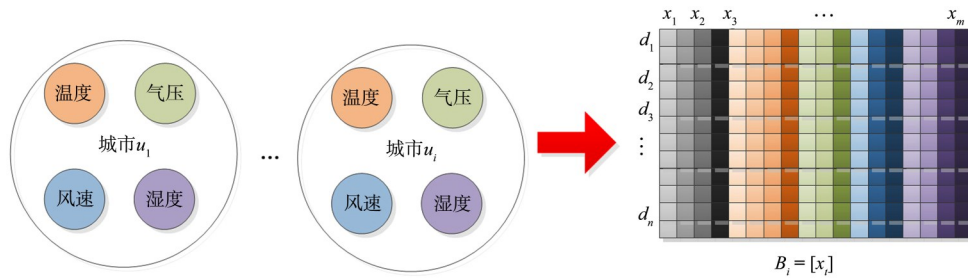


图1 关于气象因素的数学模型

Fig. 1 Mathematical model of meteorological factors

定义 $A'_i = \{u_i, d_i, y_i\}$ 为城市的污染因素样本数据网络,其中 u_i 为样本中编号为 i 的城市, d_i 为样本的天数 t , y_i 为第 t 天的污染因素 Y 的值.假

设第 i 个城市 A_i 的数据包括 h 天, l 个污染因素,则 A_i 的污染物因素构成了一个 $h \times l$ 的矩阵 $A_i = [y_i]$.图2显示了污染因素数据集合的示例.

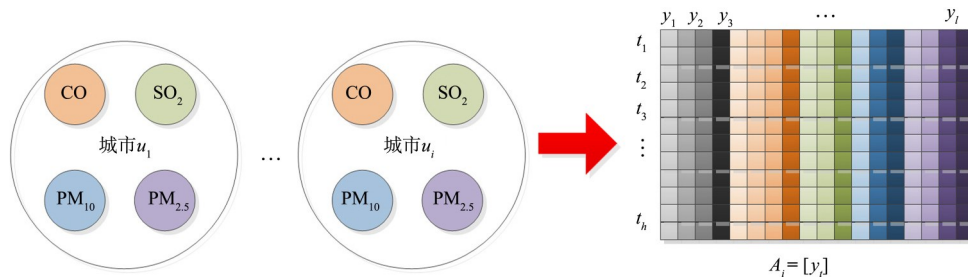


图2 关于污染因素的数学模型

Fig. 2 Mathematical model of pollution factors

2.2 CK-UNR的算法步骤

输入:多个城市前 t 天的气象因素数据集 B_i 和污染物因素数据集 A_i .

输出:目标城市第 $t+1$ 天 AQI 的值 AQI_h^{t+1} .

①对数据集 B_i 和 A_i 进行归一化,得到气象因素数据矩阵 B'_i 和污染物因素数据集 A'_i .

②通过相似度算法,计算任意两个城市 m 和 n 间气象因素的相似度矩阵 $sim(B'_m, B'_n)$.

③考虑时间因素,给相似度矩阵一个系数

ρ^{t-t+1} ,形成任意两个城市间的相似度矩阵 SA .

④对矩阵 SA 进行社区划分,选取模块度最大值 Q^* 对应的社区划分.

⑤通过社区划分,结合其他城市和目标城市的相似度,得出关于目标城市的矩阵 $A_h(Y)$.

⑥基于非线性回归方程进行模型的构建,得到 Y 个污染物的预测结果 C_Y .

⑦通过公式计算 AQI_h^{t+1} .

2.3 步骤详情描述

2.3.1 对数据集 B_i 和 A_i 进行归一化, 得到气象因素数据矩阵 B'_i 和污染物因素数据集 A'_i .

从网上采集的数据中, 将城市 i 第 t 天第 x 个气象因素的值用 $B'_i(X_x)$ 来定义, 城市 i 第 t 天第 y 个污染物因素的值用 $A'_i(Y_y)$ 来定义. 由于数据的不完整性, 需要对数据进行预处理. 在采集的气象因素数据中, 有一部分气象因素在短期内几乎不发生变化, 据此提取如下四个气象因素, 分别是气压、两分钟平均风速、温度、相对湿度. 计算空气质量指数(AQI)只需要一氧化碳(CO)、二氧化氮(NO_2)、臭氧(O_3)、 PM_{10} 、 $\text{PM}_{2.5}$ 、二氧化硫(SO_2)这六种污染物因素.

预处理后的气象因素数据集表示为:

$$B_i = [x_i] = \begin{bmatrix} B_i^1(X_1) & \cdots & B_i^1(X_4) \\ B_i^2(X_1) & \cdots & B_i^2(X_4) \\ B_i^t(X_1) & \cdots & B_i^t(X_4) \end{bmatrix}$$

污染物因素数据集表示为:

$$A_i = [y_i] = \begin{bmatrix} A_i^1(Y_1) & \cdots & A_i^1(Y_6) \\ A_i^2(Y_1) & \cdots & A_i^2(Y_6) \\ A_i^t(Y_1) & \cdots & A_i^t(Y_6) \end{bmatrix}$$

由于数据的评价尺度不一样, 使用离差归一化对原始数据样本进行处理, 函数如下:

$$x_i^* = \frac{x_i - \min}{\max - \min} \quad (5)$$

其中, x_i 为气象因素数据集, \max 为数据中的最大值, 而 \min 是数据中的最小值.

2.3.2 通过相似度算法, 计算任意两个城市 m

和 n 间气象因素的相似度矩阵 $\text{sim}(B'_m, B'_n)$.

气象因素对空气质量指数的计算存在一定的影响, 通过两个城市的气象因素计算两个城市间的相似度, 可以变相地了解两个城市间空气质量指数的相似度. 如果两个城市的气象因素相似, 那这两个城市间的空气质量指数也存在一定的相似. 式(6)是通过余弦相似性计算任意两个城市 m 和 n 关于气象因素的相似度:

$$\text{sim}(B'_m, B'_n) = \frac{B'_m \cdot B'_n}{\|B'_m\| \times \|B'_n\|} = \frac{\sum_{k=1}^z (B'_m(X_k) \times B'_n(X_k))}{\sqrt{\sum_{k=1}^z (B'_m(X_k))^2} \times \sqrt{\sum_{k=1}^z (B'_n(X_k))^2}} \quad (6)$$

其中, $B'_m(X_k)$ 表示城市 m 在不同时间 t 第 k 种气象因素的值, $B'_n(X_k)$ 表示城市 n 在不同时间 t 第 k 种气象因素的值.

2.3.3 考虑时间因素, 给相似度矩阵一个系数 ρ^{t-l+1} , 形成任意两个城市间的相似度矩阵 SA .

本算法通过其他城市前 t 天的污染物因素和气象因素来预测第 $t+1$ 天目标城市的 AQI. 为了使预测的结果更准确, 需要考虑时间因素. 因此在计算任意两个城市间的相似度时, 给每个城市每天的气象因素设置一个系数 ρ^{t-l+1} , 其中 t 表示一个常数, $l \in t$. 通过这样一个系数, 使得越靠近第 $t+1$ 天的气象因素比例越重, 可以更加精确地确定两个城市的相似度. 如式(7)所示, 通过式(6)求得任意两个城市关于天气因素的相似度, 然后形成任意两个城市间的相似度矩阵 SA :

$$SA(m, n) = \frac{\rho^t \text{sim}(B'_m, B'_n) + \rho^{t-1} \text{sim}(B'_m, B'_n) + \cdots + \rho^1 \text{sim}(B'_m, B'_n)}{t} = \frac{\sum_{l=1}^t \rho^{t-l+1} \text{sim}(B'_m, B'_n)}{t} \quad (7)$$

2.3.4 对相似度矩阵 SA 进行社区划分, 选取模块度最大值 Q^* 对应的社区划分.

本算法中, 使用 k -means 算法对多个城市形成的网络进行社区划分, 通过模块度检测算法划分的社区结构. 模块度是检测社区划分好坏的准则, 模块度的值最大, 说明此时的社区划分效果最佳, 以此来确定 k 值. 以图 3 为例, 当 k 值为 2 时所有的城市被划分为两个社区. 可以看出, 城市 $u_1, u_2, u_3, u_4, u_5, u_6, u_7$ 和 u_i 分别带有属性温度、气压、

风速和湿度. 通过社区划分, 城市 u_1, u_4, u_5 和 u_6 被划分在一个社区中, 城市 u_2, u_3, u_7 和 u_i 被划分在一个社区.

2.3.5 通过社区划分, 结合其他城市和目标城市的相似度, 得出关于目标城市的矩阵 $A'_h(Y)$.

社区划分后, 有 z 个城市和目标城市在同一个社区. 同一个社区的城市之间存在一定的相互影响, 因此结合其他城市和目标城市的相似度, 得到关于目标城市的矩阵 $A'_h(Y)$, 如式(8)所示:

$$\frac{A_h(Y) + \sum_{g=1}^z (SA(h, j_g) A_h(Y) + \dots + SA(h, j_z) A_h(Y))}{z+1} = A'_h(Y) \quad (8)$$

其中, h 表示目标城市, $A_h(Y)$ 表示目标城市 h 关于污染物因素的矩阵, j_g 表示跟目标城市在同一个社区内的 z 个城市, $SA(h, j_g)$ 表示目标城市 h 和城市 j_g 之间的相似度.

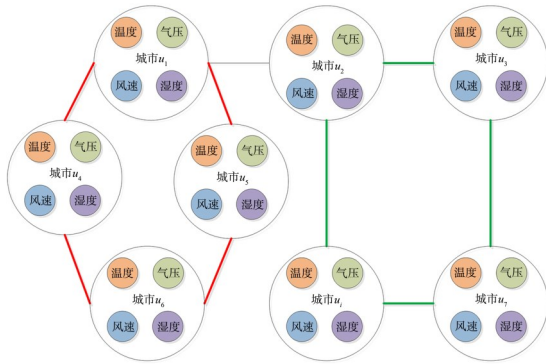


图3 社区划分的示例

Fig. 3 An example of community division

2.3.6 基于非线性回归方程进行模型的构建, 得到 Y 个污染物预测结果 C_Y .

通过绘制污染物因素的离散点图, 确定污染物因素跟时间之间是非线性的关系, 通过图的走势, 选择下列函数进行拟合:

$$y = a \sin(bt + c) + d \quad (9)$$

其中, t 表示时间, y 表示矩阵 $A'_h(Y)$ 中的值, $A'_h(Y)$ 矩阵中的每一行表示一种污染因素在不同时间的值. a, b, c 和 d 是该曲线在拟合过程中的参数. 通过非线性回归方程分别预测出 Y 个污染物因素在 $t+1$ 的值, 用 C_Y 表示.

2.3.7 通过公式计算 AQI_h^{t+1} .

把 C_Y 代入式(1), C_Y 表示 Y 个污染物因素的质量浓度值, 分别求得 Y 个污染物因素的空气质量分指数 $IAQI_Y$, 最终选取最大值作为目标城市在 $t+1$ 的空气质量指数 AQI_h^{t+1} .

3 实验结果及分析

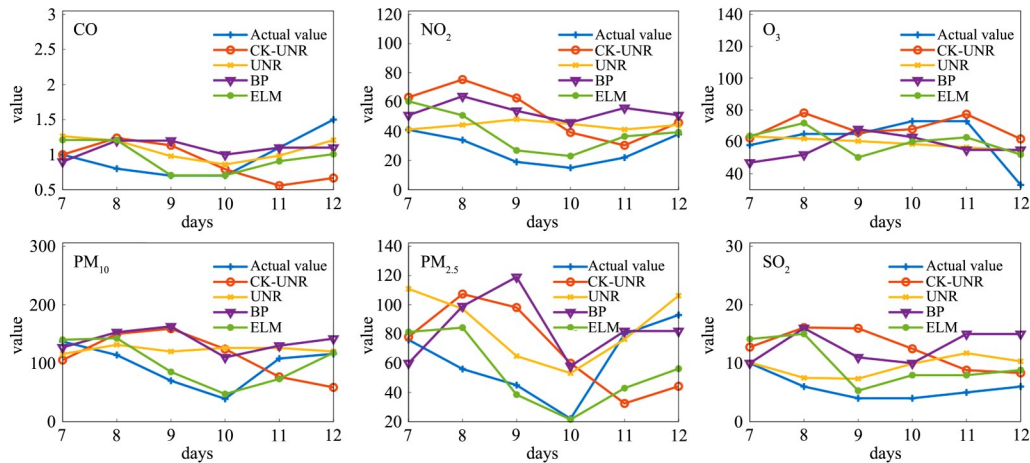
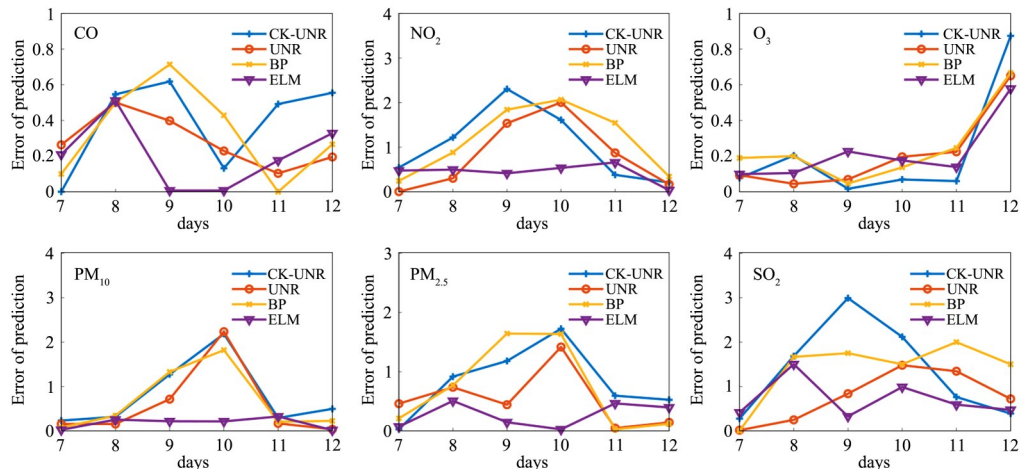
实验所用空气质量指数及污染物数据来源于中华人民共和国生态环境部信息中心 (<http://datacenter.mee.gov.cn>), 气象数据来自中国气象数据网 (<http://data.cma.cn/>). 污染物数据包括

影响空气质量指数的污染源, 选取的是日平均数据, 即一氧化碳(CO)、二氧化氮(NO₂)、臭氧(O₃)、PM₁₀、PM_{2.5}、二氧化硫(SO₂)日平均浓度; 气象数据包括气压、两分钟平均速度、温度、相对湿度, 选取的是每小时数据, 并对其处理得出日平均数据. 选取江苏省内20个城市的相关数据, 淮安是目标城市.

3.1 污染物因素预测结果 为了验证本文提出的CK-UNR算法在预测方面的准确性, 和三种算法进行了比较, 分别是非线性回归(Unlinear Regression, UNR)算法、BP(Back Propagation)神经网络算法和极限学习机(Extreme Learning Machine, ELM)算法.

图4是CO、NO₂、O₃、PM₁₀、PM_{2.5}和SO₂六种污染物因素在未来六天的预测值. 横坐标表示天数, 从第7天开始到第12天结束; 纵坐标表示不同算法预测的CO、NO₂、O₃、PM₁₀、PM_{2.5}和SO₂的值. 从图4可以看出, 四种算法的预测值和真实值虽然都存在误差, 但是在CO和PM_{2.5}的预测结果中, CK-UNR算法预测的结果与真实值最接近. 在NO₂的预测结果中, UNR算法预测的结果与真实值最接近. 在O₃的预测结果中, CK-UNR算法、UNR算法及ELM算法的预测值与真实值都很接近, 而BP算法的预测值与真实值相差较大. 在PM₁₀的预测结果中, ELM算法预测的结果与真实值最接近, BP算法的预测值其次, CK-UNR算法的预测值最差. 在PM_{2.5}的预测结果中, CK-UNR算法预测的结果与真实值最接近, ELM算法的预测值其次, UNR算法和BP算法的精确度都比较差. 在SO₂的预测结果中, CK-UNR算法和ELM算法的预测结果较差, UNR算法和BP算法的预测值与真实值很接近. 可以看出, 各算法的性能各有好坏, 但总的来说, CK-UNR算法的预测值与实际值之间的误差较小, 预测精确度比UNR算法要好, 因此在预测模型中考虑气象因素是十分必要的.

图5是CO、NO₂、O₃、PM₁₀、PM_{2.5}和SO₂的预测值在六天内与实际值之间的误差. 横坐标表示

图 4 CO, NO₂, O₃, PM₁₀, PM_{2.5} 和 SO₂ 在未来六天的预测值Fig. 4 The predicted values of CO, NO₂, O₃, PM₁₀, PM_{2.5} and SO₂ in the next six days图 5 CO, NO₂, O₃, PM₁₀, PM_{2.5} 和 SO₂ 的预测值在六天内跟实际值的误差Fig. 5 The errors of the predicted values of CO, NO₂, O₃, PM₁₀, PM_{2.5} and SO₂ in six days with the actual values

天数,从第7天开始到第12天结束;纵坐标表示各算法预测的CO, NO₂, O₃, PM₁₀, PM_{2.5}和SO₂的预测值与实际值之间的误差.可以看出,在第七天,各算法的预测值跟实际值间的误差最小. CK-UNR算法对CO和PM_{2.5}的预测值与真实值最接近, UNR算法对NO₂和O₃的预测值与真实值最接近, ELM算法对PM₁₀的预测值与真实值最接近,而BP算法的预测效果较差.

3.2 AQI预测结果展示 接下来根据表1以及各算法预测的第7天各污染物的值,计算第7天AQI的值.从表1中可以看出,不同的污染物计算AQI时有不同的标准,当天AQI的值是通过每个污染物计算AQI的值,再从其中选取最大的值

得到的.根据式(1)和式(2),各算法对未来第7天AQI的值的预测结果如表3所示. CK-UNR算法的AQI预测值与实际值最接近, ELM算法其次, UNR算法和BP算法的结果相差较大.

最后使用均方误差(MSE)、均方根误差(RMSE)、平均绝对百分误差(MAPE)、平均绝对误差(MAE)四个参数来评估模型的预测效果,如表4所示.四个评价指标的值越小表示预测值与真实值的误差越小,算法精确度越高.可以看出, CK-UNR算法预测的效果和真实值之间的误差最小(见表中黑体字),取得了较好的预测结果.

3.3 时间复杂度分析 假设有 m 个城市, CK-UNR算法的第一步,对数据集 B_i 和 A_i 进行离差

表3 第七天AQI的预测值

Table 3 The predicted AQI on the seventh day

Algorithm	CO	NO ₂	O ₃	PM ₁₀	PM _{2.5}	SO ₂	AQI
Actual value	1	41	58	137	76	10	102
CK-UNR	1	63.26	62.48	105.38	78	12.75	103.79
UNR	1.11	41.21	63.42	115.76	110.91	1.13	144.89
BP	0.90	51	47	127	60	10	88.5
ELM	1.21	60.51	63.76	140.16	81.39	14.13	107.99

表4 用来评估预测模型的参数值

Table 4 Parameters for the evaluation of the predictive model

Algorithm	MSE	RMSE	MAPE	MAE
CK-UNR	3.20	1.79	0.017	1.79
UNR	1839.55	42.89	0.42	42.89
BP	182.25	13.5	0.13	13.5
ELM	35.88	5.99	0.058	5.99

归一化,时间复杂度为 $O(m)$ 。第二步,通过余弦相似度算法,计算任意两个城市 x 和 y 间气象因素的相似度矩阵 $\text{sim}(B'_x, B'_y)$,时间复杂度为 $O(m^2)$ 。第三步,考虑时间因素,给相似度矩阵一个系数 ρ^{t-l+1} ,形成任意两个城市间的相似度矩阵 SA ,时间复杂度为 $O(m)$ 。第四步,对矩阵 SA 进行 k -means社区划分,时间复杂度为 $O(m)$ 。第五步,通过社区划分,结合其他城市和目标城市的相似度,得出关于目标城市的矩阵 $A_h(Y)$,时间复杂度为 $O(1)$ 。第六步,基于非线性回归方程进行模型的构建,通过最小二乘法进行拟合,时间复杂度为 $O(m^2)$ 。所以CK-UNR算法的总的时间复杂度为 $O(m^2)$ 。

4 结 论

空气质量指数AQI是衡量空气质量好坏的标准,实现AQI的预测可以对大气污染的治理提供指导。本文结合气象因素和周边城市对目标城市空气质量的影响,对目标城市的AQI进行预测。实验以淮安地区为例,首先计算江苏省内各个城市间特定时间内气象因素的相似度,根据相似度矩阵进行社区的划分,得到与淮安在同一社

区的城市信息,结合其气象信息和污染物信息通过非线性回归进行算法的建模。实验结果证明,该算法不但提升了预测精度,而且有效地降低了算法的时间复杂度。

参考文献

- [1] Thach T Q, Tsang H, Cao P H, et al. A novel method to construct an air quality index based on air pollution profiles. *International Journal of Hygiene and Environmental Health*, 2017, 221(1): 17–26.
- [2] Brunekreef B, Holgate S T. Air pollution and health. *The Lancet*, 2002, 360(9341): 1233–1242.
- [3] 张欣, 许建明, 王体健等. 上海市一次重霾污染过程的特征及成因分析. *南京大学学报(自然科学)*, 2015 (3): 463–472. (Zhang X, Xu J M, Wang T J, et al. Characteristics and formation mechanism of a serious haze episode in December 2013 in Shanghai. *Journal of Nanjing University (Natural Science)*, 2015(3): 463–472.)
- [4] Fu B, Xiao H G, Wu L F. Grey relational analysis for the AQI of Beijing, Tianjin, and Shijiazhuang and related countermeasures. *Grey Systems: Theory and Application*, 2018, 8(2): 156–166.
- [5] Yao W X, Zhang C X, Xiao W, et al. The research of new daily diffuse solar radiation models modified by air quality index (AQI) in the region with heavy fog and haze. *Energy Conversion and Management*, 2017, 139: 140–150.
- [6] Liu C M. Effect of PM_{2.5} on AQI in Taiwan. *Environmental Modelling & Software*, 2002, 17(1): 29–37.
- [7] Carbajal-Hernández J J, Sánchez-Fernández L P, Carrasco-Ochoa J A, et al. Assessment and prediction of air quality using fuzzy logic and

- autoregressive models. *Atmospheric Environment*, 2012, 60: 37—50.
- [8] Corani G, Scanagatta M. Air pollution prediction via multi-label classification. *Environmental Modelling & Software*, 2016, 80: 259—264.
- [9] Singh K P, Gupta S, Kumar A, et al. Linear and nonlinear modeling approaches for urban air quality prediction. *Science of the Total Environment*, 2012, 426: 244—255.
- [10] 李博群, 贾政权, 刘利平. 基于模糊时间序列的空气质量指数预测. *华北理工大学学报(自然科学版)*, 2018, 40(3): 78—86. (Li B Q, Jia Z Q, Liu L P. Air quality index forecast based on fuzzy time series models. *Journal of North China University of Science and Technology (Natural Science Edition)*, 2018, 40(3): 78—86.)
- [11] 高帅, 胡红萍, 李洋等. 基于改进的思维进化算法与 BP 神经网络的 AQI 预测. *数学的实践与认识*, 2018, 48(19): 151—157. (Gao S, Hu H P, Li Y, et al. AQI prediction based on mind evolutionary algorithm and BP neural network. *Mathematics in Practice and Theory*, 2018, 48(19): 151—157.)
- [12] 刘洪通, 冯百明, 温向慧等. 基于 Storm 的 AQI 实时预测模型. *计算机工程与设计*, 2019, 40(1): 296—301. (Liu H T, Feng B M, Wen X H, et al. Real-time AQI prediction model based on Storm. *Computer Engineering and Design*, 2019, 40(1): 296—301.)
- [13] 张春露, 白艳萍. 基于 TensorFlow 的 LSTM 模型在太原空气质量 AQI 指数预测中的应用. *重庆理工大学学报(自然科学)*, 2018, 32(8): 137—141. (Zhang C L, Bai Y P. Application of LSTM prediction model based on tensor flow in taiyuan air quality AQI index. *Journal of Chongqing Institute of Technology*, 2018, 32(8): 137—141.)
- [14] 白鹤鸣, 沈润平, 师华定等. 基于 BP 神经网络的空气污染指数预测模型研究. *环境科学与技术*, 2013, 36(3): 186—189. (Bai H M, Shen R P, Shi H D, et al. Forecasting model of air pollution index based on BP neural network. *Environmental Science & Technology*, 2013, 36(3): 186—189.)
- [15] Li X, Peng L, Hu Y, et al. Deep learning architecture for air quality predictions. *Environmental Science and Pollution Research*, 2016, 23(22): 22408—22417.
- [16] Xi X, Wei Z, Rui X G, et al. A comprehensive evaluation of air pollution prediction improvement by a machine learning method//2015 IEEE International Conference on Service Operations, Logistics and Informatics (SOLI). Hammamet, Tunisia: IEEE, 2015: 176—181.
- [17] Han L J, Zhou W Q, Li W F, et al. Meteorological and urban landscape factors on severe air pollution in Beijing. *Journal of the Air & Waste Management Association*, 2015, 65(7): 782—787.
- [18] Sharma M, Aggarwal S, Bose P, et al. Meteorology-based forecasting of air quality index using neural network//IEEE International Conference on Industrial Informatics. Banff, Canada: IEEE, 2003: 374—378.
- [19] Cui H, Ma R, Gao F. Relationship between meteorological factors and diffusion of atmospheric pollutants. *Chemical Engineering Transactions*, 2018, 71: 1417—1422.
- [20] Li Y, Chiu Y H, Lu L C. Energy and AQI performance of 31 cities in China. *Energy Policy*, 2018, 122: 194—202.
- [21] Liu M, Gu J L, Liu L J, et al. Air quality analysis of Dalian in summer based on AQI data. *Journal of Atmospheric and Environmental Optics*, 2016, 11(2): 111—117.
- [22] Newman M E J. Fast algorithm for detecting community structure in networks. *Physical Review E*, 2004, 69(6): 066133.
- [23] Newman M E J. Community detection in networks: modularity optimization and maximum likelihood are equivalent. 2016, arXiv:1606.02319.

(责任编辑 杨可盛)