



Study on BP Neural Network Fitting for High-Dimensional Small-Sample Data Based on PCA Dimensionality Reduction

Yaming Ren*, Xingzhi Xu

College of Mechanical and Control Engineering, Guilin University of Technology, Guilin, China
Email: *renyaming1981@gmail.com

How to cite this paper: Ren, Y.M. and Xu, X.Z. (2026) Study on BP Neural Network Fitting for High-Dimensional Small-Sample Data Based on PCA Dimensionality Reduction. *Open Access Library Journal*, 13: e16061.

<https://doi.org/10.4236/oalib.1116061>

Received: September 21, 2026

Accepted: October 20, 2026

Published: October 23, 2026

Copyright © 2026 by author(s) and Open Access Library Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

This paper studies the preprocessing method for high-dimensional small-sample data. High-dimensional small-sample data usually has feature redundancy and multicollinearity, which makes the generalization ability of prediction models worse. This paper uses principal component analysis (PCA) to reduce data dimension. It selects useful principal components by variance contribution rate and uses data reconstruction to check whether key information is kept. A BP neural network with fixed parameters is used for comparison tests. The results show that the test-set coefficient of determination rises from 0.94671 to 0.9715 after PCA processing. It proves that PCA can remove redundant information and improve input data quality for high-dimensional small-sample fitting.

Subject Areas

Data Preprocessing, Neural Networks, Data Fitting

Keywords

Principal Component Analysis, Data Preprocessing, Fitting, BP Neural Network

1. 引言

在非线性数据拟合领域，反向传播(BP)神经网络凭借强大的非线性映射能力，广泛用于建立输入输出样本间的量化关系[1] [2]。BP神经网络通过误差反向迭代更新网络权重具备优秀的函数逼近能力是回归任务中最常用的建模手段之一[3]-[5]。该模型无需构建复杂的物理机理方程，仅依托实测输入输出样本即可自适应挖掘数据内在规律，结构简洁、适配性强，在各类工程拟

合、预测场景中应用广泛[6] [7]。

但在实际工程试验与数据采集过程中，受试验条件、工况约束、采集成本限制，普遍存在高维特征、小样本量的数据分布特征，常出现输入特征维数远大于有效训练样本数量的极端数据结构。该类高维小样本数据直接建模时，高冗余输入特征会急剧增加网络输入层神经元数量导致网络拓扑规模大幅膨胀，模型训练参数量激增、迭代收敛速度变慢；同时高维特征普遍存在多重共线性问题，特征间信息重叠、噪声干扰严重，易导致网络权重更新紊乱，模型容易陷入局部最优解。最为突出的是会引发严重的过拟合现象，模型在训练集表现精度优异，但测试集泛化能力大幅衰减，极大限制了传统 BP 神经网络在工程实测数据拟合场景的实用性。

主成分分析(PCA)是经典的无监督线性降维算法，核心优势是通过正交变换对高维耦合特征进行解耦重构，剔除冗余保留数据信息的主成分特征[8]。该算法能够在最大限度保留原始数据分布规律的前提下，压缩特征维度适配高维小样本数据的预处理场景，可有效解决传统 BP 神经网络维度爆炸、过拟合问题[9] [10]。

为有效改善高维小样本条件下 BP 神经网络拟合精度低、泛化性差的问题，本文构建 PCA-BP 组合拟合模型。通过主成分分析对原始高维输入数据进行降维预处理，提纯有效特征、精简网络输入结构。最终通过仿真实验，对比传统 BP 模型与 PCA-BP 优化模型的拟合效果，验证本文提出的 PCA-BP 在高维小样本拟合任务中的有效性。

2. BP神经网络基本原理

BP 神经网络是一种基于误差反向迭代训练的多层前馈神经网络，具备强大的非线性函数逼近与自适应拟合能力，结构简单、容错性强，被广泛应用于各类工程数据拟合与预测建模场景中。BP 神经网络无需依赖精准的物理建模公式，仅依靠输入输出样本数据即可完成模型训练，能够有效拟合复杂数据之间的非线性映射关系。

BP 神经网络主要由输入层、隐含层、输出层三部分构成，层与层之间神经元实现全连接，同层神经元无连接关系。输入层神经元数量对应模型的输入特征维度，负责接收原始数据样本；隐含层为网络核心非线性运算层，可设置单层或多层结构，通过激活函数引入非线性特性，实现复杂数据规律的挖掘；输出层神经元对应模型拟合输出结果，完成最终的数据映射输出。

BP 神经网络训练过程分为信号正向传播与误差反向传播两个阶段。正向传播过程中，输入数据逐层加权运算并输出预测值；当预测输出与真实样本输出存在误差时，模型进入反向传播阶段，通过梯度下降算法逐层更新网络权重与偏置参数，不断缩小拟合误差，直至模型满足预设训练精度或迭代次数要求。

设网络输入样本向量为 $X = (x_1, x_2, \dots, x_n)$ ，输入层至隐含层的连接权重为 w_{ij} ，隐含层神经元偏置为 b_j ，隐含层输出计算公式为：

$$H_j = f\left(\sum_{i=1}^n w_{ji}x_i + b_j\right) \quad (1)$$

式中： H_j 为隐含层第 j 个神经元输出； $f(\cdot)$ 为神经网络激活函数，本文采用常用 Sigmoid 激活函数，其表达式为：

$$f(x) = \frac{1}{1 + e^{-x}} \quad (2)$$

隐含层数据经过加权运算后传递至输出层，最终得到网络预测输出值：

$$Y_k = \sum_{j=1}^m w_{jk}H_j + b_k \quad (3)$$

式中： w_{jk} 为隐含层到输出层权重； b_k 为输出层偏置； Y_k 为网络拟合输出结果。模型训练以降低整体拟合误差为核心目标，定义网络误差损失函数：

$$E = \frac{1}{2} \sum_{k=1}^q (T_k - Y_k)^2 \quad (4)$$

式中： T_k 为样本真实输出值， Y_k 为网络预测输出值， E 为整体拟合误差，为最小化误差函数，模型常常采用梯度下降算法反向修正权重与偏置参数，沿着误差梯度下降方向迭代更新参数，权重更新公式为：

$$\Delta w = -\eta \frac{\partial E}{\partial w} \quad (5)$$

式中： η 为网络学习率，用于控制参数迭代更新步长，取值范围为 $(0,1)$ 。

在常规样本充足、输入维度适中的建模场景中，传统 BP 神经网络可实现稳定拟合。但在输入特征维数远大于有效样本数量的高维小样本工况下，传统 BP 网络存在明显短板。输入维度过高会大幅增加网络输入层神经元数量，导致网络权重、偏置参数量急剧膨胀，训练计算量大幅提升，收敛速度显著变慢；高维原始特征普遍存在冗余性与多重共线性，无效特征会干扰网络参数迭代更新，模型易陷入局部最优解；小样本、高维度的不匹配数据结构，极易引发模型过拟合问题，模型对训练样本拟合精度极高，但对未知样本的泛化拟合能力大幅下降，无法满足实际工程建模需求。

基于上述缺陷，本文引入主成分分析算法对高维输入数据进行降维预处理，剔除冗余特征、精简网络输入结构，以此优化 BP 神经网络的拟合性能与泛化能力。

3. 主成分分析(PCA)降维原理

主成分分析(Principal Component Analysis, PCA)是工程建模中应用最广泛的线性特征降维算法，其核心思想是通过正交线性变换，将存在相关性的高维原始特征映射为一组相互独立、无冗余的全新主成分特征。在保留数据整体信息主体分布的前提下，大幅压缩特征维度、消除特征共线性、过滤无效噪声维度，非常适用于输入维数高、样本数量有限的小样本建模场景。

本文采用 PCA 对原始输入数据集进行前置降维处理，通过筛选高贡献率主成分重构低维有效特征空间，从而简化后续 BP 网络输入结构、提升模型稳定性与拟合泛化能力。

3.1. 数据预处理

工程实测采集的原始输入特征量纲、数值区间差异较大，直接参与协方差计算会导致数值偏大特征主导整体方差，造成降维失真。因此在 PCA 运算前需对原始数据进行标准化处理。

设原始样本数据集包含 n 个样本、 p 维输入特征，原始数据矩阵 X 标准化公式如下：

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{\sigma_j} \quad (6)$$

其中： $\bar{x}_j$ 为第 j 维特征均值， σ_j 为第 j 维特征标准差， x_{ij}^* 为标准化后数据。标准化后所有特征均值为 0、方差为 1，消除量纲差异对降维结果的干扰。

3.2. 协方差矩阵构建与特征分解

经过标准化处理之后，数据集记为 X ， X 是 $n \times p$ 矩阵： n 代表样本数量， p 代表特征维度，矩阵每一行对应一个样本，每一列对应一个特征。

协方差矩阵用于刻画各个特征维度之间的相关程度，协方差矩阵 C 为 p 阶方阵，计算公式：

$$C = \frac{1}{n-1} X^T X \quad (7)$$

其中 X^T 是 X 的转置矩阵；除以 $n-1$ 为无偏估计形式。协方差矩阵对角元素代表单个特征自身方差，非对角元素代表不同特征间的相关性，数值越大说明特征多重共线性越严重。对协方差矩阵进行特征值分解：

$$C v_i = \lambda_i v_i \quad (8)$$

式中： λ_i 为特征值，表征维度信息贡献量； v_i 为对应特征向量，表征数据空间变换方向。特征值越大，该维度承载的原始数据有效信息越多。

将全部特征值按照从大到小排序：

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \quad (9)$$

3.3. 主成分贡献率筛选规则

特征值大小直接对应维度信息承载量，为筛选有效特征、剔除冗余维度，本文引入单维度贡献率与累计贡献率作为降维筛选标准。

单维度贡献率计算公式：

$$\eta_i = \frac{\lambda_i}{\sum_{k=1}^p \lambda_k} \quad (10)$$

累计贡献率计算公式：

$$\eta_{sum} = \sum_{i=1}^m \eta_i \quad (11)$$

本文采用工程通用最优筛选标准：选取累计贡献率 $\geq 95\%$ 的前 m 个主成分作为降维后新特征。该标准可在最大程度保留原始数据核心规律的基础上，剔除噪声、冗余低维特征，实现高维特征高效压缩。

3.4. 生成降维后新特征矩阵

选取累计贡献率 95% 以上的前 m 个主成分对应的特征向量，构建正交变换矩阵，将原始高维数据投影至全新低维独立特征空间，得到降维后的有效数据集：

$$Y_{n \times m} = X_{n \times p} \cdot V_{p \times m} \quad (12)$$

降维后特征维度由原始高维 p 压缩为低维 m ，各主成分特征相互正交、无相关性，彻底解决原始数据多重共线性问题。

4. 仿真实验与结果分析

为验证本文构建 PCA-BP 组合模型在超高维光谱小样本回归场景下的性能，本章仿真算例选用基于近红外光谱的汽油辛烷值预测案例[11]，属于典型高维小样本回归任务。采用 MATLAB 平台完成仿真，对比原始 BP 神经网络与 PCA-BP 降维优化模型。

辛烷值是表征汽油品质的核心指标，近红外光谱可实现快速无损检测。采集 60 组汽油样品，采用傅里叶近红外光谱仪在 900~1700 nm 波段扫描，波长间隔 2 nm，每组样品得 401 个波长吸光度特征，输出标签为汽油实测辛烷值。数据集特征维度 401，样本总数仅 60 组，属于典型超高维小样本，光谱各波长特征之间存在极强多重共线性与光谱噪声。数据集随机划分：训练样本 50 组，测试样本 10 组。训练集用于模型参数学习，测试集用于评估模型泛化预测能力。

仿真环境：MATLAB R2018b；两组模型的训练环境、随机种子保持一致，消除随机因素干扰，保证对比公平。原始 BP 模型：直接将 401 维原始光谱吸光度作为网络输入；输入层神经元 401 个，单层隐含层神经元 9 个，输出层 1 个(辛烷值)。PCA-BP 模型：先对 401 维光谱数据 PCA，选取累计贡献率 $\geq 95\%$ 的主成分，将原始 401 维压缩至 4 维主成分特征；再将 4 维主成分送入 BP 网络。BP 网络结构、训练超参与对照组完全相同：隐含层 9 个神经元，仅输入特征替换为 PCA 提取的主成分。

本文选取决定系数 R^2 作为模型拟合效果的唯一评价指标，用于定量表征模型预测值与真实值的贴合程度以及模型的泛化性能。决定系数 R^2 的取值范围为[0, 1]，数值越接近 1，代表模型拟合精度越高、泛化能力越强。计算公式如下：

$$R^2 = \frac{[n \sum y_i \hat{y}_i - \sum y_i \sum \hat{y}_i]^2}{[n \sum y_i^2 - (\sum y_i)^2][n \sum \hat{y}_i^2 - (\sum \hat{y}_i)^2]} \quad (13)$$

式中： y_i 为汽油辛烷值实测真值， $\hat{y}_i$ 为模型预测辛烷值， n 为测试集样本数量。

本次实验数据集共计 60 组汽油样品，每组样品包含 401 个光谱吸光度数据点，即原始特征维度为 401 维。在 PCA 主成分分析的实际运算规则中，模型可解析的最大有效主成分维度由样本数量与特征维度的最小值决定，因此本数据集 PCA 有效最高维度仅为 60 维，而非原始的 401 维。图 1 为各主成

分方差单独贡献率, 由图 1 可以清晰看出, 所有主成分的方差贡献率呈现极度不均衡分布规律, 前序少量主成分占据了绝大部分数据信息贡献量, 曲线下降趋势陡峭、衰减速度极快; 而后续绝大多数主成分的方差贡献率趋近于零, 对应的维度几乎不承载有效光谱信息。图 2 为 PCA 累计方差贡献率曲线, 本文设定 95% 累计贡献率为核心筛选阈值, 从累计贡献曲线可以看出, 前 4 个主成分的累计方差贡献率即可达到 95% 以上, 满足本文建模的信息保留要求。这表明仅依靠前 4 维核心主成分, 就可以完整覆盖原始高维光谱

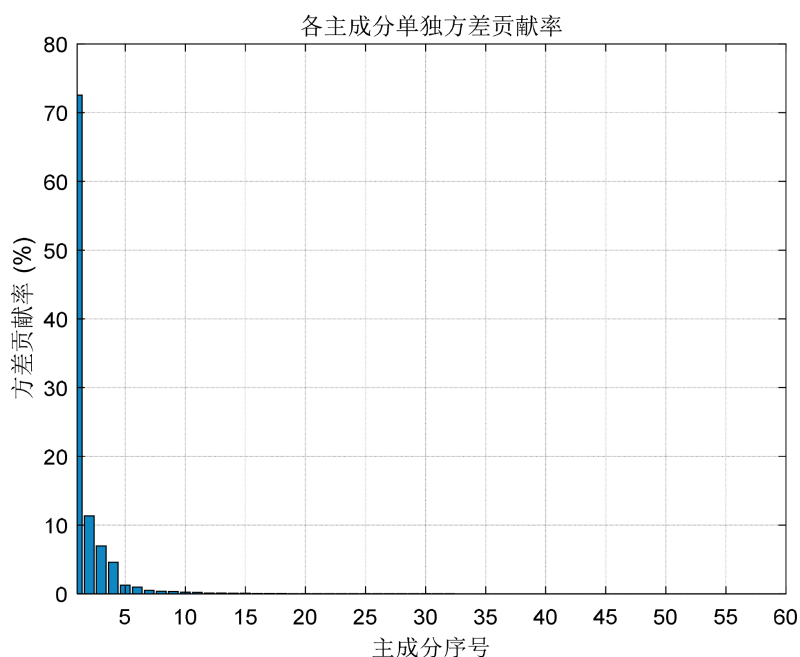


图 1. 各主成分方差单独贡献率

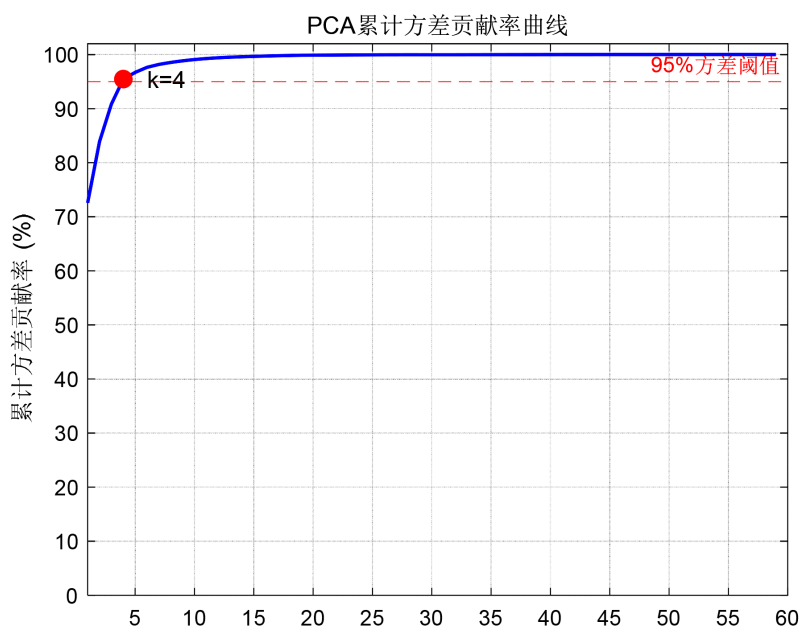


图 2. PCA 累计方差贡献率曲线

数据的核心变化规律, 无需保留 60 维最大有效主成分, 更无需使用 401 维原始特征, 极大程度压缩了数据维度, 一定程度上消除了原始光谱特征的多重共线性与维度冗余问题。

图 3 为原始光谱曲线与 PCA 重构光谱曲线对比图。为进一步验证本次 PCA 降维与主成分筛选的准确性、有效性, 本文选取数据集中第一个汽油样本开展数据重构验证。将筛选后的核心主成分反向重构还原为原始维度数据, 对比样本原始光谱曲线与 PCA 重构后的光谱曲线, 两条曲线整体走势完全一致、数值高度贴合, 几乎完全重合。该结果直观证明, 本次筛选的核心主成分保留了大部分原始光谱的有效特征信息。

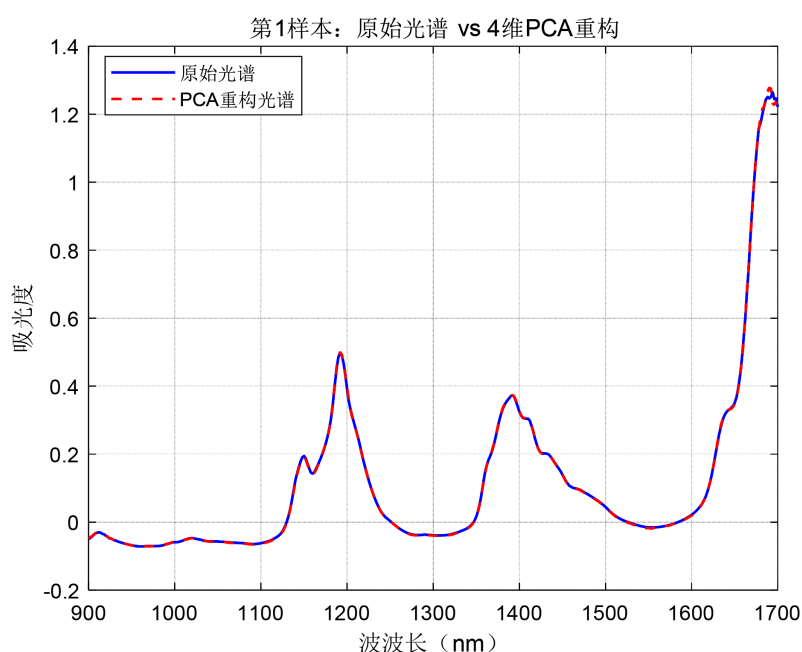


图 3. 原始光谱曲线与 PCA 重构光谱曲线对比图

结合本文仿真得到的两张测试集预测结果图可进一步直观对比传统 BP 模型与 PCA-BP 组合模型的拟合性能。图 4 为以原始 401 维光谱数据为输入的传统 BP 神经网络辛烷值预测结果, 图 5 为以 PCA 筛选的 4 维核心主成分为输入的 PCA-BP 模型预测结果。从模型拟合精度来看, 传统 BP 神经网络以 401 维原始高维光谱特征作为输入, 模型测试集决定系数仅为 0.94671, 本文所提 PCA-BP 算法仅选取 4 维核心主成分作为神经网络输入, 在极大精简输入维度的前提下, 模型测试集决定系数达到 0.9715, 相较于传统 BP 模型精度得到显著提升。

由于 BP 神经网络的权重与偏置采用随机初始化方式, 单次仿真实验结果存在随机性, 仅依靠单次实验对比难以客观评价模型性能。为降低随机因素对实验结论的干扰, 本文对传统 BP 神经网络与 PCA-BP 组合模型分别独立重复开展 10 次仿真实验, 统计得到测试集决定系数的最大值、最小值、平均值以及运算时间, 统计结果如表 1 所示。由表 1 统计结果可以看出: 在精度层面, 基于 PCA-BP 模型的决定系数的平均值、最大值、最小值均高于

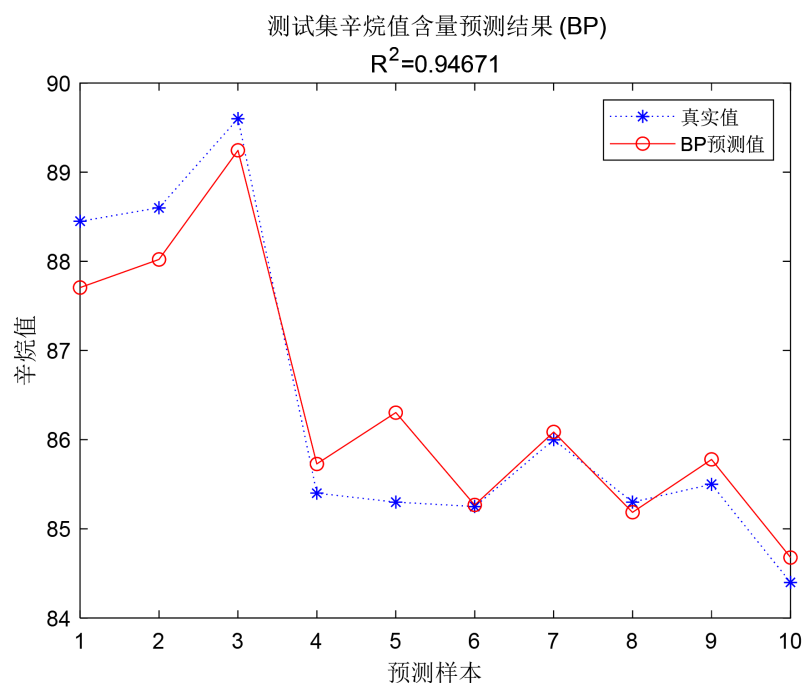


图 4. 原始 401 维光谱数据为输入的传统 BP 神经网络辛烷值预测结果

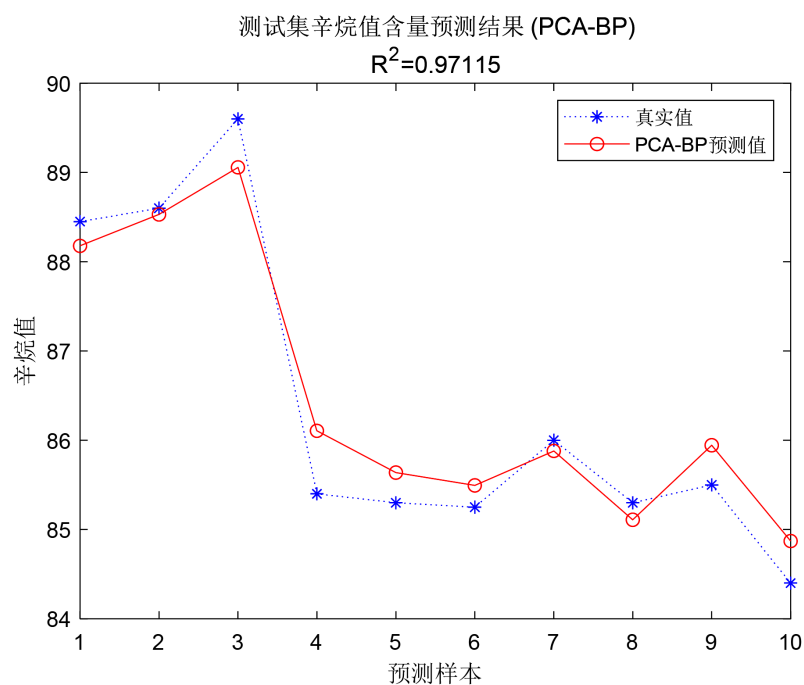


图 5. PCA 筛选的 4 维核心主成分为输入的 PCA-BP 模型预测结果

表 1. 10 次运行决定系数统计信息

模型	最大值	最小值	平均值	总运行时间/s	单次运行时间/s
传统 BP 神经网络	0.9551	0.6812	0.8450	19.7523	1.9752
PCA-BP 神经网络	0.9811	0.7882	0.8776	1.6355	0.1635

传统 BP 神经网络模型的决定系数。PCA 降维剔除了原始光谱中的冗余特征与多重共线性干扰，精简了模型输入维度，缓解了高维小样本条件下的过拟合现象，因此模型的整体拟合泛化精度得到提升。在运算效率层面，PCA-BP 模型平均计算耗时显著低于传统 BP 模型。传统 BP 模型直接以 401 维特征作为网络输入，输入层神经元数量庞大，网络待训练参数量多；而经过 PCA 降维后仅使用 4 维主成分作为输入，网络拓扑规模大幅缩减，模型迭代训练的计算开销降低，运行效率得到明显改善。综合精度与运算效率两项指标，相比于传统 BP 神经网络，PCA-BP 组合模型在高维小样本拟合任务上具备更优的预测性能与计算效率。

5. 总结

本文针对工程领域常见的高维小样本非线性建模问题，以传统 BP 神经网络为基准，构建基于 PCA 降维预处理的 PCA-BP 组合预测模型。基于 MATLAB 平台开展对照仿真实验，对传统 BP 神经网络与 PCA-BP 模型分别进行 10 次独立重复仿真，统计模型预测性能。结果表明：相较于原始光谱输入的 BP 神经网络，PCA-BP 模型利用少量核心主成分作为模型输入，模型拟合精度与结果稳定性得到改善；同时 PCA 降维大幅削减输入特征维度，简化神经网络拓扑结构，减少模型训练计算开销，加快网络收敛，并缓解高维小样本场景下的过拟合风险。仿真结果表明，PCA 预处理能够提取光谱数据中的核心有效信息，优化输入数据质量，改善 BP 神经网络的建模效果，该组合方案在工程预测场景具备一定实用价值。

基金项目

桂林理工大学科研启动基金(GLUTQD2018001)。

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] 徐锦, 笪跃武, 金一, 等. BP 神经网络模型用于水泵特性曲线拟合[J]. 中国给水排水, 2026, 42(2): 90-94.
- [2] 朱凌志, 宋云连. 基于 BP 神经网络的城镇沥青路面剪应力拟合计算[J]. 力学季刊, 2025, 46(2): 471-484.
- [3] 林瑞, 刘晓文, 丁纪昕, 等. 基于 GA-BP 的大气帆气动函数拟合与控制矩阵计算[J]. 中国空间科学技术(中英文), 2025, 45(3): 131-142.
- [4] 彭中波, 高阳. BP 神经网络在水下地形高程拟合的应用[J]. 重庆交通大学学报(自然科学版), 2018, 37(11): 64-68.
- [5] 魏巍, 师娅. 改进 PSO-BP 算法在函数拟合中的应用[J]. 微电子学与计算机, 2017, 34(9): 112-115.
- [6] 叶明壕, 牟晓玉. 基于 BP 神经网络的船舶短期电力负荷预测方法[J]. 舰船科学技术, 2026, 48(4): 84-88.
- [7] 郝彬, 田增顺, 马卫建, 等. 基于 PSO-BP 神经网络的既有桩基极限承载力预测

- 研究[J]. 山东农业大学学报(自然科学版), 2026, 57(1): 189-198.
- [8] 胡鸿鹏, 蒋水华, 孟京京, 等. 基于主成分分析的结构参数自适应降维贝叶斯更新方法[J]. 应用基础与工程科学学报, 2026, 34(4): 1135-1149.
- [9] 李晓娟, 张芳媛, 喻玲. 基于主成分分析-BP 神经网络的风电备件需求预测[J]. 科学技术与工程, 2024, 24(1): 281-288.
- [10] 罗正亮, 潘虹, 赵雷, 等. 基于 PCA-GA-BP 神经网络的水电机组状态异常辨别方法[J]. 排灌机械工程学报, 2022, 40(4): 372-377.
- [11] 王小川, 史峰, 郁磊, 等. MATLAB 神经网络 30 个案例分析[M]. 北京: 北京航空航天大学出版社, 2010.

Appendix (Abstract and Keywords in Chinese)

基于 PCA 降维的高维小样本 BP 神经网络拟合研究

摘要: 针对高维小样本数据拟合过程中存在特征冗余、多重共线性, 造成模型泛化性能下降的问题, 本文采用主成分分析开展数据预处理降维研究。通过方差贡献率准则筛选有效主成分, 并利用数据重构方式检验降维信息保留效果。采用参数保持不变的 BP 神经网络作为统一拟合模型开展对照实验。结果表明, 经降维处理后测试集决定系数由 0.94671 提升至 0.9715。证实该降维方法能够剔除冗余干扰, 改善输入数据质量, 可用于高维小样本的拟合预处理。

关键词: 主成分分析, 数据预处理, 拟合, BP 神经网络