Software Engineering and Applications 软件工程与应用, 2022, 11(3), 456-465
Published Online June 2022 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2022.113048

Hans汉斯

基于改进主成分分析的LSTM水稻产量预测模型研究

倪子凡，张云华

浙江理工大学信息学院，浙江 杭州

收稿日期：2022年4月5日；录用日期：2022年6月2日；发布日期：2022年6月10日

摘 要

作为农业大国，农业一直是我国经济发展的重要基石，而水稻作为我国主要的粮食作物，随着人口不断增长，其需求量也不断攀升，因此水稻的产量预测对我国农业的发展建设以及保障粮食安全具有重要意义。长短期记忆(LSTM)循环神经网络因其不仅能够较好地处理各因素间的非线性关系，且适合处理时间序列数据的预测问题，在作物产量预测领域应用前景良好。本文提出一种基于改进主成分分析(IPCA)的LSTM循环神经网络，对神经网络的输入进行数据降维，旨在提高神经网络训练的收敛速度，并消除输入数据间由于相关性导致的信息冗余，从而提高预测精度。

关键词

产量预测，LSTM，主成分分析

Research on LSTM Rice Yield Prediction Model Based on Improved Principal Component Analysis

Zifan Ni, Yunhua Zhang

School of Information, Zhejiang Sci-Tech University, Hangzhou Zhejiang

Received: Apr. 5th, 2022; accepted: Jun. 2nd, 2022; published: Jun. 10th, 2022

Abstract

As a major agricultural country, agriculture has always been an important cornerstone of our country's economic development. As the main food crop in our country, with the continuous growth

文章引用: 倪子凡, 张云华. 基于改进主成分分析的 LSTM 水稻产量预测模型研究[J]. 软件工程与应用, 2022, 11(3): 456-465. DOI: 10.12677/sea.2022.113048

of the population, the demand for rice is also rising. Therefore, the prediction of rice production is very important for the development and construction of agriculture and guarantee of food security. Long short-term memory (LSTM) recurrent neural network has a good application prospect in the field of crop yield prediction because it can not only deal with the nonlinear relationship between various factors, but also is suitable for dealing with the prediction problem of time series data. In this paper, a LSTM recurrent neural network based on improved principal component analysis is proposed, which reduces the data dimension of the input of the neural network, aiming to improve the convergence speed of neural network and eliminate the information redundancy caused by the correlation between the input data, thereby improving prediction accuracy.

Keywords

Yield Prediction, LSTM, PCA

1. 引言

水稻是一种主要的谷类作物，随着全球人口的持续增长，对其产量的需求也随之大幅增加。我国水稻的产量占农业总产量的五分之一[1]，因此预测其产量对农业发展与建设具有非常重要的意义[2]。同时全球变暖已经成为了一个不争的客观事实[3]，其带来的影响不可避免地对农作物生产产生了冲击[4]，结合各种气象因素来对水稻产量进行预测，不仅有助于了解气象因素对水稻产量产生的影响，还有利于提高产量预测的精度[5]。

由于受多种因素影响，使用传统方法进行水稻产量的准确预测较为困难。常规的产量预测主要包括基于统计回归原理预测、灰色理论生长模拟[6]、卫星遥感预测[7]等方法，这些方法虽然简便易行，可以用于预测产量，但实际生产中水稻产量往往与多种因素之间存在较为复杂的非线性关系[8]，使用这些传统的预测方法往往无法实现较高的预测精度。随着人工智能相关技术的不断发展与研究[9] [10]，神经网络因其对分析因素之间复杂非线性关系的优势，在作物产量预测方面得到了广泛的应用[11]，而长短期记忆(LSTM)循环神经网络结合了处理时间序列数据的能力，能够考虑到作物产量随时间波动的特点。同时影响作物产量的因素多且复杂，且各因素之间往往存在着较强的相关性，彼此之间既相互促进又相互制约，共同决定了作物的产量[12]，这为作物产量预测的建模带来了困难[13]，主成分分析是解决该问题的方法之一[14]。主成分分析是一种数据降维的方法[15]，可以将原始变量转换为一组基于原始变量的线性组合，其目标是在低维子空间中表示高维数据，使得在最小误差平方和的意义下，低维能够较好地描述原始高维数据[16]。

本文采用 LSTM 对水稻的产量进行预测，同时提出了一种改进的主成分分析方法对神经网络模型的输入进行数据降维，去除原始数据中的噪声和冗余，同时尽可能多地保留原始输入所包含的信息，提高预测精度。

2. 相关工作

2.1. LSTM 循环神经网络

LSTM 是一种特殊的循环神经网络(RNN)，通过在标准 RNN 中引入门控单元的概念来解决标准 RNN

中可能存在的梯度消失问题[17]。LSTM 循环神经网络的基本结构如图 1 所示。其中 f_t 为遗忘门，i_t 为输入门，o_t 为输出门，$\tilde{c}_t$ 是标准 RNN 单元中的结构。x_t 表示一个输入节点，用来对应一个特征参数，h_t 表示 t 时刻该单元的信息输出，C_t 表示 t 时刻记忆单元的状态。σ 一般选择 Sigmoid 作为激励函数，主要起门控作用，由于 Sigmoid 函数的输出为 0-1，当输出接近 0 或 1 时符合物理意义上的关和开。tanh 函数作为生成候选记忆单元 c 的选项，因其输出在−1 到 1 之间，符合多数场景下中心为 0 的特征分布，且梯度(求导)在接近 0 处的收敛速度快于 Sigmoid 函数。

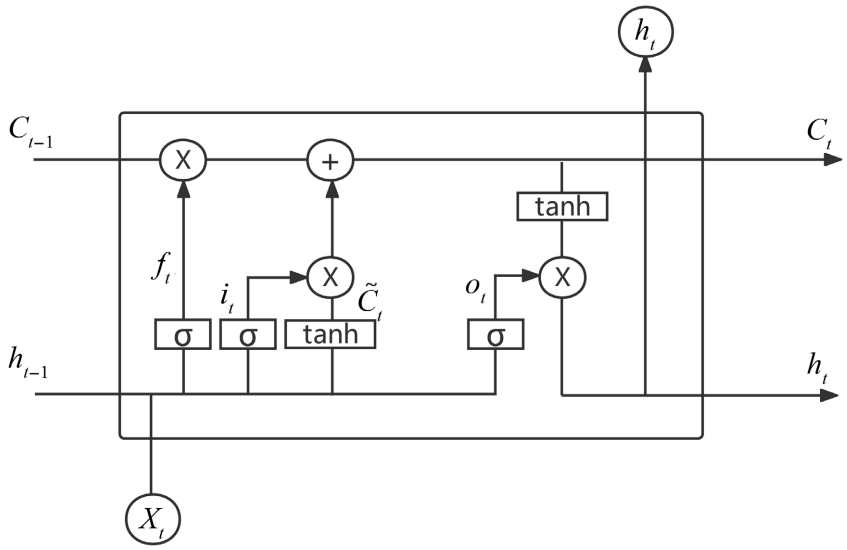

Figure 1. Basic structure of LSTM recurrent neural network
图 1. LSTM 循环神经网络基本结构

由图 1 可见 LSTM 的隐藏层通过增加神经元的复杂性，在前向计算中增加了遗忘门、输入门、输出门和记忆单元，以此保持和控制信息流动，一定程度上结局了梯度消失以及短时记忆的问题。

观察遗忘门的输出，可以得到公式(1)：

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right) \tag{1}$$

其中 W_{xf} 表示权重，下标中的 f 表示这是输入到遗忘门的权重，x 表示是 x_t 的权重，故这是 x_t 输入到遗忘门中的权重，其余权重同理，b_f 表示遗忘门偏置值，因此可以得到输入门和输出门以及 $\tilde{c}_t$ 的输出，分别见公式(2)、公式(3)和公式(4)：

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + b_i\right) \tag{2}$$

$$o_t = \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + b_o\right) \tag{3}$$

$$\tilde{c}_t = \tanh\left(W_{x\tilde{c}}x_t + W_{h\tilde{c}}h_{t-1} + b_{\tilde{c}}\right) \tag{4}$$

由公式和图 1 不难理解，遗忘门反映的是输入 x 和上一层隐藏层输出 h 被遗忘的程度的大小；输入门反映的是输入 x 和当前计算的状态更新到记忆单元的程度大小；输出门反映了输入 x 和当前输出取决于当前记忆单元的程度大小。将这些输出通过运算可得到 C_t 和 h_t，作为下一时刻的输入节点和隐藏层输入信息。计算方法见公式(5)和公式(6)：

$$C_t = C_{t-1} * f_i + \tilde{c}_t * i_t \tag{5}$$

$$h_t = o_t * \tanh\left(C_t\right) \tag{6}$$

对于一个训练好的 LSTM 模型，由遗忘门、输入门、输出门和内部记忆单元共同控制 LSTM 模型输出 h 的设计可以使整个网络更好地把握序列信息之间的关系。

2.2. 主成分分析

主成分分析是一种对数据进行降维的统计方法[18]，它借助于一个正交变换，将分量相关的原随机变量转化为分量不相关的新随机变量，以方差来衡量所包含原始变量信息的多少。在线性代数上表现为将原随机变量的协方差矩阵变换成对角矩阵，在几何上的意义则是将原坐标系通过移动和旋转变换成新的坐标系，使得所有样本点距离新的坐标轴最近，且新的坐标轴指向样本点散布最开的方向，即方差最大的方向。假设 F1 为方差最大的一个线性组合，那么称 F1 为第一主成分，如果 F1 不足以表征原变量的全部信息，则考虑选取剩下的线性组合中方差最大的线性组合 F2 作为第二主成分，依此类推可以根据需要构造出第三、第四以及更多主成分。

记原始变量 $\boldsymbol{y}=\left(Y_1,Y_2,Y_3,\cdots,Y_P\right)'$，其协方差矩阵为 Σ。主成分分析的目的即是定义一组互不相关的变量，成为 $Y_1,Y_2,Y_3,\cdots,Y_P$ 的主成分(Principal component, PC)，记为 $Z_1,Z_2,Z_3,\cdots,Z_P$，每一个主成分都是 $Y_1,Y_2,Y_3,\cdots,Y_P$ 的线性组合：

$$\begin{aligned}Z_1&=\boldsymbol{a}_1'\boldsymbol{y}=a_{11}Y_1+a_{12}Y_2+a_{13}Y_3+\cdots+a_{1p}Y_p\\Z_2&=\boldsymbol{a}_2'\boldsymbol{y}=a_{21}Y_1+a_{22}Y_2+a_{13}Y_3+\cdots+a_{2p}Y_p\\&\vdots\\Z_p&=\boldsymbol{a}_p'\boldsymbol{y}=a_{p1}Y_1+a_{p2}Y_2+a_{p3}Y_3+\cdots+a_{pp}Y_p\end{aligned}$$

由于使用线性组合的方差作为指标来衡量主成分所包含的原变量信息的多少，因此问题转化为最大化每一个主成分的方差，即 $\mathrm{var}\left(Z_j\right)=\boldsymbol{a}_j'\Sigma\boldsymbol{a}_j$ 最大，同时主成分之间应满足不相关性，否则达不到充分降维的目的，故需要同时保证 $\mathrm{cov}\left(Z_j,Z_k\right)=\boldsymbol{a}_j'\Sigma\boldsymbol{a}_k=0$。

依次对每一个主成分按照方差贡献度的指标进行排序并分析可知，第一主成分 $Z_1=\boldsymbol{a}_1'\boldsymbol{y}$ 的方差是最大的，即 $\boldsymbol{a}_1'\Sigma\boldsymbol{a}_1$ 最大，观察其方差的表达式不难看出如果 $\boldsymbol{a}_1$ 的长度越大，则会导致 $\boldsymbol{a}_1'\Sigma\boldsymbol{a}_1$ 也越大，但讨论方差贡献度时并不希望方差的大小是依赖 $\boldsymbol{a}_1$ 的长度的，因此加上限定条件 $\boldsymbol{a}_1'\boldsymbol{a}_1=1$，则 $\boldsymbol{a}_1$ 成为一个单位向量。同理可得第二主成分 $Z_2=\boldsymbol{a}_2'\boldsymbol{y}$ 需要在满足 $\boldsymbol{a}_2'\boldsymbol{a}_2=1$ 的同时最大化方差 $\boldsymbol{a}_2'\Sigma\boldsymbol{a}_2$，但由于主成分之间需要满足不相关性，因此 $\mathrm{cov}\left(Z_1,Z_2\right)=\boldsymbol{a}_1'\Sigma\boldsymbol{a}_2=0$。以此类推，第 j 主成分 $Z_j=\boldsymbol{a}_j'\boldsymbol{y}$ 则需要满足 $\boldsymbol{a}_j'\boldsymbol{a}_j=1$，且 $\boldsymbol{a}_j'\Sigma\boldsymbol{a}_k=0$，其中 $k=1,2,\cdots,j-1$ 的同时，最大化方差 $\boldsymbol{a}_j'\Sigma\boldsymbol{a}_j$。

根据线性代数定理，对于任意向量 $\boldsymbol{a}$ 和 $p\times p$ 的对称矩阵 Σ，记 $\left(\lambda_j,\boldsymbol{e}_j\right)$ 为 Σ 的特征对，其中 $j=1,2,3,\cdots,p$ 且 $\lambda_1\geq\lambda_2\geq\lambda_3\geq\cdots\geq\lambda_p\geq0$，则有 $\left(\boldsymbol{a}'\Sigma\boldsymbol{a}\right)/\left(\boldsymbol{a}'\boldsymbol{a}\right)$ 在 $\boldsymbol{a}=\boldsymbol{e}_1$ 时取得最大值，且最大值为 λ_1。当 $\boldsymbol{a}$ 与 $\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_{j-1}$ 都正交时，$\left(\boldsymbol{a}'\Sigma\boldsymbol{a}\right)/\left(\boldsymbol{a}'\boldsymbol{a}\right)$ 在 $\boldsymbol{a}=\boldsymbol{e}_j$ 时取得最大值，且最大值为 λ_j，其中 $j=2,3,\cdots,p-1$。

由之前的条件可知，以 $\boldsymbol{a}_2'$ 为例，当满足 $\boldsymbol{a}_2'\Sigma\boldsymbol{a}_1=0$ 时，由于 $\boldsymbol{a}_1=\boldsymbol{e}_1$，故可推出 $\boldsymbol{a}_2'\lambda\boldsymbol{e}_1=0$，即 $\boldsymbol{a}_2'$ 与 $\boldsymbol{e}_1$ 是垂直的，可以看出该定理可以用于解决本文的最大化方差问题。则变量 $Y_1,Y_2,Y_3,\cdots,Y_P$ 的第 j 个主成分公式如式(7)：

$$Z_j=\boldsymbol{e}_j'\boldsymbol{y}=e_{j1}Y_1+e_{j2}Y_2+e_{j3}Y_3+\cdots+e_{jp}Y_p,j=1,2,\cdots,p \tag{7}$$

其中 $\left(\lambda_j,\boldsymbol{e}_j\right)$ 为协方差矩阵Σ的特征对，$\lambda_1\geq\lambda_2\geq\lambda_3\geq\cdots\geq\lambda_p\geq0$ 且特征向量均为标准化特征向量，而且有如下性质：

$$\mathrm{var}\left(Z_j\right)=\boldsymbol{e}_j'\Sigma\boldsymbol{e}_j=\lambda_j \tag{8}$$

$$\mathrm{cov}\left(Z_j,Z_k\right)=\boldsymbol{e}_j'\Sigma\boldsymbol{e}_k=0 \tag{9}$$

进一步地可以得到式(10)：

$$\sum_{j=1}^{p}\mathrm{var}\left(Z_j\right)=\sum_{j=1}^{p}\mathrm{var}\left(Y_j\right) \tag{10}$$

由(10)式可以看出，得到的主成分可以完全描述原始变量的方差。

实际生产中，变量的方差可能会因为量纲的不同而导致部分变量方差过大或过小的问题，为了消除量纲带来的影响，可以将原始变量除以标准差，从而构造出一个新的标准化矩阵来进行主成分分析，即：

$$W_j=\left(Y_j-u_i\right)\Big/\sqrt{\sigma_{jj}} \tag{11}$$

其中分子的作用是进行中心化。而新变量的协方差矩阵即为原始变量的相关系数矩阵，因此可以通过求原始变量的相关系数矩阵的特征值和特征向量来消除量纲的影响，达到标准化的目的。但是需要注意的是，这两种方法分析得到的两组主成分的数值和方向是没有关系的，二者之间并不能通过简单的线性变换进行转化。

2.3. 主成分分析存在的问题及改进

特征与特征之间的影响程度可以由特征间的相关系数来刻画，而特征的变异程度则由变异系数(Coefficient of Variation, CV)来刻画，变异系数计算方式如公式(12)所示，其中分子为标准差，分母为平均值：

$$\mathrm{CV}=\frac{\sigma}{\mu} \tag{12}$$

如公式(11)所示的标准化过程虽然可以消除量纲对方差带来的影响，但经过标准化后的变量计算是基于原始变量的相关系数矩阵来进行的，观察相关系数矩阵可以发现其对角线元素都是 1，即标准化使得变量的方差都变成了 1，也就忽略了特征本身的变异程度，从而不能完整刻画原始变量的全部信息。

为了解决上述问题，采用如公式(13)所示的标准化方法：

$$y_{ij}=\frac{x_{ij}}{\overline{x_j}} \tag{13}$$

其中 x_{ij} 为原始变量 $n\times p$ 矩阵中的元素，y_{ij} 为标准化后的矩阵中的元素，$\overline{x_j}$ 为原始变量矩阵第 j 列的均值，即第 j 个特征的均值，计算方法如公式(14)所示：

$$x_j=\frac{1}{n}\sum_{k=1}^{n}x_{kj} \tag{14}$$

设标准化后的矩阵的协方差矩阵为 $C=\left(c_{ij}\right)_{n\times p}$，其计算方法如公式(15)所示：

$$c_{ij}=\frac{1}{n}\sum_{k=1}^{n}\left(y_{ki}-\overline{y_i}\right)\left(y_{kj}-\overline{y_j}\right) \tag{15}$$

由公式(13)和公式(14)易推得标准化后的矩阵中每个特征的均值都为 1，即 $\overline{y_i}=1$，可得：

$$c_{ij}=\frac{1}{n}\sum_{k=1}^{n}\left(\frac{x_{ki}}{\overline{x_i}}-1\right)\left(\frac{x_{kj}}{\overline{x_j}}-1\right) \tag{16}$$

化简式(16)可得：

$$c_{ij}=\frac{\frac{1}{n}\sum_{k=1}^{n}\left(x_{ki}-\overline{x_i}\right)\left(x_{kj}-\overline{x_j}\right)}{\overline{x_i}\cdot\overline{x_j}} \tag{17}$$

不难看出式(17)中分子为原始变量的协方差矩阵中的元素，特别地，矩阵 C 的对角线元素为原始变量的变异系数的平方，因此该标准化没有丢失变量变异程度的信息。同时观察标准化后矩阵的相关系数，设原始变量的相关系数为 ρ_{ij}，经过标准化后为 r_{ij}，s_{ij} 为原始变量协方差矩阵中的元素：

$$r_{ij} = \frac{c_{ij}}{\sqrt{c_{ii}} \cdot \sqrt{c_{jj}}} \tag{18}$$

由式(17)可知 $c_{ij} = \frac{s_{ij}}{\overline{x_i} \cdot \overline{x_j}}$，代入式(18)并化简，得到公式(19)如下：

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}} \cdot \sqrt{s_{jj}}} \tag{19}$$

而 $\frac{s_{ij}}{\sqrt{s_{ii}} \cdot \sqrt{s_{jj}}} = \rho_{ij}$，因此该标准化也不会丢失原始变量的相关系数信息，综上所述，该标准化方法可以更完整地刻画原始变量的信息。

3. 模型构建

对原始数据的输入进行主成分分析，将输入数据降维后作为 LSTM 的输入节点，模型结构如图 2 所示。其中 $X_1, X_2, \cdots, X_7$ 为原始输入变量，经过主成分分析后降维为两个变量，即 PC1 和 PC2，再以这两个主成分作为神经网络的输入节点进行预测并输出产量(Yield)。

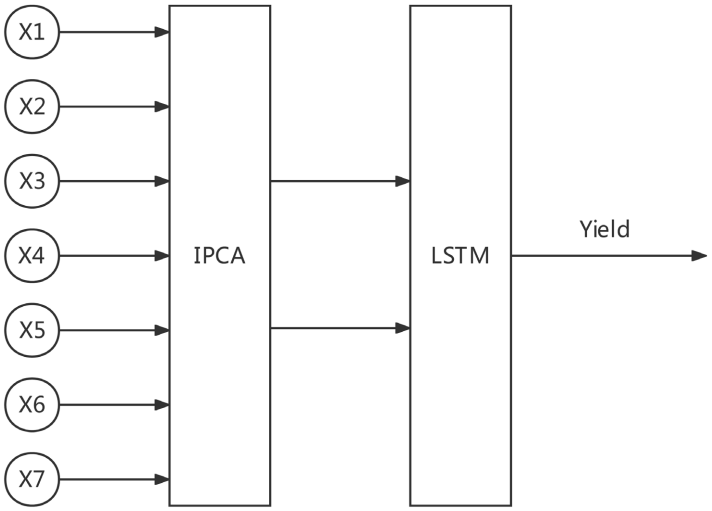

Figure 2. Structure of LSTM based on PCA
图 2. 基于主成分分析的 LSTM 结构

3.1. 主成分分析相关计算

首先计算数据集的 KMO (Kaiser-Meyer-Olkin)检验值和 Bartlett 球形检验值，以判断该数据集是否适合进行主成分分析。KMO 值大于 0.5 且越接近 1 时就越适合进行主成分分析。Bartlett 球形检验值小于显著水平 0.05 或者 0.01 时适合进行主成分分析。

在本文提出的改进主成分分析中，先将数据进行标准化后使用其协方差矩阵的特征值作为指标来选取主成分，选取主成分一般遵循以下原则：特征值大于 1；累计方差贡献率大于 85%。

3.2. 神经网络模型构建

根据 Kolmogorov 定理，含有输入层、隐含层和输出层的三层神经网络可以精确表征任意一个连续函数[19]，因此本文同样采用输入层、隐含层和输出层的三层结构来构建神经网络，隐含层激活函数采用 Softmax 函数，输出层激活函数采用恒等式。训练的参数通过比较不同取值对预测结果的影响来确定，最终结果如表 1 所示。

Table 1. Parameters
表 1. 参数设置

参数	值
学习率	0.01
Batch-size	8
交叉检验数据比例	0.3
BP 隐藏层节点数	8
LSTM 隐藏层节点数	64
迭代次数	500

由最终确定的节点个数可以看出，BP 神经网络的最优节点数显著小于 LSTM 的最优节点数，原因在于过于复杂的网络拓扑结构虽然可以提高拟合的精度，但也可能导致这两种神经网络的过拟合问题，而过拟合问题对于 BP 神经网络而言较为突出，LSTM 因其本身独特的门控结构使得其的权重更新不会受到所有数据的影响，而是可以选择记住或者遗忘前时刻的特征，因此一定程度上降低了产生过拟合的可能性[20]。

4. 实验结果及分析

4.1. 数据集

为了验证模型的有效性，本文使用的某地水稻产量数据集中原始变量包含降雨量(Precipitation)、最低温度(Minimum Temperature)、平均温度(Mean Temperature)、最高温度(Maximum Temperature)、相关作物需水量(Reference Crop Evapotranspiration)、种植面积(Area)和总产量(Production)。

4.2. 相关参数计算及数据预处理

由表 2 数据可知，该数据集 KMO > 0 且 Sig < 0.01，因此变量之间相关性较高，存在信息重叠，故可以进行主成分分析。

Table 2. KMO and Bartlett's test of sphericity
表 2. KMO 和巴特利特球形检验

KMO 检验值	巴特利特球形检验值(Sig)
0.676	0.000

将数据集按改进的标准化方法处理后计算其协方差矩阵及其特征值来进行主成分的选取，计算结果如表 3 所示。

Table 3. Variance contribution rate
表 3. 方差贡献率

成分	方差	初始方差贡献率/%	累计方差贡献率/%
1	3.991	57.019	57.019
2	2.282	32.606	89.625
3	0.491	7.009	96.634
4	0.112	1.569	98.229
5	0.080	1.139	99.369
6	0.030	0.423	99.791
7	0.015	0.209	100.000

根据选取主成分的原则可知，该数据集应选取的主成分个数为 2，设第一主成分为 F1，第二主成分为 F2，其成分矩阵如表 4 所示。

Table 4. Component matrix
表 4. 成分矩阵

原始变量	第一主成分 F1	第二主成分 F2
Precipitation (X_1)	−0.781	0.490
Minimum Temperature (X_2)	0.796	0.325
Mean Temperature (X_3)	0.949	0.157
Maximum Temperature (X_4)	0.991	0.023
Reference Crop Evapotranspiration (X_5)	0.925	−0.192
Area (X_6)	0.058	0.973
Production (X_7)	0.076	0.963

由表 4 中主成分系数可得主成分表达式如(20)式和(21)式所示。

$$F1 = -0.781X_1 + 0.796X_2 + 0.949X_3 + 0.991X_4 + 0.925X_5 + 0.058X_6 + 0.076X_7 \tag{20}$$

$$F2 = 0.490X_1 + 0.325X_2 + 0.1573_1 + 0.023X_4 - 0.192X_5 + 0.973X_6 + 0.963X_7 \tag{21}$$

根据(12)式和(13)式计算主成分 F1 和 F2，则将原变量由 7 维降到了 2 维。

4.3. 模型对比

通过在相同数据集上对比标准 BP 神经网络的预测效果和经过改进主成分分析后的 LSTM 神经网络(IPCA-LSTM)预测效果，以验证本文模型的有效性。评价指标采用均方根误差(RMSE)和平均绝对百分比误差(MAPE)，计算结果见公式(22)和公式(23)，其中 $h(x_i)$ 为预测值，y_i 为观测值。图 3 和图 4 分别为 BP 神经网络和 PCA-LSTM 在训练集上随着迭代次数增加的损失(Loss)变化情况，横坐标为迭代次数，纵坐标为损失。实验结果如表 5 所示。

$$\mathrm{RMSE}(X,h) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(h(x_i) - y_i\right)^2} \tag{22}$$

$$\mathrm{MAPE}(X,h)=100\%\times\frac{1}{n}\sum_{i=1}^{n}\left|h(x_i)-y_i\right| \tag{23}$$

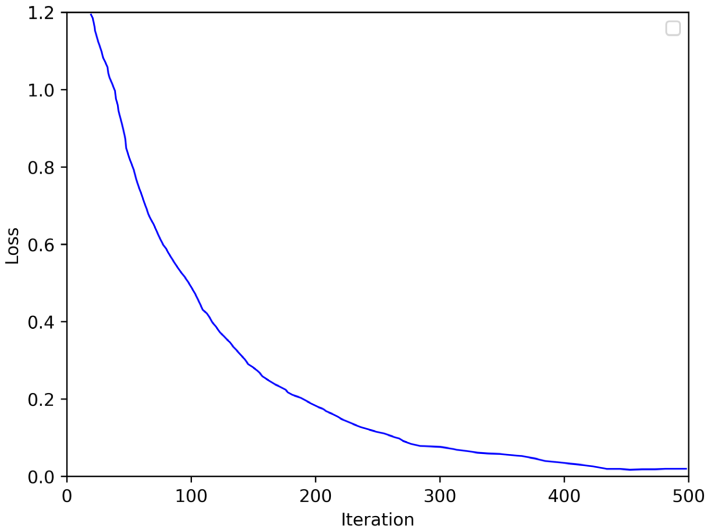

Figure 3. Loss of BP neural network on training set
图 3. BP 神经网络训练模型损失

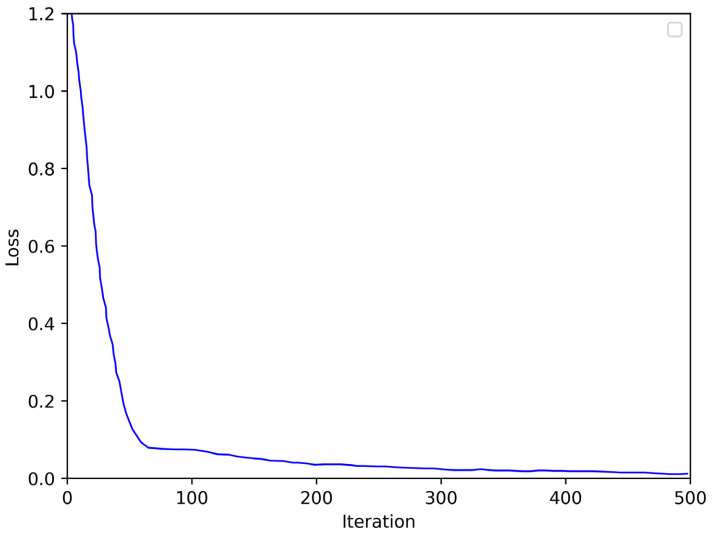

Figure 4. Loss of IPCA-LSTM on training set
图 4. IPCA-LSTM 训练模型损失

Table 5. System results of different models on the dataset
表 5. 不同模型在数据集上的结果

指标	RMSE	MAPE/%
BP	0.0542	3.57
PCA-LSTM	0.0261	1.26

从实验结果不难看出，本文提出的基于改进主成分分析的 LSTM 神经网络模型相较于仅使用 BP 神经网络的模型具有更高的预测精度，误差更小且模型训练收敛更快。

5. 结论

本文针对作物产量预测研究中数据维度高和包含时间序列信息的问题，提出了基于改进主成分分析的 LSTM 神经网络预测模型，在能够处理复杂输入变量之间的非线性关系的同时，可以有效降低原始输入变量的维度，消除其冗余性，提高模型训练时的收敛速度，提高预测精度，应用前景良好。

参考文献

[1] Guo, Y., Fu, Y., Hao, F., Zhang, X., Wu, W., Jin, X., et al. (2021) Integrated Phenology and Climate in Rice Yields Prediction Using Machine Learning Methods. *Ecological Indicators*, **120**, Article ID: 106935. https://doi.org/10.1016/j.ecolind.2020.106935

[2] 王雨晨. 基于灰色模型的江西水稻产量预测研究[J]. 粮食科技与经济, 2020, 45(4): 29-30. https://doi.org/10.16465/j.gste.cn431252ts.20200404

[3] 李红艳, 徐建强, 许甫金, 肖玉苹, 沈足金, 张乐平, 方明. 气象因素对水稻产量的影响及预测模型的建立[J]. 浙江农业科学, 2018, 59(7): 1104-1107+1110. https://doi.org/10.16178/j.issn.0528-9017.20180707

[4] 沈陈华. 气象因子对江苏省水稻单产的影响[J]. 生态学报, 2015, 35(12): 4155-4167. https://doi.org/10.5846/stxb201309212315

[5] 赵桂涛, 刘中聚, 冯尚宗, 赵理, 王世伟, 娄华敏. 基于气象因素的临沂水稻产量评估预测模型[J]. 江西农业学报, 2016, 28(10): 71-74. https://doi.org/10.19386/j.cnki.jxnyxb.2016.10.16

[6] 向昌盛, 张林峰. 灰色理论和马尔可夫相融合的粮食产量预测模型[J]. 计算机科学, 2013, 40(2): 245-248.

[7] 薛利红, 曹卫星, 罗卫红. 基于冠层反射光谱的水稻产量预测模型[J]. 遥感学报, 2005, 9(1): 100-105.

[8] Schlenker, W. and Roberts, M.J. (2009) Nonlinear Temperature Effects Indicate Severe Damages to US Crop Yields under Climate Change. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 15594-15598. https://doi.org/10.1073/pnas.0906865106

[9] 汪小昱, 丁为民, 罗卫红, 戴剑锋. 利用 BP 神经网络对江淮地区梅雨季节现代化温室小气候的模拟与分析[J]. 农业工程学报, 2004, 20(2): 235-238.

[10] 任守纲, 刘鑫, 顾兴健, 王浩云, 袁培森, 徐焕良. 基于 R-BP 神经网络的温室小气候多步滚动预测模型[J]. 中国农业气象, 2018, 39(5): 314-324.

[11] 宗宸生, 郑焕霞, 王林山. 改进粒子群优化 BP 神经网络粮食产量预测模型[J]. 计算机系统应用, 2018, 27(12): 204-209. https://doi.org/10.15888/j.cnki.csa.006651

[12] 李灿东. 黑龙江省北部大豆主栽品种产量性状主成分分析[J]. 现代化农业, 2020(12): 4-8.

[13] Borrego, C., Tchepel, O., Costa, A.M., Amorim, J.H. and Miranda, A.I. (2003) Emission and Dispersion Modelling of Lisbon Air Quality at Local Scale. *Atmospheric Environment*, **37**, 5197-5205. https://doi.org/10.1016/j.atmosenv.2003.09.004

[14] Sousa, S.I.V., Martins, F.G., Alvim-Ferraz, M.C.M. and Pereira, M.C. (2007) Multiple Linear Regression and Artificial Neural Networks Based on Principal Components to Predict Ozone Concentrations. *Environmental Modelling & Software*, **22**, 97-103. https://doi.org/10.1016/j.envsoft.2005.12.002

[15] 肖枝洪, 冉小华. 运用主成分分析法的过程控制和诊断[J]. 重庆理工大学学报(自然科学), 2014, 28(1): 96-101.

[16] 梁佐堂, 刘欣涛. 基于主成分分析的手写体数字识别方法研究[J]. 信息技术, 2016(8): 121-124. https://doi.org/10.13274/j.cnki.hdzj.2016.08.030

[17] 张驰, 郭媛, 黎明. 人工神经网络模型发展及应用综述[J]. 计算机工程与应用, 2021, 57(11): 57-69.

[18] 郭凯维, 郭传超, 史耀凡, 于水. 基于主成分分析的 GA-BP 神经网络地表下沉系数预测[J]. 北京测绘, 2021, 35(11): 1374-1379. https://doi.org/10.19580/j.cnki.1007-3000.2021.11.003

[19] 和树栋. 基于 BP 神经网络的保护层卸压高度预测方法研究[J]. 煤炭技术, 2020, 39(6): 73-75. https://doi.org/10.13301/j.cnki.ct.2020.06.022

[20] 崔巍, 顾冉浩, 陈奔月, 王文. BP 与 LSTM 神经网络在福建小流域水文预报中的应用对比[J]. 人民珠江, 2020, 41(2): 74-84.