【数学与应用数学 / Mathematics and Applied Mathematics】

基于 EMD-LSTM 的时间序列预测方法

戴邵武¹, 陈强强^{1,2}, 刘志豪³, 戴洪德⁴

1) 海军航空大学岸防兵学院, 山东烟台 264000; 2) 海军 92728 部队, 上海 200040;
3) 海军 92214 部队, 浙江宁波 315000; 4) 海军航空大学航空基础学院, 山东烟台 264000

摘要: 工程应用中的时间序列多为非线性、非平稳序列, 直接对其进行预测难度较大. 本研究通过经验模态分解算法将原始时间序列分解为多个相对平稳, 并具有不同特征尺度的本征模态函数及趋势项, 在一定程度上降低时间序列的复杂程度; 同时, 在预测过程中, 针对递归神经网络模型难以训练及梯度消失等问题, 引入长短期记忆网络算法. 利用长短期记忆网络算法对分解的本征模态函数分量及趋势项进行分别预测, 叠加预测结果得到最终预测结果. 以中国北京市 PM_{2.5} 浓度为例进行预测分析, 并将本预测算法与单一预测算法进行比较, 结果表明, 所提方法具有更高的模型预测精度, 达到预测要求.

关键词: 数理统计学; 时间序列预测; 经验模态分解; 长短期记忆网络; PM_{2.5}; 机器学习; 时间序列分解
中图分类号: C81 **文献标志码:** A **doi:** 10.3724/SP.J.1249.2020.03265

Time series prediction based on EMD-LSTM model

DAI Shaowu¹, CHEN Qiangqiang^{1,2}, LIU Zhihao³, and DAI Hongde⁴

1) College of Coastal Defense, Naval Aviation University, Yantai 264000, Shandong Province, P. R. China
2) Naval 92728, Shanghai 200040, P. R. China
3) Naval 92214, Ningbo 315000, Zhejiang Province, P. R. China
4) College of Basic Sciences for Aviation, Naval Aviation University, Yantai 264000, Shandong Province, P. R. China

Abstract: The time series in engineering applications are mostly non-stationary and non-linear, which are difficult to be directly predicted. Based on the empirical model decomposition (EMD) method, we decompose the original time series into a number of intrinsic mode functions (IMFs) and trend series with the different feature scales in order to reduce the complexity of time series. Meanwhile, in the prediction process, in order to solve the problems of training difficulty and the gradient disappearance in recurrent neural network (RNN) model, we introduce a long-short term memory (LSTM) network algorithm to predict both the results of the decomposed IMF components and trend series respectively, and obtain the final prediction result by surposing the respective prediction results. Taking the PM_{2.5} concentration in Beijing as an example for prediction and analysis, we compare our prediction algorithm with the single prediction algorithm. The results show that the proposed prediction model has higher accuracy and can meet the prediction requirements.

Key words: mathematical statistics; time series prediction; empirical mode decomposition; long-short term memory network; PM_{2.5}; machine learning; time series decompose

时间序列预测在工程实践中具有重要意义, 通 有针对性地制定相应计划措施, 把握时间序列特
过对时间序列的准确预测, 可对其趋势进行管理, 征, 从而应对未来情形的不确定性^[1].

Received: 2019-03-13; **Revised:** 2019-10-09; **Accepted:** 2019-10-22

Foundation: Natural Science Foundation of Shandong Province (ZR2017MF036); National Defense Science and Technology Foundation of China (F062102009)

Corresponding author: Engineer CHEN Qiangqiang. E-mail: 1195275597@qq.com

Citation: DAI Shaowu, CHEN Qiangqiang, LIU Zhihao, et al. Time series prediction based on EMD-LSTM model [J]. Journal of Shenzhen University Science and Engineering, 2020, 37(3): 265-270. (in Chinese)

传统的时间序列预测方法主要基于数学和统计学基础,通过对序列进行线性回归或最小二乘回归分析,建立预测数据和历史数据之间的联系.自回归移动平均(auto regression moving average, ARMA)模型作为经典的时间序列预测模型,具有严谨的理论支撑,发展较为成熟. ARMA 模型从数学意义上可看做自回归(auto regression, AR)模型与移动平均(moving average, MA)模型的组合^[2],其扩展形式还有自回归积分移动平均(autoregressive integrated moving average, ARIMA)模型等^[3].传统的预测分析方法模型简单,理论精确,但针对复杂非线性非平稳时间序列的预测存在一定精度限制.

神经网络理论的成熟完善,使其在工程中取得重大应用.一些浅层的神经网络,如支持向量回归(support vector regression, SVR)在时间序列预测中具有广泛应用^[4];LI 等^[5]采用极限学习机(extreme learning machine)对经济学中的时间序列进行预测;丁亮等^[6]利用 BP(back propagation)神经网络,实现自然方面的时间序列预测.浅层神经网络的局限之处在于模型过于简单,在实际训练中,模型精度制约了预测精度的提升.递归神经网络(recurrent neural network, RNN)作为一种循环反馈的神经网络框架,能够充分考虑时间序列中的相关性,通过分析历史信息,建立完整的时间序列模型^[7].但在使用过程中,RNN 容易出现梯度消失及梯度爆炸等问题,而长短期记忆(long-short term memory, LSTM)网络作为一种特殊的 RNN 模型,能够有效规避这些问题,实现对历史信息真实有效的利用^[8].

复杂非线性非平稳时间序列在预测模型的构建中具有难度较大及模型不准确的问题,本研究通过经验模态分解(empirical mode decomposition, EMD)方法对原信号进行处理,以降低时间序列的复杂度;针对 EMD 分解后的本征模态函数(intrinsic mode function, IMF)分量及趋势项,采用精度更高的 LSTM 模型分别进行预测;通过对预测结果进行汇总后,得到原始时间序列的预测结果.

1 长短期记忆网络

1.1 递归神经网络原理

递归神经网络由输入层、隐藏层及输出层构成,通过将上一时刻的状态信息传递到当前状态,从而将时间序列展开为一系列具有相互性的神经元,其结构如图 1^[9].其中,节点表示每一时间点

上的神经元; U 为输入层与隐藏层之间的权重连接; W 为隐藏层与神经元之间的权重连接; V 为隐藏层与输出层之间的权重连接.随着时间累加,信息在传递过程中不断损失,初始信息能力逐渐弱化,将会出现梯度消失等问题,也就丧失对时间序列的长时间处理能力.

图 1 递归神经网络神经元^[9]

Fig. 1 Structure of recurrent neurons^[9]

1.2 LSTM 原理

针对 RNN 训练过程中的梯度消失问题, LSTM 网络被提出并不断完善. LSTM 的主要理论是设计记忆模块,该模块可受控于多个控制门,从而实现长时间记忆功能^[10]. LSTM 的结构如图 2.

图 2 长短期记忆网络

Fig. 2 Long-short term memory network

LSTM 网络更新^[11-12]为: 设 h 为 LSTM 单元输出, c 为 LSTM 记忆模块值, x 为单元输入, W_{xc} 及 W_{hc} 分别为输入数据及上一时刻的输出权值, 时刻 t 的记忆单元值为

$$\tilde{c}_t = \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (1)$$

输入门可控制当前数据输入对记忆模块状态的影响, 表示为

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (2)$$

遗忘门可控制历史学习对当前记忆模块状态值的影响, 表示为

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (3)$$

当前时刻记忆单元状态值为

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (4)$$

其中, 符号 $\odot$ 表示逐点乘积. 输出门可控制记忆模块状态值的输出, 表示为

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (5)$$

LSTM 网络单元输出为

$$h_t = o_t \tanh c_t \quad (6)$$

2 EMD-LSTM 算法

时间序列的预测需要对训练数据建立准确的预测模型, 从而建立时间序列内部关系, 为后续预测过程提供准确依据. 工程应用中的绝大多数时间序列模型具有典型的非线性、非平稳性特征, 针对这些复杂性较高的时间序列, 若直接建立预测模型, 很难达到预测精度要求. 本研究采用 EMD-LSTM 算法, 通过对复杂非平稳时间序列进行 EMD 分解, 降低时间序列的复杂程度, 从而完成对复杂程度较低 IMF 分量的 LSTM 建模.

EMD 是一种新型的信号分解方法, 在非线性、非平稳信号的处理中具有良好效果^[13]. EMD 算法将复杂信号分解为多个 IMF 和 1 个趋势项(trend), 其表达式为

$$s_t = \sum_{i=1}^n \text{imf}_i(t) + r_n(t) \quad (7)$$

其中, s_t 为原始信号; $\text{imf}_i(t)$ 为第 i 个 IMF 分量; $r_n(t)$ 为趋势项.

在分解过程中, IMF 分量的取值标准为: ① 信号中零点数和极值点数相等或至多相差 1; ② 极大值和极小值包络线的均值相等且为 0.

采用 EMD 方法进行时间序列分解后, 其误差量级为 1×10^{-14} , 可以完整表达出时间序列包含的全部信息, 分解后的各 IMF 分量及趋势项组合, 可以完全拟合原时间序列. 对 IMF 分量进行特性分析可知, 随着分解次数的增加, IMF 的阶数不断增加, 相应 IMF 分量的波动频率不断降低, 即低阶 IMF 分量对应高频信号. 利用 EMD 算法处理原时间序列, 可以准确得到时间序列在不同时间段的波动特征, 简化时间序列预测模型, 在一定程度上提高时间序列预测精度.

文献[14]通过经验模态分解与支持向量机相结合的方法, 对氢原子钟差序列进行预测分析; 然而, 对于工程实践中大部分复杂时间序列, 其低阶(高频) IMF 分量在平稳性及复杂性上虽然比原时间序列具有一定优化, 但仍具有很强的波动性, 此

时利用普通的支持向量回归难以取得满意的预测效果. 低阶 IMF 分量仍保留时间序列的部分信息, 不能通过简单的舍弃处理进行降噪. 本研究在经验模态分解基础上, 针对相对平稳但仍具有较强波动性的 IMF 分量, 构建 LSTM 网络预测模型, 提高了预测模型的准确性及预测精度. 根据 EMD 分解的完备性可知, 分别预测后的结果符合时间序列完备性的需求. EMD-LSTM 预测的整体流程图如图 3.

图3 EMD-LSTM 预测流程图

Fig. 3 Flow chart of EMD-LSTM

图3左半部分为直接使用 LSTM 进行预测的结果. 本研究预测过程具体步骤为: ① 将待分析的时间序列进行经验模态分解, 得到分解后各分量 ($\text{imf}_1, \text{imf}_2, \dots, \text{imf}_n$) 及趋势项 r_n ; ② 对所有 IMF 分量及趋势项分别建立 LSTM 预测模型, 并与 SVR 预测模型进行对比验证; ③ 对各个分量进行预测; ④ 将预测值累加得到最终预测值; ⑤ 分析预测结果, 与文献[14]中提出的 EMD-SVR 方法对比, 验证模型可行性.

3 仿真分析

以 2014-05-13 至 2015-09-25 中国北京市 $\text{PM}_{2.5}$ 浓度数据(来自网站公开数据^[15])为例进行仿真分析. 受供暖等环境因素影响, 原始数据序列的波动较大, 复杂程度较高. 气象局记录数据间隔为 1 h, 且有部分缺失. 数据处理时将以天(d)为单位对每天的 $\text{PM}_{2.5}$ 浓度做平均值处理, 同时补充缺失值, 共得到 501 个数据点^[16], 如图 4.

由图 4 可见, $\text{PM}_{2.5}$ 浓度的原始时间序列呈现很强的非线性与非稳定性, 序列的复杂程度较高. 采用 EMD 分解方法处理原始 $\text{PM}_{2.5}$ 数据序列, 以降低序列复杂程度, 结果如图 5.

针对每个 IMF 分量及趋势项, 构建预测模型.

图 4 PM_{2.5} 浓度数据序列Fig. 4 The data series for the concentration of PM_{2.5}

图 5 原始序列的 EMD 分解结果

Fig. 5 The decomposition result of original series by EMD

选取前 450 个数据作为训练集, 余下 51 个数据作为测试集, 以均方根误差 (root mean squared error, RMSE) 及平均绝对误差 (mean absolute error, MAE)

为指标, 衡量 EMD-LSTM 算法对时间序列的预测能力^[17].

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (X_{\text{real},i} - X_{\text{pre},i})^2}{n}} \quad (8)$$

$$\text{MAE} = \frac{\sum_{i=1}^n |X_{\text{real},i} - X_{\text{pre},i}|}{n} \quad (9)$$

其中, $X_{\text{real},i}$ 为数据序列中第 i 个数据的真实值; $X_{\text{pre},i}$ 为第 i 个数据的预测值; n 为数据点个数.

排列熵作为一种检测时间序列随机性和动力学突变行为的方法, 具有计算简单、抗噪声能力强等优点. 因此, 可以通过排列熵来衡量信号复杂度^[18], 排列熵值越大, 其对应时间序列复杂程度越高. 根据 EMD 分解得到各 IMF 分量的排列熵如图 6.

图 6 序列的排列熵

Fig. 6 Permutation entropy of series

根据图 5 中 EMD 分解后的 IMF 序列及趋势项, 分别构建 LSTM 预测模型. 为对比验证, 构建各自对应的 SVR 预测模型, 选择 RBF 核函数及线性核函数进行时间序列预测, 利用交叉验证方法确定核函数参数及惩罚参数. 两种方法下各 IMF 分量及趋势项的 RMSE 及 MAE 计算结果如表 1.

由表 1 及图 6 可见, 随着 IMF 阶数的增加, 各分量对应的排列熵逐渐降低, 预测误差随之减小, 证明了 EMD 算法对于降低 IMF 序列复杂程度的合理性. 在表 1 中, 由于趋势项的时间序列过于简单, 因此, 未进行区分对比预测. 对 IMF 分量采用 LSTM 与 SVR 方法对比, 结果表明, 对于高阶 IMF 分量, 随着时间序列复杂程度的降低, 两种预测方法的差别逐渐降低, 相比 SVR 预测模型, LSTM 预测模型的优越性不断降低; 对于低阶 IMF, LSTM 表现出良好的预测效果, 预测精度较 SVR 有明显提

高. 结合对 EMD 算法完备性的分析, 分解得到的所有子序列重构之后, 与原时间序列的误差量级为 1×10^{-14} . 所以, 将表 1 中各序列的预测结果相加, 即可得到原始 $PM_{2.5}$ 数据序列的 EMD-LSTM 和 EMD-SVR 预测结果, 如图 7. 可见, EMD-LSTM 预测结果更贴近真实值.

表 1 各序列的预测误差
Table 1 Prediction errors of series

序列	LSTM		SVR	
	RMSE	MAE	RMSE	MAE
imf ₁	15.883 1	12.814 8	17.025 6	11.630 7
imf ₂	6.259 6	3.826 3	13.252 7	10.152 2
imf ₃	0.429 2	0.313 8	3.978 6	3.573 3
imf ₄	0.223 5	0.177 6	2.881 9	2.615 0
imf ₅	0.153 6	0.126 1	1.242 8	1.171 7
imf ₆	0.094 5	0.081 0	1.816 0	1.115 7
imf ₇	0.063 1	0.062 9	0.213 8	0.167 0
trend	0.113 4	0.109 7	0.113 4	0.109 7

图 7 真实值及对比预测结果
Fig. 7 The original data and the contrast of prediction results

为验证 EMD-LSTM 算法的准确性, 采用单个 SVR 模型及 LSTM 模型对原始 $PM_{2.5}$ 时间序列建立预测模型, 并进行预测分析. 将图 7 中得到的预测结果用数值形式进行分析, 其 RMSE 及 MAE 指标如表 2.

由表 2 可见, 由于原始 $PM_{2.5}$ 数据序列的非线性及非平稳性, 单一预测方法构建的预测模型不能对其进行准确预测, 且由于序列的复杂程度较高, 单一 LSTM 方法劣于 SVR 方法的预测效果; 在利用

EMD 算法对时间序列进行分解之后, 一定程度上降低了原序列的复杂程度, 此时, 对分解得到的 IMF 分量及趋势项采用 LSTM 模型和 SVR 模型进行预测分析, 其结果比单一 LSTM 及 SVR 模型均有显著提高, 证明利用 EMD 方法对原始时间序列进行分解的可行性. 相比 EMD-SVR 方法而言, 采用 EMD-LSTM 方法预测结果的 RMSE 降低 30.95%, MAE 降低 26.05%, 预测精度得到明显提高.

表 2 四种方法的预测指标
Table 2 Prediction errors of four algorithms

预测方法	RMSE	MAE
LSTM	40.624 0	31.909 5
SVR	29.016 4	25.113 8
EMD-SVR	21.518 2	16.941 6
EMD-LSTM	14.858 1	12.527 8

结 语

工程实践中的时间序列具有典型的非线性及非平稳特性, 此时传统时间序列预测方法存在预测模型不精确及预测精度低等缺点, 直接采用机器学习方法进行预测时, 在预测精度方面也存在一定限制, 影响时间序列的预测精度. 本研究在对北京市 $PM_{2.5}$ 数据序列预测过程中, 通过 EMD 方法对原始数据序列进行分解, 得到与原始数据序列相关的 IMF 分量及趋势项, 有效降低原始数据序列的非线性及非平稳性; 对分解得到的 IMF 分量及趋势项分别建立对应预测模型, 并将所得结果进行汇总分析, 结果认为: ① 以排列熵作为时间序列复杂程度指标, 通过 EMD 方法对原始数据序列进行分解, 可有效降低原始序列的复杂程度, 有利于建立更精确的预测模型; ② 与单一的预测算法对比, 基于 EMD 的预测算法能够有效提高预测模型的准确性, 从而提高预测精度; ③ 针对分解后的时间序列, 与 SVR 模型相比, 采用深度学习理论中的 LSTM 模型具有较高的准确性, 深度学习理论为时间序列预测提供理论支撑, 在工程实践中具有潜在应用价值.

基金项目: 山东省自然科学基金资助项目 (ZR2017MF036); 国防科技基金资助项目 (F062102009)

作者简介: 戴邵武 (1966—), 海军航空大学教授. 研究方向: 飞行器综合导航、时间序列分析. E-mail: 13386386392@163.com

引 文: 戴邵武, 陈强强, 刘志豪, 等. 基于 EMD-LSTM 的时间序列预测方法 [J]. 深圳大学学报理工版, 2020, 37 (3): 265-270.

参考文献 / References:

- [1] 陈艳华. 基于机器学习的时间序列预测方法研究与应用[D]. 兰州: 兰州大学, 2017.
CHEN Yanhua. The research and applications of time series forecasting methods based on machine learning [D]. Lanzhou: Lanzhou University, 2017. (in Chinese)
- [2] 李 祥, 彭 玲, 邵 静, 等. 基于小波分解和 ARMA 模型的空气污染预报研究[J]. 环境工程, 2016, 34(8): 110-113.
LI Xiang, PENG Ling, SHAO Jing, et al. Air pollution forecast based on wavelet decomposition and ARMA model [J]. Environmental Engineering, 2016, 34(8): 110-113. (in Chinese)
- [3] 牟敬锋, 赵 星, 樊静洁, 等. 基于 ARIMA 模型的深圳市空气质量指数时间序列预测研究[J]. 环境卫生学杂志, 2017, 7(2): 102-107.
MOU Jingfeng, ZHAO Xing, FAN Jingjie, et al. Time series prediction of AQI in Shenzhen based on ARIMA model [J]. Journal of Environmental Hygiene, 2017, 7(2): 102-107. (in Chinese)
- [4] 王 威, 陈熙源. 一种基于多尺度和改进支持向量机的光纤陀螺温度漂移建模与补偿方法 [J]. 中国惯性技术学报, 2016, 24(6): 793-797.
WANG Wei, CHEN Xiyuan. Modeling and compensation method of FOG temperature drift based on multi-scale and improved support vector machine [J]. Journal of Chinese Inertial Technology, 2016, 24(6): 793-797. (in Chinese)
- [5] LI Guohui, MA Xiao, YANG Hong. A hybrid model for monthly precipitation time series forecasting based on variational mode decomposition with extreme learning machine [J]. Information, 2018, 9(7): 1-13.
- [6] 丁 亮, 武 林, 刘清民. 基于边缘保持中值滤波和 BP 神经网络的自然手势识别研究 [J]. 现代计算机, 2017(25): 10-13.
DING Liang, WU Lin, LIU Qingmin. Research on natural gesture recognition based on edge preserving median filter and BP neural network [J]. Modern Computer, 2017(25): 10-13. (in Chinese)
- [7] LUKOŠEVIČIUS M, JAEGER H. Reservoir computing approaches to recurrent neural network training [J]. Computer Science Review, 2009, 3(3): 127-149.
- [8] 毛景慧. 基于 LSTM 深度神经网络的股市时间序列预测精度的影响因素研究[D]. 广州: 暨南大学, 2017.
MAO Jinghui. Research on influencing factors of stock market time series prediction accuracy based on LSTM deep neural network [D]. Guangzhou: Jinan University, 2017. (in Chinese)
- [9] LEV G, SADEH G, KLEIN B, et al. RNN fisher vectors for action recognition and image annotation [J]. Computer Science, 2016: 833-850.
- [10] FU Rui, ZHANG Zuo, LI Li. Using LSTM and GRU neural network methods for traffic flow prediction [C]// The 31st Youth Academic Annual Conference of Chinese Association of Automation. Wuhan, China: IEEE, 2017: 324-328.
- [11] ZHANG Boning. Foreign exchange rates forecasting with an EMD-LSTM neural networks model [J]. Journal of Physics Conference, 2018, 1053(1): 5-12.
- [12] 杨训政, 柯余洋, 梁 肖, 等. 基于 LSTM 的发电机组污染物排放预测研究[J]. 电气自动化, 2016, 38(5): 22-25.
YANG Xunzheng, KE Yuyang, LIANG Xiao, et al. A study on the prediction of generator set pollutant emissions based on LSTM [J]. Electrical Automation, 2016, 38(5): 22-25. (in Chinese)
- [13] HUANG M L, LONG S E, SHIH H H, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis [J]. Proceedings: Mathematical, Physical and Engineering Sciences, 1998, 454(1971): 903-993.
- [14] HUANG Yiming, WU Di, ZHANG Zhifen, et al. EMD-based pulsed TIG welding process porosity defect detection and defect diagnosis using GA-SVM [J]. Journal of Materials Processing Technology, 2017, 239: 92-102.
- [15] 秦喜文, 刘媛媛, 王新民, 等. 基于整体经验模态分解和支持向量回归的北京市 PM_{2.5} 预测[J]. 吉林大学学报地球科学版, 2016, 46(2): 563-568.
QIN Xiwen, LIU Yuanyuan, WANG Xinmin, et al. PM_{2.5} prediction of Beijing city based on ensemble empirical mode decomposition and support vector regression [J]. Journal of Jilin University Earth Science Edition, 2016, 46(2): 563-568. (in Chinese)
- [16] 刘媛媛. 基于整体经验模态分解和随机森林的城市 PM_{2.5} 预测[D]. 长春: 长春工业大学, 2017.
LIU Yuanyuan. A hybrid model for PM_{2.5} forecasting of cities based on ensemble empirical mode decomposition and a random forest [D]. Changchun: Changchun University of Technology, 2017. (in Chinese)
- [17] WILLMOTT C J, MATSUURA K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance [J]. Climate Research, 2005, 30(1): 79-82.
- [18] 冯辅周, 饶国强, 司爱威, 等. 排列熵算法的应用与发展[J]. 装甲兵工程学院学报, 2012, 26(2): 34-38.
FENG Fuzhou, RAO Guoqiang, SI Aiwei, et al. Application and development of permutation entropy algorithm [J]. Journal of Academy of Armored Force Engineering, 2012, 26(2): 34-38. (in Chinese)

【中文责编: 方 圆; 英文责编: 溯 心】