计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190700158

基于 CEEMD-Pearson 和深度 LSTM 混合模型的 $PM_{2.5}$ 浓度预测方法

丁子昂　乐曹伟　吴玲玲　付明磊
浙江工业大学理学院　杭州 310023
(785441595@qq.com)

摘　要　$PM_{2.5}$ 是衡量空气污染物浓度的核心指标。通过挖掘 $PM_{2.5}$ 历史数据的时序特性,完成对未来 $PM_{2.5}$ 浓度值的精确预测具有较强的学术意义和应用价值。然而,原始 $PM_{2.5}$ 浓度值时间序列数据相关性对模型的预测精度产生了较大的影响。为了解决这个问题,文中提出一种基于补充总体经验模态分解-皮尔逊相关分析(CEEMD-Pearson)和深度长短期记忆神经网络(Long Short Term Memory,LSTM)混合模型的 $PM_{2.5}$ 浓度预测方法。该方法利用补充总体经验模态分解(Complementary Ensemble Empirical Mode Decomposition,CEEMD)对 $PM_{2.5}$ 浓度历史数据进行不同频率的分解,增强数据中体现的时序特性。然后通过 Pearson 相关性检验方法对分解后的不同频率子波(IMFs)进行筛选,将筛选后的增强数据输入到多隐含层的深度 LSTM 网络的输入层进行训练并预测。实验数据表明,CEEMD-LSTM 混合模型的预测精度为 80%,但是该模型在训练次数为 7000 次左右才收敛;而经过 Pearson 二次筛选后的模型在训练 800 次左右就已经收敛,并且精度提升到 87%;CEEMD-Pearson 与深度 LSTM 神经网络混合模型的训练效果最优,在训练 650 次左右就已经收敛,并且预测精度达到了 90%。实验结果说明,CEEMD 模态分解方法可以展现出历史数据中的隐藏时序特性,结合 Pearson 相关性分析进行的二次筛选可有效地提升模型训练的收敛速度和预测精度。因此,基于 CEEMD-Pearson 和深度 LSTM 的混合模型可以获得最佳的训练效果、最快的收敛速度以及最精准的预测结果,可以有效解决 $PM_{2.5}$ 浓度预测问题。

关键词:深度神经网络;$PM_{2.5}$;LSTM;CEEMD;Pearson;混合模型
中图法分类号　TP183

$PM_{2.5}$ Concentration Prediction Method Based on CEEMD-Pearson and Deep LSTM Hybrid Model

DING Zi-ang,LE Cao-wei,WU Ling-ling and FU Ming-lei
College of Sciences,Zhejiang University of Technology,Hangzhou 310023,China

Abstract　$PM_{2.5}$ is well-known as the key indicator for measuring the concentration of air pollutants. It is of great significance for both academic study and applications to make accurate prediction of future $PM_{2.5}$ concentration values by excavating the time series characteristics of $PM_{2.5}$ historical data. However,the correlation of time series data of the original $PM_{2.5}$ concentration value has great influence on the prediction accuracy of the model. In order to solve this problem,a $PM_{2.5}$ concentration prediction method based on CEEMD-Pearson and deep LSTM hybrid model was proposed in this paper. The CEEMD modal decomposition method is adopted to decompose the $PM_{2.5}$ concentration historical data at different frequencies,and to enhance the timing characteristics of the data. Then,the Pearson correlation test method is used to screen the different frequency IMFs after decomposition,and the filtered enhancement data is input to the input layer of the deep LSTM network of multiple hidden layers for training and prediction. Experimental data shows that the prediction accuracy of the CEEMD-LSTM hybrid mo-del is 80%. However,the model converges after 7000 training times. While by means of the secondary screening of Pearson correlation test,the model converges after 800 training times,and the prediction accuracy is improved to 87%. At last,the hybrid model combines CEEMD-Pearson with deep LSTM neural network has the best training effect. It converges after 650 training times,and the prediction accuracy reaches 90%. Experimental results show that the CEEMD modal decomposition method can show the hidden time series characteristics in historical data. The secondary screening combined with Pearson correlation analysis can effectively improve the convergence speed and prediction accuracy of model training. Therefore,based on the CEEMD-Pearson and deep LSTM hybrid models,the best training result,the fastest convergence speed and the most accurate prediction result can be obtained,which can effectively solve the $PM_{2.5}$ concentration prediction problem.

Keywords　Deep neural network,$PM_{2.5}$,LSTM,CEEMD,Pearson,Hybrid model

1　引言

如今,空气污染已经成为人们关注的焦点,而空气污染物以 $PM_{2.5}$ 为主。$PM_{2.5}$ 是指大气中直径小于或等于 2.5 微米的颗粒物,它能较长时间悬浮于空气中,其在空气中含量浓度越高,就代表空气污染越严重。世界卫生组织统计资料显示:

基金项目:浙江省科技厅"一带一路"专项(2015C04005)
This work was supported by the Special Project of "One Belt and One Road" of Zhejiang Science and Technology Department (2015C04005).
通信作者:付明磊(fuml@zjut.edu.cn)

全球每年由城市室外和室内空气污染而导致过早死亡的人数超过 200 万。严重的 $PM_{2.5}$ 污染问题已经引起了环境科学界的广泛关注[1]。当前，挖掘 $PM_{2.5}$ 浓度历史数据的时序特性，对未来时间段的 $PM_{2.5}$ 浓度值进行及时的预警已经成为具有较强学术意义和应用价值的研究问题。

早期，学者们对于空气质量数据进行预测的研究方法主要集中在线性回归、随机森林、差分整合移动平均自回归模型（Autoregressive Integrated Moving Average model，ARIMR）以及小波分析等基础数值方法。近年来，随着神经网络在非线性问题领域不断取得显著的效果，一些学者开始尝试利用神经网络作为空气质量数据预测的数学模型。最早使用神经网络对 $PM_{2.5}$ 浓度预测进行研究的是 Patricio 等[2]。他们利用多层神经网络对智利圣地亚哥市中心的一个固定点进行了 24 小时的预测，并与线性回归及持久化方法进行了对比。

在接下来的研究中，一些学者逐渐使用多层感知机（MLP，Multi-layer Perceptron）、支持向量机（SVM，Support Vector Machine）、误差反向传播（BP，Back Propagation）神经网络等经典神经网络对 $PM_{2.5}$ 等污染物浓度进行预测。例如，Ordieres 等利用多层感知器、径向基函数（RBF，Radial Basis Function）和方形多层感知器（SMLP，Square Multilayer Perceptron）对 $PM_{2.5}$ 浓度进行了预测，并与线性回归等模型进行了比较[3]。将神经网络与一些时序分析方法结合进行预测也是其中一个研究方向。如 Luis 等提出了结合差分整合移动平均自回归模型和人工神经网络（Artificial Neural Network，ANN）的新型混合模型，用于提高空气质量和气象数据有限的区域的预报准确性[4]。Al-Alawi 等使用多元回归结合主成分分析（Principle Component Regression，PCR）和人工神经网络（ANN）得到 BPNN 模型，用于预测低层大气中的臭氧浓度水平[5]。Wang 提出了一种结合小波神经网络和遗传算法（GA-WNN）的混合模型，并与 BPNN 做了比较[6]。Fu 等提出了一种灰度模型与前馈神经网络结合的改进模型（RM-GM-FFNN）用来预测 $PM_{2.5}$ 的浓度[7]。还有一些学者对神经网络的参数进行了系列研究[8-11]。

此外，利用模态分解混合的方法应用在神经网络中可以起到对数据特性进行增强的作用。这种混合模型在一定程度上能够解决数据特性不明显时暴露出的神经网络对于数据量以及训练次数的依赖缺陷。Gan 等就利用模态分解的方法结合 LS-SVM 进行了 $PM_{2.5}$ 浓度的预测[12]。Zhu 等提出了两种混合模型（EMD-SVR-Hybrid 和 EMD-IMFs-Hybrid）来预测 AQI 数据[13]。Niu 等结合通过 CEEMD 与 SVM 进行预测[14]。Xu 等提出了一种名为 ICEEMD-WOA-SVM 的混合模型[19]。

同时，更适用于处理时序序列数据的循环神经网络和 LSTM 网络逐渐走入研究者们的视野[2]。Liu 等利用自组织 LSTM 网络对 $PM_{2.5}$ 浓度进行了预测[15]。Huang 等将卷积神经网络（CNN）和长短期记忆（LSTM）组合并应用于 $PM_{2.5}$ 预测系统[16]。Loy-Benitez 在研究中实现了标准 RNN（SRNN）、长短期存储器（LSTM）和门控循环单元（GRU）结构[17]。Soh 等结合多个神经网络预测空气质量长达 48 h，包括人工神经网络、卷积神经网络和长短期记忆，以提取时空关系[18]。

本文基于对 $PM_{2.5}$ 浓度历史数据的时序特性挖掘，提出了一种基于 CEEMD-Pearson 和深度 LSTM 混合模型的 $PM_{2.5}$ 浓度预测方法。具体而言，首先利用 CEEMD 模态分解方法对 $PM_{2.5}$ 浓度历史数据进行不同频率的分解，增强数据中体现的时序特性；然后利用 Pearson 相关性检验方法对分解后的不同频率 IMFs 进行筛选，将筛选后的增强数据输入到深度 LSTM 网络的输入层。此处，本文通过多隐层的深度 LSTM 网络对于 $PM_{2.5}$ 浓度预测值和真实值进行学习训练。实验结果表明，本文所提方法能够获得更为准确的 PM2.5 浓度预测数据。

2 相关工作

2.1 LSTM

本文模型是基于 LSTM 长短期记忆网络进行的改进，而 LSTM 又是在 RNN 循环神经网络基础上的变形，比 RNN 更适用于处理和预测时间序列中间隔和延迟相对较长的重要事件。对于 $PM_{2.5}$ 浓度数据此类复杂多变的实时数据而言，LSTM 显然是更具适应性的预测模型。

RNN 模型是由输入层、隐含层和输出层组成的三层神经网络，但与传统神经网络不同，RNN 模型所有的输入（包括输出）之间都并非独立：在每一次训练中，它针对每一个神经元的每一个操作都依赖于之前的计算结果。该特性使其具备良好的序列预测性能。下式对这一过程进行抽象：

$$h^{(t)}=\sigma(z^{(t)})=\sigma(Ux^{(t)}+Wh^{(t-1)}+b) \tag{1}$$

$$o^{(t)}=Vh^{(t)}+c \tag{2}$$

$$\hat{y}^{(t)}=(o^{(t)}) \tag{3}$$

其中，$x^{(t)}$ 代表 t 时刻训练样本的输入；$h^{(t)}$ 代表 t 时刻模型的隐藏状态，$h^{(t)}$ 由 $x^{(t)}$ 和 $h^{(t-1)}$ 共同决定；$o^{(t)}$ 代表 t 时刻模型的输出，$o^{(t)}$ 只由模型当前的隐藏状态 $h^{(t)}$ 决定；$y^{(t)}$ 代表 t 时刻训练样本序列的真实输出；U，W，V 是 RNN 模型的线性关系参数。

LSTM 区别于 RNN 之处，主要在于它在算法中加入了一个判断信息有用与否的结构，这个结构中放置了 3 扇门，分别为输入门、遗忘门和输出门。其结构如图 1 所示。

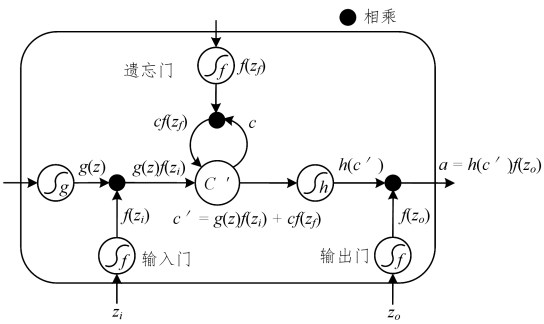


图 1 LSTM 模型

Fig. 1 LSTM model

其中，历史信息 c 在隐含层经遗忘门函数 $f(z_f)$ 作用后衰减保留；而当输入层的一个新的信息到达 LSTM 单元后，先经激活函数激活，然后被输入门函数 $f(z_i)$ 抑制输入到隐含层；隐含层将衰减后的历史信息与新输入的信息加在一起得到新的输出信息 c'；新的输出信息先经激活函数激活，然后

被输出门函数 $f(z_o)$选择输出到输出层。这样 LSTM 克服了 RNN 训练中存在的梯度爆炸和梯度消失的问题。LSTM 神经网络总架构的数学公式如下：

$$i^{(t)}=\sigma(W_i\cdot[h^{(t-1)},x^{(t)}]+b_i) \tag{4}$$

$$f^{(t)}=\sigma(W_f\cdot[h^{(t-1)},x^{(t)}]+b_f) \tag{5}$$

$$o^{(t)}=\sigma(W_o\cdot[h^{(t-1)},x^{(t)}]+b_o) \tag{6}$$

$$h^{(t)}=o^{(t)}*\tanh(C^{(t)}) \tag{7}$$

$$C^{(t)}=f^{(t)}*C^{(t-1)}+i^{(t)}*\tilde{C}^{(t)} \tag{8}$$

$$\tilde{C}^{(t)}=\tanh(W_C\cdot[h^{(t-1)},x^{(t)}+b_C]) \tag{9}$$

其中，$x^{(t)}$代表 t 时刻训练样本的输入；$h^{(t)}$代表 t 时刻当前单元的输出；$i^{(t)}$代表 t 时刻的输入门限；$f^{(t)}$代表 t 时刻的遗忘门限；$o^{(t)}$代表 t 时刻的输出门限；W 和 b 分别是模型的连接权值和偏置。$\tilde{C}^{(t)}$表示前一时刻单元状态，$C^{(t)}$表示单元状态，$h^{(t)}$表示当前单元的输出，$h^{(t-1)}$表示前一时刻单元的输出。

2.2 CEEMD

完备总体经验模态分解(CEEMD)是经验模态分解(EMD)算法的改进算法。EMD 算法是一种分析非线性和非平稳序列的方法。该算法能把一个序列分解成一系列不同频率的子序列，称之为固有模态函数(IMF)。然而，EMD 算法存在模态混叠的缺点，模态混叠会影响分解的固有模态函数的精确度。

于是本文采用 CEEMD 算法对非平稳的 $PM_{2.5}$浓度值时间序列进行预处理，将其分解为一系列相对平稳的分量。其分解过程如下。

将一个数据序列组 X 记为$\{X\}=\{x_1,x_2,\cdots,x_M\}$。选取其中一个序列 $x_1\in A$ 作为处理对象并简记为 x，对 x 进行分解。首先对 x 加入 i 组高斯白噪声生成新序列 x^i，可表示为：

$$x^i=x+\beta_k w^i \tag{10}$$

其中，w^i 为一组高斯白噪声变量，β_k 为分解信号与所添加噪声信噪比的倒数。

然后对 x^i 进行模态分解处理后可以得到模态 $IMF(x^i)$，得到第 1 个余项 r_1，以及第 1 个分解波($\widetilde{IMF}_1$)如下：

$$r_1=\langle IMF(x^i)\rangle \tag{11}$$

$$\widetilde{IMF}_1=x-r_1 \tag{12}$$

对第 k 个余项 $r_k(k=2,3,\cdots,K)$加不同噪声并进行以上处理，K 是模拟函数的总数，可以得到第 k 个分解波($\widetilde{IMF}_k$)：

$$r_k=\langle M(r_{k-1}+\beta_{k-1})E(w^i))\rangle \tag{13}$$

$$\widetilde{IMF}_k=r_{k-1}-r_k \tag{14}$$

2.3 Pearson

Pearson 相关度分析是一种准确度量两个变量之间的关系密切程度的统计学方法，将两组 IMF_1，IMF_2 记为(x_i,y_i) $(i=1,2,\cdots,n)$，则相关系数的数学表达式为：

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{15}$$

其中，$\bar{x}$，$\bar{y}$ 分别是 n 个数据的均值。Pearson 相关度系数 r 是用来衡量两个变量之间的线性相互关系的，取值范围在[−1，+1]之间，将相关度系数分级，分级越细越能够表示两个时间序列不同的相关程度。Pearson 相关度系数与序列间相关性的描述如表 1 所列。

表 1 相关度取值与相关程度

Table 1 Correlation degree and relevance

相关度绝对值	意义
0.00～0.19	极低相关
0.20～0.39	低度相关
0.40～0.69	中度相关
0.70～0.89	高度相关
0.90～1.00	极高相关

本文利用 Pearson 相关性分析方法对 IMFs 序列进行二次筛选，进一步增强神经网络输入数据的时序与相关特性。

3 CEEMD-Pearson 和深度 LSTM 混合模型

为了实现历史数据的深层次挖掘，本文在传统神经网络结构中加入两层 LSTM 单元层，通过多层神经网络结构来挖掘序列的深层特征，并结合 CEEMD 模态分解和 Pearson 相关度过滤，提出了一种基于 CEEMD-Pearson 和深度 LSTM 的混合模型，模型结构如图 2 所示。

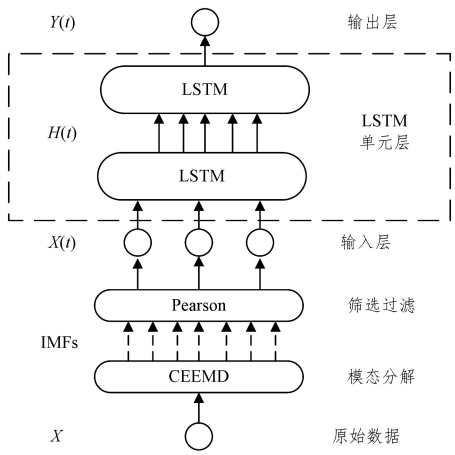


图 2 基于 CEEMD-Pearson 和深度 LSTM 的混合模型

Fig. 2 Hybrid model based on CEEMD-Pearson and deep LSTM

由于 $PM_{2.5}$浓度值的历史数据极不平稳，短期内的变化性较大，因此本文首先对历史数据按照时序利用第 2.2 节给出的 CEEMD 经验模态分解方法进行多模态的分解，分解后得到不同模态下的多个分解波$\widetilde{IMF}_k$。处理后的分解波在保留原数据非线性的同时，其数据的时序变化相较更为平稳，有利于进一步挖掘时序特性。传统的 CEEMD 模态分解方法会将高频波作为杂波直接去除，而本文则是依照 2.3 节中的 Pearson 相关度分析对各模态分解波进行二次过滤，筛选出与原序列相关度较强的分解波序列，作为最终的神经网络输入序列组。经由以上增强和筛选处理后，可以有效增强神经网络的预测精度，优化网络收敛速度。序列信息在 LSTM 网络中的传输过程如图 3 所示。

图 3 展示了深度 LSTM 神经网络的单元结构，其是由一层输入层、一层输出层和两层 LSTM 单元层组成的四层神经网络结构。其中 $x^{(t)}$为 t 时刻经 CEEMD-Pearson 处理后的输入序列，而第一层 LSTM 单元在 t 时刻的输出 $h'(t)$，将作为 t 时刻第二层 LSTM 单元层的输入 $x'(t)$，经与 2.1 节中所述相同的传输之后，得到 t 时刻的最终输出 $h^{(t)}$。据此进行不断训练和学习，得到最终的预测模型。

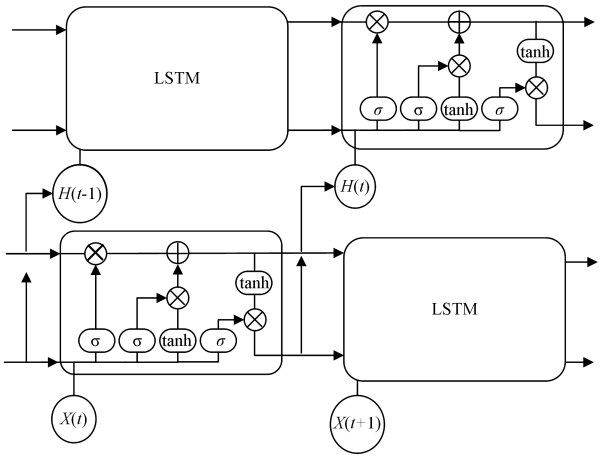


图 3　双层 LSTM 单元结构

Fig. 3　Double layer LSTM unit structure

4　实验与分析

4.1　实验设置

本文在真气网[1]上收集了 2018 年 1 月 1 日至 2018 年 12 月 31 日间每天每小时内的 $PM_{2.5}$ 平均浓度值，共 8 760 条历史数据。数据序列如图 4 所示。

1）数据集划分：选取其中的 80％作为训练集数据，另外 20％作为测试集数据。

2）CEEMD 模态数设置：对数据进行多模态的分解测试，测试显示当模态数为 8 时各分解波的得分最平稳。

3）Pearson 筛选标准：参考表 1 的信息，本文在 Pearson 相关性分析的筛选中，以相关度系数 $r>0.40$ 作为标准选取符合条件的模态分波输入神经网络。

4）神经网络基本参数：经调研及实例分析，本文将神经网络的学习率设为 x；为了使数据的训练过程更为充分，每批次只训练一组序列，故将时间步和样本数均设为 1；输入层单元数与筛选后的输入模态序列数相同，输出层单元数为 1，两层 LSTM 单元层的单元数均为 x。

5）优化算法：对于模型的优化算法，本文中采用 Adam 优化算法，相比随机梯度下降，Adam 优化算法的收敛速度更快而且更稳定。本文使用了 Adam 优化算法的默认超参数设置（$\beta_1=0.9$，$\beta_2=0.999$，$\varepsilon=10^{-8}$）。

6）学习目标：每一次训练中，本文在输出层设置的对比真值均为当前时刻下一时刻的 $PM_{2.5}$ 浓度值，即通过前一时刻的历史数据预测下一时刻的数据。

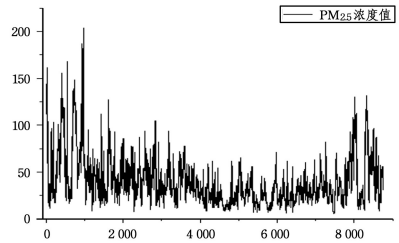


图 4　$PM_{2.5}$ 浓度值历史数据

Fig. 4　Historical data of $PM_{2.5}$ concentration values

4.2　实验结果与分析

4.2.1　CEEMD 模态分解结果

按照 CEEMD 模态分解方法将 PM2.5 浓度数据按照 8 个模态进行分解后，得到 8 组 IMFs 分解波，如图 5 所示。

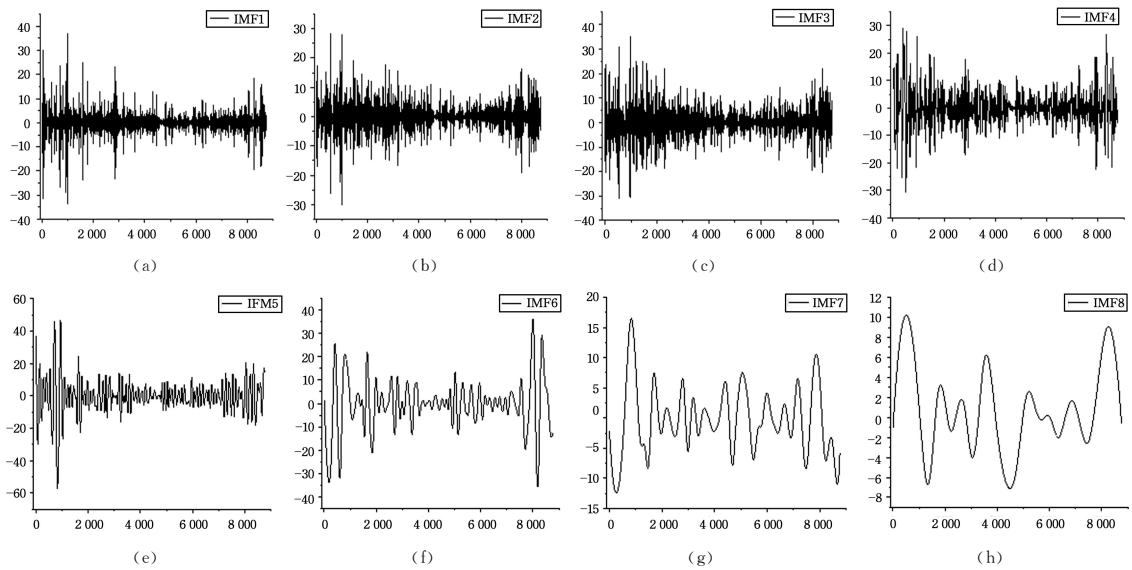


图 5　CEEMD 模态分解波 IMFs

Fig. 5　CEEMD mode decomposition wave IMFs

CEEMD 可以将复杂的 $PM_{2.5}$ 浓度序列分解成包含不同尺度信息且噪声逐渐减少直至消除的 IMFs。由图 5 可知：$PM_{2.5}$ 历史浓度值序列经过 CEEMD 分解得到的子序列，从 IMF1 到 IMF8 呈现频率降低、波长变长、振幅变小的趋势。

综上可知，经 CEEMD 分解后的 IMFs 逐渐表现出一定的变化规律及周期，表明噪声因素已被逐步消除，进而展现出原序列中不同尺度的信息，于是本文将上述 IMFs 分别用于 LSTM 模型进行直接预测和 Pearson 相关性分析方法进行序列的二次清洗处理。

4.2.2　Pearson 相关性分析结果

根据上节中得到的模态分解波 IMFs，将其与 $t+1$ 时刻的 $PM_{2.5}$ 浓度值序列做 Pearson 相关性分析，结果如表 2 所列。

[1] https://www.aqistudy.cn

表 2 IMFs 的 Pearson 相关度得分

Table 2 Pearson correlation scores for IMFs

分解波	IMF1	IMF2	IMF3	IMF4
相关度	0.498429	0.479795	0.432312	0.425830
分解波	IMF5	IMF6	IMF7	IMF8
相关度	0.331135	0.219099	0.167735	0.059112

根据以上结果，IMF1－IMF4 均表现出了与 $t+1$ 时刻 $PM_{2.5}$ 浓度值的中度相关性，IMF5 和 IMF6 只有低度相关性，而 IMF7 和 IMF8 基本不具备相关性。故而，本文选取中度相关性及以上的模态分解波进行 CEEMD-Pearson 与 LSTM 神经网络混合模型的预测。

4.2.3 模型对比分析

对于模型训练误差的判断，本文采用均方误差作为损失函数 Loss 值来评判每一批训练样本的训练误差，公式如下：

$$Loss=\sum_{i=1}^{n}\frac{1}{n}(h_i-y_i)^2 \tag{16}$$

对于训练后模型对测试集数据进行测试的预测精度判断，本文选取了平均绝对百分比误差 MAPE、平均绝对误差 MAE、均方根误差 RMSE 以及 R 平方来综合评判，公式如下：

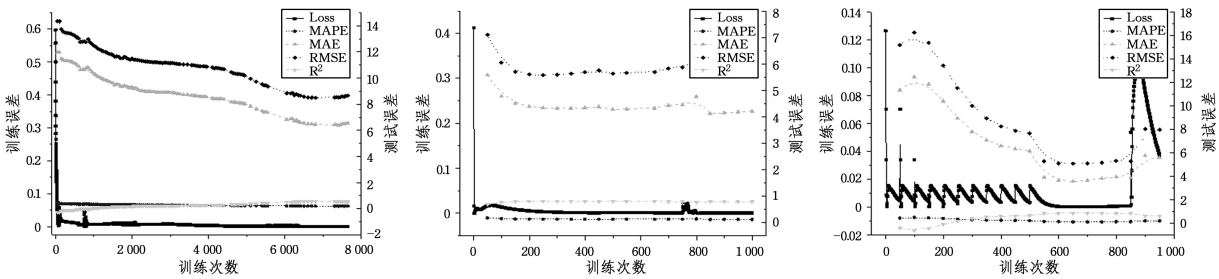
$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\frac{|h_i-y_i|}{y_i} \tag{17}$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|h_i-y_i| \tag{18}$$

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(h_i-y_i)}{n}} \tag{19}$$

$$R^2=1-\frac{\sum_{i=1}^{n}(h_i-y_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{20}$$

将分割后的训练集和测试集数据分别利用 CEEMD-LSTM、CEEMD-Pearson-LSTM 以及双隐含层的 CEEMD-Pearson-Double-LSTM 3 种模型进行训练、学习和预测后，得到以上各评判指标的收敛趋势，如图 6 所示。



(a)CEEMD-LSTM (b)CEEMD-Pearson-LSTM (c)CEEMD-Pearson-Double-LSTM

图 6 模型训练收敛速度对比图

Fig. 6 Comparison chart of model training convergence speed

如图 6 所示，CEEMD 与单隐层 LSTM 混合模型虽在一定程度上取得了相当精准的预测结果，但是其在训练 7000 次左右才收敛；而经过 Pearson 二次筛选后的模型，则在收敛速度上有了极大提升，在训练 800 次左右就已经收敛，并且精度也略有提升；进一步，CEEMD-Pearson 与双隐含层 LSTM 神经网络混合模型的训练效果最优，在训练 650 次左右就已经完全收敛，并且预测精度最高。由此可知，本文提出的 CEEMD-Pearson 与深度 LSTM 神经网络混合模型可有效地提高模型训练的收敛速度，大大提升学习效率。

同时，利用训练好的模型对未来 100 h 的 $PM_{2.5}$ 浓度值进行预测，对比结果如图 7 所示。

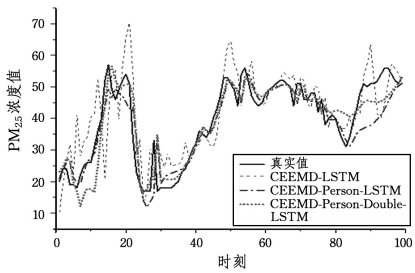


图 7 模型预测结果对比图

Fig. 7 Comparison chart of model prediction results

如图 7 所示，对于未来时刻 100 h 内的 $PM_{2.5}$ 浓度值变化的预测，基于 CEEMD-Pearson 与双隐层 LSTM 神经网络混合模型表现最优，与真值曲线的贴合性最强，CEEMD-Pearson-LSTM 模型次之。而简单的 CEEMD 与单层 LSTM 神经网络的组合模型预测效果相对较差。具体数值如由表 3 所列。

表 3 模型预测误差对比

Table 3 Comparison of model prediction error

	CEEMD-LSTM	CEEMD-Pearson-LSTM	CEEMD-Pearson-Double-LSTM
MAPE	0.200293	0.124229	0.097060
MAE	6.405416	4.273214	3.232446
RMSE	8.516661	5.633898	4.479592
R^2	0.549076	0.802675	0.889593

如表 3 所列，3 种混合模型均取得了较高的预测精度，都在 80%以上，说明本文提出的混合模式十分适用于 $PM_{2.5}$ 浓度值预测。但是简单的 CEEMD-LSTM 混合模型的 MAE 和 RMSE 都普遍偏高，且 R^2 仅有 0.549076，故而相对较差。此外，经过 Pearson 的二次筛选后的模型预测精度有了较大提升，各项误差都有所降低。进一步，基于 CEEMD-Pearson 和双隐层深度 LSTM 混合模型的预测结果最好，精度最高，达到了 90%以上，R^2 的值也达到了 0.889593。

结束语 本文利用每小时的 $PM_{2.5}$ 浓度值历史数据对 $PM_{2.5}$ 浓度值的时许演变特性进行了研究和预测。为了充分挖掘 $PM_{2.5}$ 浓度值数据的时序特性，提出一种基于 CEEMD-Pearson 和深度 LSTM 混合模型。实验结果表明：CEEMD 模态分解方法可以展现出历史数据中的隐藏时序特性，基于 CEEMD-LSTM 的混合模型可以有效地对 $PM_{2.5}$ 浓度值进行

精准预测。此外，结合 Pearson 相关性分析进行的二次筛选可有效地提升模型训练的收敛速度和预测精度，基于 CEEMD-Pearson 和 LSTM 的混合模型体现出了更优的训练效果和预测结果。而本文最终提出的基于 CEEMD-Pearson 和深度 LSTM 混合模型明显优于前两种混合模型，可以获得最佳的训练效果、最快的收敛速度以及最精准的预测结果。

除了 $PM_{2.5}$ 浓度值之外，天气因素（如温度、湿度、风速、降水量等）以及其他颗粒物指标（如 CO，NO，SO_2 等）可能也会对 $PM_{2.5}$ 浓度值产生影响。因此，下一步将考虑加入更多的天气因素和其他颗粒物指标数据，以提升 PM2.5 浓度的预测精度。

参 考 文 献

[1] 陈宁，毛善君，李德龙，等. 多基站协同训练神经网络的 $PM_{2.5}$ 预测模型[J]. 测绘科学，2018，43(7)：87-93.

[2] PEREZ P，TRIER A，REYES J. Prediction of $PM_{2.5}$ concentrations several hours in advance using neural networks in Santiago，Chile[J]. Atmospheric Environment，2000，34(8)：1189-1196.

[3] ORDIERES J B，VERGARA E P，CAPUZ R S，et al. Neural network prediction model for fine particulate matter (PM 2.5) on the US-Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua) [J]. Environmental Modelling and Software，2005，20(5)：547-559.

[4] LUIS A，ORTEGA J C，FU J S，et al. A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas：The case of Temuco，Chile[J]. Atmospheric Environment，2008，42(35)：8331-8340.

[5] AL-ALAWI S M，ABDUL-WAHAB S A，BAKHEIT C S. Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone[J]. Environmental Modelling and Software，2008，23(4)：396-403.

[6] WANG Z，LU F，HE H，et al. Fine-scale estimation of carbon monoxide and fine particulate matter concentrations in proximity to a road intersection by using wavelet neural network with genetic algorithm[J]. Atmospheric Environment，2015，104：264-272.

[7] FU M，WANG W，LE Z，et al. Prediction of particular matter concentrations by developed feed-forward neural network with rolling mechanism and gray model[J]. Neural Computing and Applications，2015，26(8)：1789-1797.

[8] YETILMEZSOY K，OZKAYA B，CAKMAKCI M. Artificial intelligence-based prediction models for environmental engineering [J]. Neural Network World，2011，21(3)：193-218.

[9] DONG M，YANG D，KUANG Y，et al. $PM_{2.5}$ concentration prediction using hidden semi-Markov model-based times series data mining [J]. Expert Systems with Applications，2009，36(5)：9046-9055.

[10] KURT A，OKTAY A B. Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks[J]. Expert Systems with Applications，2010，37(12)，7986-7992.

[11] GUPTA P，CHRISTOPHER S A. Particulate matter air quality assessment using integrated surface，satellite，and meteorological products：Multiple regression approach[J]. Journal of Geophysical Research Atmospheres，2009，114(14)：1-13.

[12] GAN K，SUN S，WANG S，et al. A secondary-decomposition-ensemble learning paradigm for forecasting PM2. 5concentration [J]. Atmospheric Pollution Research，2018，9(6)：989-999.

[13] ZHU S，LIAN X，LIU H，et al. Daily air quality index forecasting with hybrid models：A case in China[J]. Environmental Pollution，2017，231(Pt 2).

[14] NIU M，WANG Y，SUN S，et al. A novel hybrid decomposition-and-ensemble model based on CEEMD and GWO for short-term $PM_{2.5}$ concentration forecasting[J]. Atmospheric Environment，2016，134：168-180.

[15] LIU X，LIU Q，ZOU Y，et al. A Self-organizing LSTM-Based Approach to $PM_{2.5}$ Forecast[C]// Cloud Computing and Security (ICCCS 2018). Lecture Notes in Computer Science. Springer，Cham，2018.

[16] HUANG C J，KUO P H. A deep cnn-lstm model for particulate matter (Pm2. 5) forecasting in smart cities[J]. Sensors，2018，18(7)：2220.

[17] LOY-BENITEZ J，VILELA P，LI Q，et al. Sequential prediction of quantitative health risk assessment for the fine particulate matter in an underground facility using deep recurrent neural networks[J]. Ecotoxicology and Environmental Safety，2019，169：316-324.

[18] SOH P W，CHANG J W，HUANG J W. Adaptive Deep Learning-Based Air Quality Prediction Model Using the Most Relevant Spatial-Temporal Relations [J]. IEEE Access，2018，6：38186-38199.

[19] XU Y，YANG W，WANG J. Air quality early-warning system for cities in China[J]. Atmospheric Environment，2017，148：239-257.

[20] 赵雪花，桑宇婷，祝雪萍. 基于 CEEMD-GRNN 组合模型的月径流预测方法[J]. 人民长江，2019(4)：117-123.

[21] 王礼敏，严倩，李寿山，等. 基于双通道 LSTM 模型的用户性别分类方法研究[J]. 计算机科学，2018，45(2)：121-124.

[22] 吕永强，闵巍庆，段华，等. 融合三元卷积神经网络与关系网络的小样本食品图像识别[J]. 计算机科学，2020(1)：1-8.

[23] 曾蒸，李莉，陈晶. 用于情感分类的双向深度 LSTM[J]. 计算机科学，2018，45(8)：213-217，252.

DING Zi-ang, born in 1995, postgraduate. His main research interests include data processing and deep learning.

FU Ming-lei, born in 1981, Ph.D, associate professor. His main research interests include Signal processing, deep learning and intelligent robot.