李昊坤. 基于 EMD-LSTM 模型的植被净初级生产力预测[J]. 智能计算机与应用,2025,15(9):64-68. DOI:10.20169/j.issn.2095-2163.250910

基于 EMD-LSTM 模型的植被净初级生产力预测

李昊坤
(山东农业大学 农学院, 山东 泰安 271018)

摘 要: 为研究在气候变化与人类活动影响下植物的净初级生产率(Net Primary Productivity, NPP)并加以预测,论文首先收集 2000~2020 年的 NPP 数据,并对数据集进行预处理;进而构建起了基于时间顺序的 BP 神经网络预测模式和 EMD-BP 模型预测模式,并利用 2 种模式对 NPP 及其影响因素进行了分类与预测;最后,再将 2 种预测模式的实测结果加以了比较。结果显示,EMD-BP 神经网络模型在估计 NPP 方面的偏差,较 BP 神经网络模型更小。

关键词: 净初级生产力; BP 神经网络; EMD-BP 预测模型
中图分类号: Q948 **文献标志码**: A **文章编号**: 2095-2163(2025)09-0064-05

Prediction of Net Primary Productivity of plants based on EMD-LSTM model

LI Haokun
(College of Agriculture, Shandong Agricultural University, Tai'an 271018, Shandong, China)

Abstract: To study the Net Primary Productivity (NPP) of plants under the influence of climate change and human activities and to predict it, the NPP data from 2000 to 2020 are firstly collected and the dataset is preprocessed; then the time-order-based BP neural network prediction model and the EMD-BP prediction model are constructed, and the NPP and its influencing factors are classified and predicted using the two models; finally, the measured results of the two prediction models are compared. The results show that the EMD-BP neural network model has less bias in estimating NPP than the BP neural network model.

Key words: Net Primary Productivity; BP neural network; EMD-BP prediction model

0 引 言

植被净初级生产力(Net Primary Productivity, NPP)是指在单位面积、单位时间中绿化植株由于光合作用形成的有机质数量减去绿化植物因自养呼吸所耗费能源后的存留部分[1],也可以被称为净第一性生产率。一般来说,植物 NPP 代表了植物通过吸取和运用光合效应获得物质的能力,对于评价整个生态系统的质量情况以及植物种群的生产能力等方面都具有很重要的意义[2]。NPP 是由植物本身的生物特征和外部环境因素共同影响的结果,是对世界碳循环过程进行生物质交流的重要物质基础,同时也是评估世界或地区生态安全的重要性指标[4-10]。植物 NPP 是整个陆地生态系统碳循环的主要原动力,为整个生态系统的二次生长提供了材料和能源,同时还反映了植物和外部条件之间的相互作用过程。所以,植物 NPP 已成为研究天气的演变以及对陆地生态系统的改变中的一个关键因素和必须重视的内容,对进行陆地生态系统植物干净初级生产力的研究、探讨地区碳循环作用、保障陆地生态系统的稳定和平衡,有着重大价值[11-14]。

因此,本文结合机器学习方法预测植被净初级生产力,提出了集成经验模态分析(经验模态分解, EMD)与长短期记忆神经网络(长短时记忆法)的植被净初级生产力预测方法,有效提高了预测的准确度。

1 EMD 和 BP 神经网络基本原理

1.1 EMD 基本原理

EMD (Empirical Mode Decomposition)是一种时间-频率分析技术,可以用于将非平稳信号分解成若干个平稳信号。EMD 基于经验模态,按照信号的时间和频率特征,通过迭代对信号进行分解和重组,提取信号的基本特征和关键模态。这些基本特征通

作者简介: 李昊坤 (2002—),男,本科生,主要研究方向:种子科学与工程。Email:280790413@qq.com。

收稿日期: 2023-12-22

常被认为是信号的主要成分，可以用于进一步的分析和处理。EMD 的优点在于可以分解出信号的非平稳特征，并且不需要对信号的时域和频域的特征进行任何假定。由于 EMD 分解可以分解出信号的非平稳特征，因此可以捕捉信号中隐藏的相关性和趋势。通过使用 EMD 分解得到的信息，就能对时间序列进行更精确和有效的预测。

EMD 分解是一种用于 NPP 时间分布预测的方法。其初衷是将复杂的 NPP 时间序列分解成多个本征模态函数(IMF)，每个 IMF 代表不同尺度上的信号成分，以揭示 NPP 时间分布中的局部特征和趋势。这样可以提供更精确的时间序列成分，为后续的 NPP 预测建模提供更有价值的特征。

1.2　BP 神经网络基本原理

BP 算法是一个监测学习结果的算法。算法通过使用均方误差和梯度变化的方式，来对网络连接权重进行调整。对网络权重的调整主要是为了达到最小的误差平方率。在这种方法中，先给网络的最大连接系数设置一个小值，随后再选择某个训练样本，从而得到了这个相应数据的最大误差梯度弥散。BP 神经网络在 NPP 时间分布预测中的核心作用是利用神经网络的强大非线性建模能力，以自动学习复杂的 NPP 时间序列数据的关联，从而实现对未来 NPP 的准确预测。其目的是通过训练神经网络来最小化预测误差，以提高预测精度。

BP 模型由输入层、隐藏层和输出层组成。每个神经元在隐藏层和输出层都有相应的权重和偏差，通过激活函数(通常为 S 型函数)进行非线性变换。首先是数据预处理过程，将 NPP 时间序列数据进行标准化，然后搭建具有输入层、隐藏层和输出层的神经网络。在前向传播阶段，通过计算每个神经元的输出，从输入到输出逐层传播。损失函数通常采用均方误差(MSE)来衡量预测值与真实值之间的差异。接下来，在反向传播阶段，根据损失函数的梯度，利用梯度下降算法更新神经网络的权重和偏差。这一过程迭代多次，直到损失收敛或达到指定的迭代次数。

BP 的调整方法可说明如下。

(1)操作信息的前向传送：输入消息由输入层通过隐藏层传送到出口层。在操作信息的前向传输流程中，由于网络的权重值与位移值都维持恒定，因此各级神经元的状况都只对下一级神经元的状况产生影响。但若在输出层上没有实现预期的传输，则可以转换至误差消息的反向传输。

(2)错误信息的正反向传输：网络的真实输入输出与理想输入输出间的偏差被确定为错误信息；在错误信息的正反向传输中，错误信息以逐层的形态由输入输出端反馈给输入层。在错误信息的正反向传递过程中，系统的权重值由错误反馈来控制。而经过对权重值和偏离值的进一步调整，可以使得系统的真实数据更接近于预测的。

指导 BP 网络学习规律的基本思路是：对网络的权重值和阈值的调整都要按负梯度方式执行，反映函数的最快下降。这里需用到的公式为：

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_k \boldsymbol{g}_k \tag{1}$$

其中，$\boldsymbol{x}_k$ 表示当前权重值和阈值的矩阵；$\boldsymbol{g}_k$ 表示当前函数的梯度；η_k 表示学习率。以 3 层 BP 网络为例，假设其输入节点为 x_i，隐藏层的节点为 y_j，输出层的节点为 z_l。输入节点与隐藏层节点之间的网络权重值为 w_{ji}，隐藏层节点与输出层节点之间的网络权重值为 v_{lj}。当输出节点的期望值为 tl 时，$f(-)$ 为激活函数。该模型的计算公式具体如下。

(1)前向传播：计算网络的输出、隐藏层节点的输出，可由如下公式计算求得：

$$y_j = f\left(\sum_i w_{ji} x_i - \theta_j\right) = f(\mathrm{net}_j) \tag{2}$$

$$\mathrm{net}_j = \sum_i w_{ji} x_i - \theta_j \tag{3}$$

(2)输出节点的计算输出。可由如下公式计算求得：

$$z_l = f\left(\sum_i v_{lj} y_j - \theta_l\right) = f(\mathrm{net}_l) \tag{4}$$

$$\mathrm{net}_l = \sum_j v_{lj} y_j - \theta_l \tag{5}$$

(3)输出节点的误差。可由如下公式计算求得：

$$E = \frac{1}{2}\sum_l (t_l - z_l)^2 = \frac{1}{2}\sum_l \left(t_l - f\left(\sum_j v_{lj} y_j - \theta_l\right)\right)^2 = \frac{1}{2}\sum_l \left(t_l - f\left(\sum_j v_{lj} f\left(\sum_i w_{ji} x_i - \theta_j\right) - \theta_l\right)\right)^2 \tag{6}$$

在人工神经网络的实际应用中，虽然大部分神经网络模型都使用了 BP 网络的变化模型，但这并不代表 BP 网络是完整的，该方法也存在不容忽视的问题。比如，在训练过程中的局部变化最小部分，且收敛速率较慢，网络往往有较多的冗余，新加入的样本可能影响学习的样本，或者其他。研究人员提出了许多改进的算法来解决这些缺陷。其改进方法一般可分为 3 类。一是提高神经网络训练的速度；二是提高训练的精度；三是避免陷入局部的最小点。

在这些方法中,比较典型的是附加动量法和可变学习率法。本文实验中采用了附加动量法与可变学习率法。

2 实证分析

2.1 数据来源与处理

本文使用的甘肃省 NPP 数据获取自美国国家航空航天局网站、美国国家海洋和大气管理局网站和美国国家环境信息中心网站,使用 BIOME-BGC 生态系统模型、基于 MODIS/TERRA 卫星参数模拟得到全球植被净初级生产力数据集。该数据集已被广泛应用于全球或区域植被生产力研究,具备科学性。本研究使用的数据时间跨度为 2000 年至 2020 年,空间分辨率为 500 m, 时间分辨率为月。

首先进行数据集的预处理。由于原始数据为 hdf 格式,需要借助 MRT 工具完成 hdf 格式文件的批量拼接、重采样和重投影。然后在数据集上乘以比例因子 0.1,降低数据集的数值范围,以便于后续处理和分析。使用聚类方法将数据点进行分组,对每组数据点计算其中心点和标准差,将每个数据点与其所属组的中心点进行比较,识别出异常值并将其赋值为空值。为了消除数据中的噪声,提高数据质量和可靠性,采用高斯滤波的方法进行数据去噪声。定义高斯核,将原始数据与高斯核进行卷积运算,将原始数据中的噪声降到最小。

2.2 评价指标

本文实验选用的评价指标为 RMSE、MAE、SMAPE、MAD 与 PCC,在 5 个不同层面对于预测效果进行刻画。

(1)均方根偏差(Root-Mean-Square Deviation, RMSD)或均方根误差(Root-Mean-Square Error, RMSE)是一个经常使用的衡量模型或估计器预测的值(样本或群体值)与观察到的值之间的差异。RMSE 是准确度的衡量标准,用于比较不同模型对特定数据集的预测误差,而不是在数据集之间进行比较,因为该值与尺度有关。其公式为:

$$\mathrm{RMSE}(y,x)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-x_i)^2} \tag{7}$$

(2)均值绝对误差(Mean Absolute Error, MAE)在统计上是指代表同一事件的成对观察值间偏差的计算。Y 与 X 的例子包括预测与观察的比较、后续时间与初始时间的比较、以及一种测量技术与另一种测量技术的比较。MAE 的计算方法是绝对误差之和除以样本量,可以表示为:

$$\mathrm{MAE}(y,x)=\frac{\sum_{i=1}^{n}|y_i-x_i|}{n} \tag{8}$$

(3)对称平均绝对百分比误差(Symmetric Mean Absolute Percentage Error , SMAPE)是一个基于百分比(或相对)误差的精度测量。y_i 和 x_i 之间的绝对差值除以实际值 y_i 和预测值 x_i 的绝对值之和的一半。这个计算值对每个拟合点 i 进行求和,再除以拟合点的数量 n。通常被定义为:

$$\mathrm{SMAPE}=\frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i-x_i|}{(|x_i|+|y_i|)/2} \tag{9}$$

(4)中位数绝对偏差(Median Absolute Deviation, MAD)在统计学中是对定量数据单变量样本的变异性的一种稳健测量。也可以指由样本计算的 MAD 所估计的群体参数。对于一个单变量数据集 $X_1,X_2,\cdots,X_n$, MAD 被定义为与数据中位数的绝对偏差的中值,定义公式如下:

$$\mathrm{MAD}=\mathrm{median}(|X_i-\tilde{X}|) \tag{10}$$

也就是说,从数据的中位数的残差(偏差)开始,MAD 是其绝对值的中位数。

(5)皮尔逊相关系数(Pearson Correlation Coefficient, PCC)的统计学中是 2 组数据之间线性相关的测量。皮尔逊相关系数是 2 个变量的协方差与其标准差的乘积之间的比率;因此,基本上是协方差的归一化测量,这样的结果总是有一个在-1 和 1 之间的值。与协方差本身一样,该测量只能反映变量的线性相关,而忽略了许多其他类型的关系或相关性。其定义为:

$$\mathrm{PCC}(X,Y)=\frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y} \tag{11}$$

其中, $\mathrm{cov}(\cdot)$ 表示协方差计算公式,σ_i 表示 i 的标准差。

2.3 植被净初级生产力的 EMD 分解

利用 EMD 法对 20 年的植被净初级生产力进行分解,由分解结果可知 3 个 EMD 分量均呈现频率降低、波长变长、振幅变小的特征。对各 NPP 的分解结果,做阐释分述如下。

(1)EMD 分量 1 (见图 1(a)):该分量展示了最高频率的变动,具有很强的周期性。由于该分量代表了最高频率的变动,这可能反映了 NPP 数据中的短期波动或可能是某些外部因子,如天气变化、土壤湿度等,对植物净初级生产力的短期影响。

(2)EMD 分量 2 (见图 1(b)):该分量显示了中频率的变动,波动幅度减小,并且周期性有所扩

展。这可能代表了植物净初级生产力中的季节性影响，例如随着季节更替导致的生态系统内生物活动的增减。

(3)EMD 分量 3（见图 1(c)）：该分量代表了更低的频率变动，展现了一个平滑的趋势，波动非常小。这可能表示了植物净初级生产力中的长期趋势或背景变化，比如由于全球变暖、土地利用变化或其他长期环境因子造成的影响。

综上所述，EMD 的分解明确地揭示了 NPP 数据中的不同频率变动。从最高的频率到最低的频率，每个分量都代表了不同的时间尺度上的生态和环境影响。最高频率的变动可能与日常或月度的环境因素有关、如降雨和温度。中频率可能关联到季节性因子，如温度和降水模式的季节性变化。而最低频率的变动则可能与多年或十年的长期环境趋势或地区性变化有关。这种分解方法为研究者们提供了一个深入了解植物净初级生产力的动态和驱动因子的工具，有助于更好地理解不同时间尺度上的生态变化和其背后的原因。

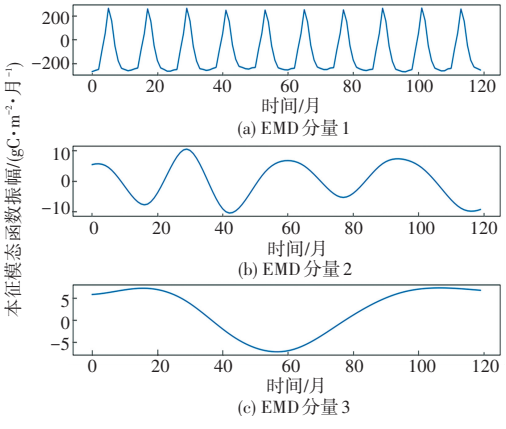

图 1　EMD 分解结果

Fig. 1　EMD decomposition results

2.4　预测结果分析

本文采用随机分割的方法，将数据分为训练集和测试集，然后进行预测。模型对比结果如图 2 所示。由图 2 可以发现，BP 神经网络、EMD-BP 预测模型的预测结果与 NPP 实际的走向情况大体相符，误差也较小，而且很明显地 EMD-BP 模型的拟合能力更佳；由图 2 中还可发现 EMD-BP 预测模型的预测值与 NPP 实际数值同时最符合的，而且预测值与 NPP 实际数值之间的偏差也很小。

不同模型预测结果对比见表 1。由表 1 可以看出，EMD-BP 模型所预测的 RMSE、MAE、SMAPE、MAD、PCC 等的数据，与 BP 神经网络的模型相比显然都更小，大小依次为 79.022、43.324、22.746、0.981、0.981，说明了 BP 神经网络的预测结果比较精确，由此也再次表明了 BP 神经网络对该数据的检测结果，相比于 LSTM 神经网络更佳。

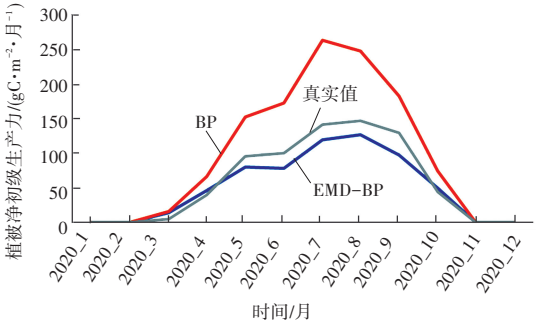

图 2　模型对比结果

Fig. 2　Comparison results of the models

表 1　不同模型预测结果对比

Table 1　Comparison of prediction results of different models

模型	平均 RMSE	平均 MAE	平均 SMAPE	平均 MAD	平均 PCC
BP 模型	141.403	75.325	28.474	4.443	0.991
EMD-BP 模型	**79.022**	**43.324**	**22.746**	**0.981**	**0.981**

综上所述，EMD-BP 模型的模拟预测效果都比单一 BP 模型的要好，这进一步说明将其用于非线性、非平稳的径流序列预测是可行、有效的。

3　结束语

全球气候变化逐渐影响到生态系统的平和稳定，这一现象已日渐成为各学科进行学术研究的重要背景。通过研究甘肃省 NPP，能为全球气候变化及环境保护提供重要依据。本文采用 BP 算法和 EMD-BP 算法，对 NPP 进行预测。通过对这 2 种算法的对比，可以看出 EMD-BP 预测模型可以较好地预测 NPP。

参考文献

[1] 许静，陈迪，李文龙. 基于光能利用率模型的甘南州植被净初级生产力研究[J]. 草业科学，2019，36(10)：2455-2465.
[2] 登科，范建忠，董金芳. 1981～2000 年陕西省植被净初级生产力时空变化[J]. 西北植物学报，2011，31(9)：1873-1877.
[3] 刘真真，张喜旺，陈云生. 基于 CASA 模型的区域冬小麦生物量遥感估算[J]. 农业工程学报，2017，33(4)：225-233.
[4] 赵国帅，王军邦，范文义. 2000-2008 年中国东北地区植被净初级生产力的模拟及季节变化[J]. 应用生态学报，2011，22(3)：621-630.
[5] 徐勇，郑志威，戴强玉，等. 顾及时滞效应的西南地区植被 NPP 变化归因分析[J]. 农业工程学报，2022，38(9)：297-305.
[6] 杨丹，王晓峰. 黄土高原气候和人类活动对植被 NPP 变化的影

响[J]. 干旱区研究,2022,39(2):584-593.
[7] 张黎明,蔡琦,宋梅村. 基于 RBF 神经网络的 NPP 运行状态趋势预测[J]. 原子能科学技术,2013,47(11):2103-2107.
[8] 韩红珠. 黄土高原植被物候和净初级生产力(NPP)的关系及其对气候变化的响应[D]. 西安:陕西师范大学,2020.
[9] 杨晶. 大沽河自然保护区红松与落叶松 NPP 变化趋势与其树轮年表间相关研究[D]. 哈尔滨:哈尔滨师范大学,2019.
[10] 刘智勇,张鑫,周平. 广东省未来温度、降水及陆地生态系统 NPP 预测分析[J]. 广东林业科技,2011,27(1):59-65.
[11] YADAV A, JHA C K, SHARAN A. Optimizing LSTM for time series prediction in Indian stock market[J]. Procedia Computer Science, 2020, 167: 2091-2100.
[12] KAREVAN Z, SUYKENS J A K. Transductive LSTM for time-series prediction: An application to weather forecasting[J]. Neural Networks, 2020, 125: 1-9.
[13] GERS F A, ECK D, SCHMIDHUBER J. Applying LSTM to time series predictable through time-window approaches[M]// DORFFNER G, BISCHOF H, HORNIK K. Artificial Neural Networks. ICANN2001. Lecture Notes in Computer Science. Cham:Springer, 2001,2130:669-676.
[14] BEYENEW T. Application of artificial neural networks to statistical analysis and nonlinear modeling of high - speed interconnect systems[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2007, 26(1): 166-176.