

基于机器学习的细粒度空气质量时间预测器

曹鑫磊, 冯 锋

(宁夏大学信息工程学院, 宁夏 银川 750021)

摘 要: 针对空气质量状况的预防和治理, 提出了基于线性回归的空气质量预测方法。作为时间预测器, 更多的是考虑本地的历史数据、过去与未来的状况, 挖掘空气质量随时间变化的关系以及变化趋势, 为了选取最优的特征, 对数据进行分析, 找到与空气质量变化关系密切且存在线性关系的特征, 以此特征来建立线性回归模型, 并使用 Cross-Validation 方法进行评估验证。实验结果表明线性回归方法能够有效地预测到空气质量的变化趋势, 效果良好, 具有较强的实用性。

关键词: 空气质量预测; 线性回归; 时间预测器; Cross-Validation; 特征选择

中图分类号: X831; TP311.1

文献标志码: A

DOI: [10.16803/j.cnki.issn.1004-6216.2020.02.016](https://doi.org/10.16803/j.cnki.issn.1004-6216.2020.02.016)

Fine-grained Air Quality Time Predictor Based on Machine Learning

CAO Xinlei, FENG Feng

(Information Engineering Institute, Ningxia University, yinchuan 750021, China)

Abstract: Aiming at the prevention and treatment of air quality conditions, the air quality prediction method based on linear regression is proposed. As a time predictor, it should consider the local historical data, past and future conditions in more details, and explore the relationship between the air quality and time. In order to select the optimal characteristics, the data is analyzed to find the characteristics that are closely related to the air quality variation and have a linear relationship. The linear regression model is established based on the characteristics, and it is then further evaluated and validated by the Cross-Validation method. The experimental results show that the linear regression method can effectively predict the trend of the air quality. Thus, the proposed method in this paper has a good effect and a strong practicability.

Keywords: Air Quality Prediction; Linear Regression; Time Predictor; Cross-Validation; Feature Selection

CLC number: X831; TP311.1

长期以来, 空气质量状况只能通过气象站点得知, 况且空气质量受多种因素影响, 所以在预测时不仅要考虑自身周边环境的影响, 还要考虑空气质量指数自身的变化、随时间推移的变化趋势。本文对细粒度空气质量的时间预测器进行研究, 目的是研究空气质量指数与时间的内在联系, 并将其应用到预测中。

何晓飞等^[1]总结了我国学者在机器学习等领域的一部分研究成果, 其中有很多对优秀的机器学习算法的总结。星敏传感器在轨标定算法^[2]结合了机器学习的预测建模思想, 使得预测的效果提升较为明显。目前广泛应用的预测算法有随机森林^[3]、回

归树^[4]、局部加权线性回归^[5]、岭回归^[6]、前向逐步回归^[7]、高斯混合^[8]以及神经网络^[9]等。吴有训等^[10]采用径向基函数神经网络进行短期的空气质量预测, 该方法具有优秀的函数逼近能力和学习速度, 但是存在结果收敛速度慢和容易陷入局部极小点等缺陷。

时间预测器是基于线性回归, 并通过大数据来挖掘空气质量在气象影响下随时间变化的变化趋势, 以相关的特征建立两者之间的线性关系, 并使用 Hold-out Validation 或者 Cross-Validation 的方法进行验证评估^[11]。此方法简化了预测过程, 也提高了预测的准确性。

收稿日期: 2019-07-22

基金项目: 国家自然科学基金项目(7156102); 宁夏重点研发计划重点项目(2018BFG02003)

作者简介: 曹鑫磊(1995-), 男, 硕士研究生。研究方向: 信息系统工程。E-mail: caoxinlei888@hotmail.com

通信作者: 冯 锋(1971-), 男, 博士、教授、硕士生导师。研究方向: 信息系统工程、RFID 射频识别及无线传感器网络。
E-mail: feng_f@nxu.edu.cn

引用格式: 曹鑫磊, 冯 锋. 基于机器学习的细粒度空气质量时间预测器[J]. 环境保护科学, 2020, 46(2): 81-84.

1 空气质量预测的总体方案

1.1 线性回归预测的原理

运用一个线性函数进行预测的过程就是建立模型的过程,模型建立后解决问题。即求解系数和误差的过程^[12]。模型函数见式(1)。

$$h(x) = w_0 + w_1x_1 + w_2x_2 + w_ix_i + \dots + w_nx_n \quad (1)$$

式(1)中: x_i 为第 i 个特征的值; w_i 为 x_i 对应的系数; $h(x)$ 代表预测结果。

假设有训练数据集 D 见式(2)。

$$D = \{(X_1, Y_1), (X_2, Y_2), (X_i, Y_i), \dots, (X_n, Y_n)\} \quad (2)$$

式(2)中: X_i, Y_i 表示各横纵坐标的特征值。

对第 i 条数据进行预测, 函数见式(3)。

$$h(x^i) = w_0 + w_1x_1^i + w_2x_2^i + \dots + w_nx_n^i \quad (3)$$

式(3)中: x^i 为第 i 条数据里面的各特征值。

损失函数是一种评估成本或者损失的函数, 定义损失函数的目的就是让求解模型函数系数时有好的放矢, 即找到的系数应该是让损失函数越小越好。函数式见(4)。

$$J(W) = \frac{1}{2M} \sum_{i=0}^M (h(x^i) - y^i)^2 \quad (4)$$

式(4)中: y^i 为真实值; $h(x^i)$ 为预测值; M 为总数据记录数, 前面 $\frac{1}{2M}$ 是一个方便分析的自定义常数; W 为损失函数的系数。将上式用矩阵表示见式(5)。

$$J(W) = \frac{1}{2M} (XW - y)^T (XW - y) \quad (5)$$

式(5)中: X, y 分别为 $h(x^i), y^i$ 的矩阵, 化简求导得式(6)。

$$W = (X^T X)^{-1} X^T y \quad (6)$$

式(6)中出现了求逆运算^[13], 常用的求逆方法有最小二乘法、梯度下降法。

求逆方法: 最小二乘法最简单, 但是效率是较低的, 本文不推荐此方法。

梯度下降法的流程如下:

1) 对 W 赋值(可随机), 或者使 W 为一个全 0 的向量;

2) 改变 W 的值, 使 $J(W)$ 按梯度下降的方法进行减少;

3) 梯度方向由 $J(W)$ 对 W 的偏导数确定, 由于求极小值, 因此梯度方法是偏导数的反方向。结果见式(7)。

$$W_j = W_j + a(y^i - h(x^i))x_i^j \quad (7)$$

迭代更新的方式有 2 种: 一种是批梯度下降, 此方法结果能够不断收敛; 另一种是增量梯度下降, 此方法结果可能不断在收敛处徘徊。

1.2 建立模型

本设计完成的是基于线性回归的空气质量预测的时间预测器模块, 目的是预测一个气象站未来 48 h 空气质量指数的变化量。使用线性回归预测模型进行预测, 基于 3 方面的数据: ①过去 2 h 空气质量指数和现在的空气质量指数; ②当地的气象状况; ③要预测最近的天气状况, 首先获取到气象站一年内的空气质量, 见图 1。

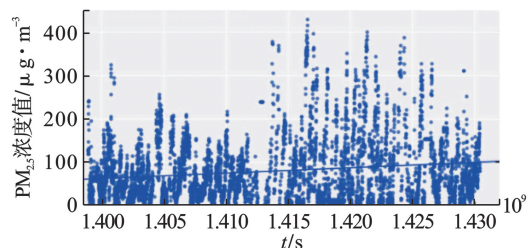


图 1 1001 气象站 1 年内的空气质量

当前时间自然界的不同状态对未来不同时间段的影响程度是不一样的。把输入数据(虚线矩形部分)分别和不同的时间段的空气质量进行匹配以规划出不同的训练集, 见图 2。

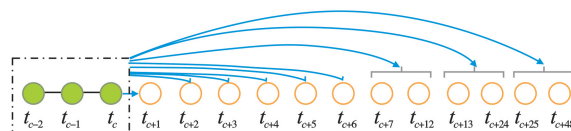


图 2 时间预测器的描述

图 2 可见, 前 6 h 中每个小时都有一个预测器, 但是从第 7 ~ 48 h 被分成了 3 部分, 每一部分都能预测出 2 个值, 一个最大值, 一个最小值。对于 7 ~ 24 h 的模型, 有一种思路是, 每个小时都建立一个模型, 分别预测出每个小时的空气质量指数, 取得最大值和最小值。但这种方法无论在结果数量上还是预测效率上都不够优秀。所以根据相邻时间点空气质量指数变化量不大的特点, 挑取离散时间点(如 7, 9 和 12), 这样就可以尽量覆盖到多个取值, 使效率和准确度得到了均衡提高。

1.3 特征选择

根据预测所需要的 3 类数据, 用到的预测特征包括: 过去 2 h 的空气质量指数; 过去 1 h 的空气质量指数; 当前时间的空气质量指数; 当前时间以及未来时间的天气、气压、温度、风速及风向等状况。

由于要验证线性回归算法, 考量一个特征是否真正的适合作为预测特征, 最直观的方法是模拟该特征与污染物浓度是否呈现线性关系, 特征选取过程见图 3(a~g)。

图 3(a~c)可见, 过去时间的 $\text{PM}_{2.5}$ 浓度和当前时间的 $\text{PM}_{2.5}$ 浓度都与未来 1 h 的 $\text{PM}_{2.5}$ 浓度呈正相关, 空气质量变化是连续的。由图 3d 可见, 当天气不是强对流天气时, 未来 1 h 浓度的最大值接近 400; 而天气取值在 5~13 h, 浓度的最大值在 300 左右; 当天气数值再增大时, 很明显浓度的范围再次增大, 不存在线性关系, 可以放弃该特征。天气和浓度的线性关系表现差的原因可能是天气的离散化表示, 因为天气在数据中的表示采用的是离散数值, 这样的表示方式一方面会降低天气描述的准确性, 另一方面会让天气对浓度造成的影响不能以线性的关系体现。

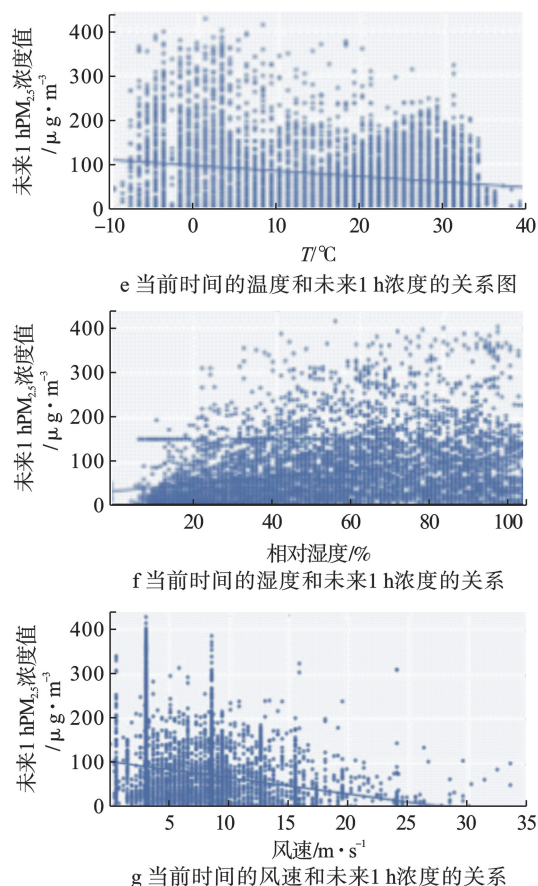
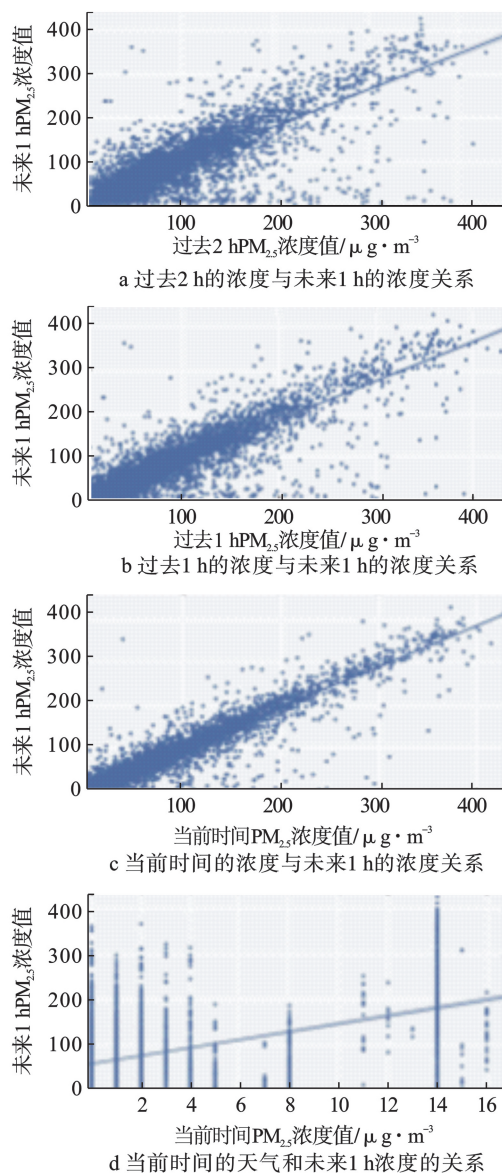


图 3 线性回归算法特征选取过程

图 3(e~g)分别反映了当前时间的温度、湿度和风速在未来 1 h 的浓度关系, 除了噪声点外, 温度和风速与浓度呈现负线性相关, 温度与浓度呈正线性相关。

所以最终确定选取的特征为: 过去 2 h 和当前时间的浓度, 当前时间本地的温度、湿度、风速, 预测时间的温度、湿度和风速。

使用随机森林中数据降维和特征选择的方法^[14], 借助随机森林来验证或者量化一些特征的重要性。其中主要实现代码(Python 语言)为:

```
Feature, target = data[['temperature', 'humidity',
'wind_speed', 'next_temperature', 'next_humidity',
'next_wind_speed']], data[next_PM25]
```

```
ss = RandomForestRegressor().fit(feature, target)
```

其中 data 为输入, 每条记录是由各个特征值和未来 1 h 的浓度组成, 输出为每个特征的重要性。

2 结果与分析

在确定了特征和建立模型之后, 最后的工作就是要对模型进行评价, 采用 K 折交叉验证的方式^[15], 将全部数据分为 K 份, 先选择其中 1 份作为测试

集,其他 $k-1$ 份作为训练集,进行训练和验证,同样的过程循环 k 次即可。

实验选取 3 折交叉验证,准确率通过取其平均值得到,见图 4。红色线是浓度的真实值,蓝色线是浓度的预测值。

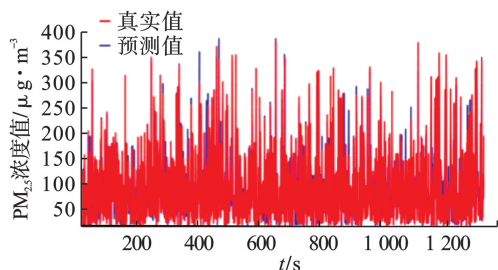


图 4 未来 1 h 预测结果与真实值的对比

用准确率函数 score 计算得到准确率为 0.908 5,其中 score 返回的是该次预测的 R^2 系数。方法见式(8)。

$$R^2 = 1 - \frac{\sum_{i=0}^M (\hat{y}_i - y_i)^2}{\sum_{i=0}^M (y_i - \bar{y})^2} \quad (8)$$

式(8)中: $\hat{y}_i$ 为预测值; y_i 为真实值; $\bar{y}$ 为真实值的平均值; M 为数据量。

每一段都要预测一个最大值和最小值,求出平均值 $\hat{m}$,在求出这段区间的真实值的平均值 m ,计算准确率方法见式(9)。

$$p = 1 - \frac{\hat{m}}{m} \quad (9)$$

同时,每个时间段还要计算误差^[11],见式(10)。

$$e = \frac{\sum_{i=0}^M |\hat{y}_i - y_i|}{M} \quad (10)$$

分别对未来的 1, 2, 3, 4, 5, 6, 7, 9, 12, 13, 18, 24, 25, 36, 48 h 建立模型进行预测。每个小时预测的准确率情况见图 5。

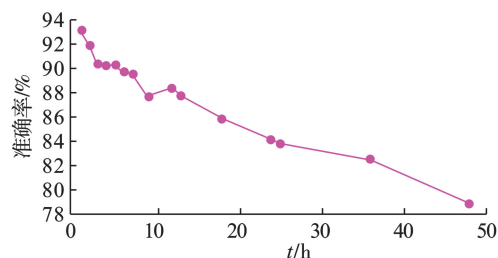


图 5 未来每个小时的预测准确率

实验结果表明,天气是由大自然千万种变化因素综合影响的,我们只抓住了几个最重要的特征进行预测,难以保证准确率达到 100%。在预测未来 5 h 以内的空气质量变化时,准确率能够保持在 90% 以

上。实验选定的是预测未来 48 h 内的变化,由图 5 可见,准确率即使在减小,依然能达到 75%,表现良好。

3 结论

为了实现对空气质量的预测,在选定的主要特征条件下,对空气质量变化做进一步研究,经过 Cross-Validation 方法、准确率函数 score 并将误差考虑在内,对线性回归方法进行评估验证,预测结果能够达到应用的标准。因此,空气质量由温度、湿度和风速这 3 个主要特征共同影响,为空气治理与预防提供了参考,具有现实意义。水源如同大气一样也是我们赖以生存的条件,借此方法,还可以选定水质的预测特征,对水质在未来的变化趋势做一个预测,为水质的预防和治理提供参考。

参考文献

- [1] 何晓飞,郭茂祖,张敏灵. 大数据时代的机器学习研究专刊前言[J]. 软件报, 2015, 26(11): 2749 - 2751.
- [2] 刘源,谢睿达,赵琳,等. 基于机器学习的大视场星敏感器畸变在轨标定技术[J]. 红外与激光工程, 2016, 45(12): 282 - 290.
- [3] 杨思琪,赵丽华. 随机森林算法在城市空气质量预测中的应用[J]. 统计与决策, 2017(20): 83 - 86.
- [4] 沈夏炯,张俊涛,韩道军. 基于梯度提升回归树的短时交通流预测模型[J]. 计算机科学, 2018, 45(6): 222 - 227, 264.
- [5] 卢月明,王亮,仇阿根,等. 局部加权线性回归模型的 $PM_{2.5}$ 空间插值方法[J]. 测绘科学, 2018, 43(11): 79 - 84, 91.
- [6] 杨楠. 岭回归分析在解决多重共线性问题中的独特作用[J]. 统计与决策, 2004(3): 14 - 15.
- [7] 张维刚,廖兴涛,钟志华. 基于逐步回归模型的汽车碰撞安全性多目标优化[J]. 机械工程学报, 2007, 43(8): 142 - 147.
- [8] 乔少杰,金琨,韩楠,等. 一种基于高斯混合模型的轨迹预测算法[J]. 软件学报, 2015, 26(5): 1048 - 1063.
- [9] 刘豹,胡代平. 神经网络在预测中的一些应用研究[J]. 系统工程学报, 1999, 14(4): 338 - 344.
- [10] 吴有训,彭慕平,刘勇. 基于径向基函数网络的宣城市空气质量预测[J]. 安徽师范大学学报(自然科学版), 2011, 34(4): 374 - 379.
- [11] 杨柳,王钰. 泛化误差的各种交叉验证估计方法综述[J]. 计算机应用研究, 2015, 32(5): 1287 - 1290, 1297.
- [12] 冷建飞,高旭,朱嘉平. 多元线性回归统计预测模型的应用[J]. 统计与决策, 2016(7): 82 - 85.
- [13] 刘颖超,张纪元. 梯度下降法[J]. 南京理工大学学报(自然科学版), 1993(2): 12 - 16, 22.
- [14] 吴晓婷,闫德勤. 数据降维方法分析与研究[J]. 计算机应用研究, 2009, 26(8): 2832 - 2835.
- [15] ALIPPI C, ROVERI M. Virtual k-fold cross validation: An effective method for accuracy assessment[P/OL]. Neural Networks (IJCNN), The 2010 International Joint Conference on, 2010. <https://ieeexplore.ieee.org/document/5596899>.