Software Engineering and Applications 软件工程与应用, 2021, 10(1), 24-29
Published Online February 2021 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2021.101004

Hans汉斯

一种基于随机配置网络的PM2.5浓度预测方法

王前进，蔡佳浩

盐城工学院电气工程学院，江苏 盐城
Email: 794318558@qq.com

收稿日期：2021年1月26日；录用日期：2021年2月7日；发布日期：2021年2月24日

摘　要

作为评价空气质量的重要指标，细颗粒物(Fine Particulate Matter, PM2.5)浓度是实现环境治理的基础。降低PM2.5浓度可有效改善空气质量，减少各种呼吸道病况的发生。因此，PM2.5浓度的预测变得尤为重要。本文基于随机配置网络算法，建立一个非线性回归模型用于预测PM2.5的浓度。实验结果表明：采用随机配置网络算法建立的PM2.5浓度预测模型具有较高的预测精度。

关键词

PM2.5，随机配置网络，预测模型

A Method of Predicting PM2.5 Using Stochastic Configuration Network

Qianjin Wang, Jiahao Cai

School of Electrical Engineering, Yancheng Institute of Technology, Yancheng Jiangsu
Email: 794318558@qq.com

Received: Jan. 26th, 2021; accepted: Feb. 7th, 2021; published: Feb. 24th, 2021

Abstract

As a key index of measuring the air quality, fine particulate matter (PM2.5) plays an important role in realizing the environmental treatment. The air quality can be effectively improved by reducing the PM2.5 concentration, which can prevent the occurrence of various kinds of respiratory diseases. Hence, it is very important to predict the PM2.5. In this paper, using the stochastic configuration network algorithm, a nonlinear regression model is established to predict the PM2.5. Experiment result illustrates that the established PM2.5 prediction model using the stochastic configuration network has a high precision.

文章引用: 王前进, 蔡佳浩. 一种基于随机配置网络的 PM2.5 浓度预测方法[J]. 软件工程与应用, 2021, 10(1): 24-29.
DOI: 10.12677/sea.2021.101004

Keywords

PM2.5, Stochastic Configuration Network, Prediction Model

1. 引言

近年来，随着现代化进程的快速推进，空气质量问题(特别是雾霾问题)越发凸显，严重影响公众的身体健康[1]。雾霾的主要成分是细颗粒物(Fine Particulate Matter, PM2.5)，当其浓度越高时，说明空气的质量越差。PM2.5 的体积小，需要较长时间才能在空气中消散，而且活性强，可吸附含有毒物质的污染物[2] [3]。当 PM2.5 进入到人体后，会造成各种呼吸道病况的发生[4] [5]。因此，有必要对 PM2.5 浓度进行准确预测，为环境治理人员和公众提供 PM2.5 浓度信息。

目前，作为一种随机增量学习方法，随机配置网络因其引入了监督机制对将要添加的隐含层节点进行筛选，使其具有强大的预测能力和极快的学习速度[6] [7]。随机配置网络的构造分两步进行：1) 采用监督机制来随机生成一组隐含层节点，从中选择一个使当前网络具有最高精度的隐含层节点作为新增节点；2) 采用优化算法来确定整个网络的输出权值。因此，随机配置网络不仅结构简单、易于实现，而且具有较强的预测能力。因此，借助于随机配置网络，本文建立一种新型的 PM2.5 浓度预测方法，致力于构建一种结构简单、易于实现，且能够用于 PM2.5 浓度预测的方法。

2. 预备知识

随机配置网络

随机配置网络采用点增量的方式，即一个一个地添加隐含层节点的方式，来构建网络，直至达到事先设定的停止条件。其停止条件是指：根据运行系统内存的要求，同时考虑计算复杂性，设置最大的隐层节点和期望达到的预测精度，只要满足上述两个条件之一，就结束网络的构建。

随机配置网络的监督机制有如下形式：

$$\langle e_{L-1}, g_L \rangle^2 \geq g_L^2 (1-r-u_L) \|e_{L-1}\|^2 \tag{1}$$

式中，e_{L-1} 为当前网络的残差，L 为当前网络隐含层节点数，$0<r<1$，$\lim_{L\to\infty} u_L = 0$，且 $0<u_L \leq 1-r$。再添加完隐含层节点后，随机配置网络采用如下的权值更新算法来获得整个网络的输出权值[8] [9]：

$$\left[\beta_1^*, \beta_2^*, \cdots, \beta_L^*\right] = \arg\min_{\beta} \left\| f - \sum_{j=1}^{L} \beta_j g_j \right\| \tag{2}$$

3. 基于随机配置网络的 PM2.5 浓度预测模型

本文选择来自于盐城市在 2016 年下半年到 2017 年下半年的空气检测数据来建立 PM2.5 浓度的预测模型。其中，将 2016 年下半年至 2017 年上半年的数据用于训练所提 PM2.5 浓度预测算法，剩余的数据作为测试集，来验证所提 PM2.5 浓度预测模型的预测能力。其他的空气质量指标还包括 CO、NO_2、O_3、SO_2、和 PM10 这 5 种污染物质。

3.1. 数据预处理

由于数据集中存在缺失值，不能直接用于模型的训练。数据集的处理方法一般有两种，一种是直接将缺失的数据删除，另一种是采用数据填补算法来补充缺失的数据。为了简单起见，本文采用数据删除法来处理缺失值。具体操作如下：

本文利用 Matlab 的 isnan 函数来删除数据中的缺失数据，程序如下：

```
for i=size(Xtraino, 1):-1:1
    if sum(isnan(Xtraino(i, :))) > 0
        Xtraino(i, :)=[];
    end
end
delete(' Xtrain.csv');
dlmwrite('Xtrain.csv', Xtraino, 'precision', 12);
```

其中，Xtraino 为原始的训练数据，Xtrain 为不存在缺失值的训练数据。部分原始训练数据如表 1 所示，删除缺失值之后的训练数据如表 2 所示。

Table 1. Some original training data
表 1. 部分原始训练数据

0.933	14	80	21	53	35
0.921	24	77	21	60	41
0.887	27	69	21	73	47
0.87	30	49	22	80	48
0.411	4	79	7	78	9
0.429	4	70	4	44	8
0.496	4	91	4	9	8
0.494	2	117	4	10	8
0.441	3	128	3	NaN	9
0.442	6	126	3	21	9
0.445	12	119	3	NaN	8
0.454	23	119	50	NaN	8
0.463	33	120	97	NaN	8
0.471	44	120	144	20	8
0.48	12	116	30	9	8

Table 2. The corrected training data
表 2. 部分不存在缺失值的训练数据

0.933	14	80	21	53	35
0.921	24	77	21	60	41
0.887	27	69	21	73	47
0.87	30	49	22	80	48

Continued

0.411	4	79	7	78	9
0.429	4	70	4	44	8
0.496	4	91	4	9	8
0.494	2	117	4	10	8
0.442	6	126	3	21	9
0.471	44	120	144	20	8
0.48	12	116	30	9	8

3.2. 数据的归一化

在建立 PM2.5 浓度预测模型之前，要对所有的训练数据集和测试数据集进行归一化处理。本文采用 Matlab 的 mapminmax 函数对所有数据进行归一化，将其缩放到 $[0,1]$ 区间内。具体操作如下：

```
[Xntrain,inputps]= mapminmax(Xtrain);
Xntest=mapminmax('apply',Xtest,inputps);
[Yntrain,outputps]=mapminmax(Ytrain);
Yntest=mapminmax('apply',Ytest,outputps);
```

其中，Ytrain，Xtest 和 Ytest 分别为训练集的输出，测试集的输入和输出；Xntrain，Yntrain，Xntest 和 Yntest 为归一化后的结果。

3.3. 训练集与测试集

每个时刻采集 6 个数据，将前两个时刻采集的 12 个数据作为输入，当前时刻的 PM2.5 的浓度作为输出。那么，经上述处理后，训练集包含 8218 个样本，测试集包含 4018 个样本，分别用于所建 PM2.5 浓度预测模型的训练和验证。

3.4. PM2.5 浓度预测模型的实现

由前面内容可知，基于随机配置网络的 PM2.5 浓度预测模型从 $L=1$ 开始添加隐含层节点，首先利用监督机制来指导隐含层参数的生成，在添加完隐含层节点后，再采用优化算法(2)来获得网络的输出权值。因此，基于随机配置网络的 PM2.5 浓度预测模型的构建分为两部分：隐含层节点的添加和输出权值的更新。

令 $X=\{x_1,x_2,\cdots,x_{8218}\}, x_i=\{x_{i1},x_{i2},\cdots,x_{i12}\}\in R^{12}$ 为训练模型的输入，则对应的模型的输出为 $Y=\{y_1,y_2,\cdots,y_{8218}\}, y_i=\{y_{i1}\}\in R^1$。给定两个空集 Ω 和 W，用于存储监督机制所产生的隐层参数 w 和 b。模型的具体实现步骤如下：

1) 隐含层节点的增加。首先在监督机制下随机配置一组隐层参数，若找不到满足监督机制的节点，则对监督机制的约束力度进行放松，即 $r=r+\tau$，$\tau\in(0,1-r)$，直至找到监督机制下的隐层参数为止。将获得的隐性参数存储于 Ω 和 W 中，再从中找出一组隐性参数，即 w^* 和 b^*，使得当前模型具有最高的预测精度，并将其代入事先指定的激活函数，形成新增节点。

2) 输出权值的更新。在确定激活函数的随机参数后，需要对隐含层与输出层之间的连接权值进行更新。本文采用全局优化算法(2)来得到网络的输出权值。最后，重复步骤 1)和 2)，直至到达设置的停止条件，模型建立完毕。

3.5. 模型参数选择

所建 PM2.5 浓度预测模型包含以下训练参数：$L_{\max}$ 为模型训练设定的最大节点数；$T_{\max}$ 为隐含层节点配置的次数；r 为学习参数；ε 为网络容忍度；$\Upsilon := \{\lambda_{\min} : \Delta\lambda : \lambda_{\max}\}$ 为随机参数 w (输入权值)和 b (偏置)的选取范围，$\Delta\lambda$ 为变化步长。另外，选择 $g(x) = 1/(1+\exp(-x))$ 为网络的激活函数。具体参数设置如下：

$L_{\max} = 100$； $T_{\max} = 200$； $r = 0.9$； $\varepsilon = 0.01$； $\Upsilon := \{1:1:20\}$。

4. 模型验证

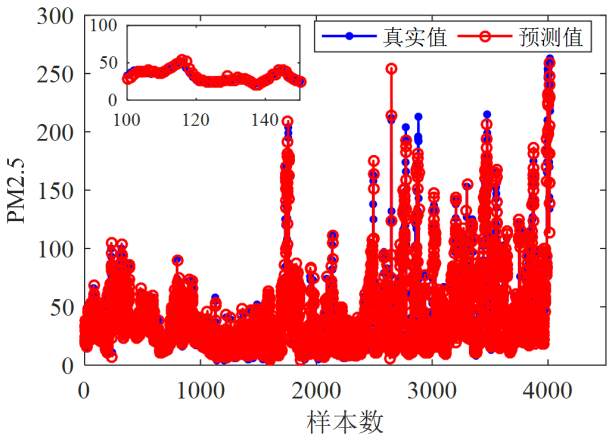

Figure 1. Prediction result of the PM2.5 based on stochastic configuration network
图 1. 随机配置网络的 PM2.5 浓度模型预测结果

基于随机配置网络的 PM2.5 浓度模型的预测结果如图 1 所示。从图 1 可以看出，基于随机配置网络的 PM2.5 浓度预测模型的输出与真实的 PM2.5 浓度是基本匹配的，这说明本文所建 PM2.5 浓度预测模型具有较高的预测精度，可以用于实际中的 PM2.5 浓度的预测。

5. 结论

考虑到 PM2.5 对公众健康和空气监控的重要性，本文集成随机配置网络技术，建立了一种新型的 PM2.5 浓度预测模型。所建立的预测模型，结构简单，便于实现，具有良好的应用前景。然而，本文并没有将数据的时序特征和干扰考虑在内，以致模型的紧致性不够。未来的研究工作是引入时序鲁棒算法来提升预测模型的紧致性。

参考文献

[1] 刘基伟, 闵素芹, 金梦迪. 基于分布式感知深度神经网络的高分辨率 PM2.5 值估算[J]. 地理学报, 2021, 76(1): 191-205.

[2] 王磊, 杨翠丽, 乔俊飞. 基于回声状态网络的 PM2.5 预测研究[J]. 控制工程, 2019, 26(1): 1-5.

[3] 段大高, 赵振东, 梁少虎, 等. 基于 LSTM 的 PM2.5 浓度预测模型[J]. 计算机测量与控制, 2019, 27(3): 215-219.

[4] 郭云, 李瑞娟, 黄炳昭, 等. 未来 10 年保定市大气 PM2.5 的健康效益预测[J]. 中国环境科学, 2020, 40(12): 5459-5467.

[5] 米馨, 张云婷, 胡立文, 等. 大气 PM2.5 长期暴露与儿童血压升高的关联研究[J]. 中华预防医学杂志, 2019, 53(1): 45-50.

[6] 代伟, 李德鹏, 杨春雨, 等. 一种随机配置网络的模型与数据混合并行学习方法[J]. 自动化学报, 2019(45): 1-11.

[7] 张洁, 庞丽萍, 曲洪权, 等. 基于随机配置网络的机载电子吊舱多工况热模型[J]. 化工学报, 2020, 71(S1): 441-447.

[8] 张万栋, 李庆忠, 黎明, 等. 基于最优误差自校正极限学习机的高频地波雷达 RD 谱图海面目标检测算法[J]. 自动化学报, 2021, 47(1): 108-120.

[9] 夏平凡, 倪志伟, 朱旭辉, 等. 基于双错测度的极限学习机选择性集成方法[J]. 电子与信息学报, 2020, 42(11): 2756-2764.