基于 PCA-BPNN 的学生写作成绩预测模型研究*

胡 帅 顾 艳 姜 华 曲巍巍
(渤海大学大学外语教研部 锦州 121013)

摘 要：针对传统学生英语写作成绩预测方法准确率偏低的情况，提出一种基于主成分分析(PCA)和 BP 神经网络相结合的写作成绩预测模型。首先，用 PCA 对所建立的学生写作评价体系作数据降维处理，提取前 3 个主成分，构建了新的样本矩阵，再对 BP 神经网络进行训练和泛化能力测试。仿真结果表明：单一的 BPNN 预测最大相对误差为 -2.165%，PCA-BPNN 预测最大相对误差仅为 -0.824 2%，PCA-BPNN 简化了网络结构，提高了单一 BPNN 的训练速率、预测精度和泛化能力，验证了所提出的模型的有效性。

关键词：主成分分析；BP 神经网络；成绩预测
中图分类号：TN957.52+9 TP183 **文献标识码**：A **国家标准学科分类代码**：520.2060

Study of student score prediction model based on PCA-BPNN

Hu Shuai Gu Yan Jiang Hua Qu Weiwei
(Teaching and Research Institute of Foreign Languages, Bohai University, Jinzhou 121013, China)

Abstract: In view of the low accuracy of traditional prediction method of students' English writing scores, a prediction model based on principal component analysis (PCA) and BP neural network was proposed. First, the dimensions of the evaluation system of students' writings were reduced by PCA. The first three principal components were extracted to create a new sample matrix. Then the BP neural network was trained and its generalization ability was tested. The simulation results show that the maximum relative error of prediction produced by the simple BPNN is -2.165%, while the one produced by the PCA-BPNN is only -0.824 2%. The PCA-BPNN simplifies the network structure. It also improves the training rate, prediction accuracy and generalization ability of the simple BPNN. The effectiveness of the proposed model is verified.

Keywords: principal component analysis (PCA); BP neural network; score prediction

1 引 言

对学生英语写作成绩进行准确预测可以为教师教学方案的调整和学生有针对性的训练提供重要依据。传统的预测方法采用简单的线性模型来预测，由于学生英语写作成绩受诸多因素影响，使写作成绩预测呈现高维、非线性特性[1-2]，所以，传统方法的预测结果误差较大，难以满足实际需要。伴随人工神经网络技术的飞速发展，基于神经网络的数据挖掘方法为解决学生英语写作成绩预测提供了新的方法。因为 BP 神经网络(back propagation neural networks, BPNN)算法简单、逼近精度高、具有良好的非线性映射能力[3-5]，所以被广泛应用于各个领域的预测，但 BPNN 存在收敛速度慢、容易陷入局部极小值等缺点[6-9]，文献[10]和文献[11]都利用单一的 BPNN 对学生成绩进行预测，通过改进传统 BP 算法，改善了网络逼近能力，但由于未考虑各评价指标之间存在的信息重叠现象，从而导致预测准确度不高。文献[12]采用改进的遗传算法对 BPNN 参数进行优化，提高了 BPNN 的预测精度，但算法复杂度较高且未考虑评价指标之间的权重，各因子之间仍然存在多重共线性。传统的成绩预测方法多数都只关注于 BPNN 自身算法的改进，而对于作为预测对象主体的评价指标之间的关联度考虑不够，所以，尝试利用信息熵方法计算出原始数据指标的权值，将 12 个指标的线性加权和作为最终评价结果，再用主成分分析方法(principal component analysis，PCA)对原始评价体系中的 12 项指标作数据降维处理，将提取的前 3 个主成分得分系数矩阵输入 BPNN，构建了 3 层 PCA-BPNN 写作成绩预测模型，并与单一的动量 BP 算法改进的 BPNN 模型作对比，验证 PCA-BPNN 预测模型的有效性。

收稿日期：2015-08
* 基金项目：辽宁省教育厅科学研究一般项目(W2015015)、辽宁省社会科学基金(L14CYY022)资助项目

2 PCA算法基本原理[13-15]

PCA实质是用数学统计的方法将多个相关程度较大的变量转化为少数几个互相无关的综合变量，降低原始样本的数据维数。假设样本矩阵 $X=(X_1,X_2,\cdots,X_n)$ 的样本容量为 n，每个样本有 m 个特征指标 $X_i=(X_{i1},X_{i2},\cdots,X_{in})$，$i=(1,2,\cdots,m)$，计算样本矩阵 X 的相关系数矩阵，如式(1)所示。

$$R_X=\frac{\sum_{i=1}^{n}(X_i-E(X))\cdot(X_i-E(X))^{\mathrm{T}}}{n}=a\cdot a^{\mathrm{T}} \quad (1)$$

式中：$E(x)=\frac{\sum_{i=1}^{n}X_i}{\sqrt{n}}$ 表示样本矩阵均值，$a=\sqrt{\frac{\sum_{i=1}^{n}(X_i-E(X))}{n}}$ 表示标准化后处理的样本矩阵。令 R_X 的特征值为 $\lambda_1,\lambda_2,\cdots,\lambda_m$，且 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m\geqslant0$，特征向量为 $w_1,w_2,\cdots,w_m$，则式(2)可以成立。

$$R_X\cdot w_i=\lambda_i\cdot w_i \quad i=(1,2,\cdots,m) \quad (2)$$

设特征向量矩阵 $w=[w_1,w_2,\cdots,w_m]$，按照式(3)计算可以建立新的样本矩阵 M。w 将 m 维样本矩阵变换为等维的样本矩阵，M 中任一元素 M_{ij} 表示 X 中 x_i 样本的第 j 个主分量。前 p 个主成分的累计贡献率计算方法如式(4)所示。保留重构样本矩阵中的主要分量，剔除次要分量，当累计贡献率 $C_{1-p}>0.85$ 时，可以用前 p 个主成分代替初始的 m 个特征指标，从而达到数据降维目的。

$$M=a\cdot w_i^{\mathrm{T}} \quad (3)$$

$$C_{1-p}=\frac{\sum_{n=1}^{p}\lambda_n}{\sum_{m=1}^{m}\lambda_m} \quad (p<m) \quad (4)$$

3 PCA-BPNN预测模型的建立

3.1 原始数据的采集

要对学生英语写作成绩做出准确预测，前提是具备完整可靠的历史数据，所以原始数据的采集过程至关重要。首先，本文综合考虑了影响写作成绩的诸多因素，查阅大量的相关文献并结合英语教学领域的专家意见，建立了包含有12项评价指标的学生英语写作成绩评价体系。各个指标含义如下：X_1－口语水平、X_2－听力测试、X_3－词汇量、X_4－语法技能、X_5－阅读能力、X_6－翻译水平、X_7－学习动机、X_8－学习兴趣、X_9－跨文化交际能力、X_{10}－写作策略、X_{11}－语篇知识、X_{12}－英美文化知识。然后，按照每项指标满分为10分制定评分表，并给出详细评分标准。最后，请25位英语专业教师采用面试、笔试的形式对某高校非英语专业的2个教学自然班学生(共计60人)的统一命题作文(包括议论文、图表类作文、信函类作文)按照评分表逐项打分，为确保数据的有效性，将获得的原始数据中的倾向性数据剔除，对60个样本的每一项指标得分均去掉3个最低分和3个最高分后取平均值，依次得到60位学生的12项指标的得分值[16]，为了尽可能避免评价过程中主观因素的影响，采用信息熵方法计算出各评价指标的权重，将12项指标的线性加权和作为最终评价结果。学生英语写作评价原始数据如表1所示。

表1 学生英语写作评价原始数据

样本编号	X_1	X_2	X_3	…	X_{10}	X_{11}	X_{12}	评价结果
1	9.32	9.56	9.78	…	8.59	8.42	8.24	8.72
2	8.47	9.18	8.92	…	8.22	8.18	6.71	8.44
3	9.10	9.21	9.63	…	8.47	9.23	8.20	8.94
4	8.63	8.30	9.26	…	8.06	8.23	6.49	8.24
5	8.31	8.69	9.58	…	6.74	6.72	5.30	7.17
⋮	⋮	⋮	⋮	…		⋮	⋮	⋮
56	9.11	8.27	9.09	…	8.50	7.11	7.23	8.16
57	9.31	8.02	9.10	…	6.92	7.21	7.26	7.70
58	9.08	7.30	8.81	…	8.03	6.96	6.86	7.45
59	7.56	8.29	7.33	…	5.9	6.91	6.04	7.23
60	7.64	8.11	7.55	…	6.00	7.21	6.11	7.27

3.2 主成分的提取与样本集的重构

3.2.1 计算原始样本数据的相关系数矩阵

对表1中的原始数据作标准化处理，按式(1)计算可以获得如表2所示主成分相关系数矩阵 R。从表2可以看出，指标 X_1、X_2、X_3 与其他指标之间的相关系数均较大，指标 X_{10} 与指标 X_1、X_2、X_3、X_7、X_8、X_9、X_{11}、X_{12} 之间的相关系数较大，X_{12} 与指标 X_{10}、X_{11} 之间的相关系数也较大等。这表明原始指标之间的信息相互干扰现象严重，若直接将原始指标输入BPNN，势必影响BPNN的收敛速度和预测准确率，所以，有必要用PCA对原始数据作降维处理。

表2 相关系数矩阵

指标	X_1	X_2	X_3	…	X_{10}	X_{11}	X_{12}
X_1	1.000 0	0.730 0	0.772 8	…	0.768 6	0.727 9	0.366 5
X_2	0.730 0	1.000 0	0.616 8	…	0.677 8	0.830 6	0.428 3
X_3	0.772 8	0.616 8	1.000 0	…	0.780 1	0.677 5	0.493 9
X_4	0.808 4	0.646 0	0.665 8	…	0.554 3	0.567 6	0.028 8
X_5	0.823 4	0.693 8	0.701 2	…	0.575 7	0.664 4	0.069 1
X_6	0.823 1	0.778 5	0.639 3	…	0.582 9	0.714 7	0.085 8
X_7	0.782 5	0.824 6	0.637 4	…	0.678 6	0.779 6	0.255 4
X_8	0.878 8	0.821 6	0.771 2	…	0.776 2	0.867 7	0.374 9
X_9	0.905 2	0.748 7	0.776 6	…	0.777 5	0.819 1	0.348 4
X_{10}	0.768 6	0.677 8	0.780 1	…	1.000 0	0.807 8	0.690 8
X_{11}	0.727 9	0.830 6	0.677 5	…	0.807 8	1.000 0	0.645 6
X_{12}	0.366 5	0.428 3	0.493 9	…	0.690 8	0.645 6	1.000 0

3.2.2 计算 R 的特征值、贡献率与主成分提取

根据式(3)和式(4)计算相关系数矩阵R的主成分特征值、贡献率和累计贡献率，如表3所示。有3个较大的

特征值，其值分别为 8.949 04、1.559 19、0.544 019。这说明第 1 个主成分的贡献率最大，它包含 74.58%的原始变量所含信息；第 2 个主成分的贡献率次之，它包含 12.99%的原始变量所含信息；第 3 个主成分的贡献率相对较小，它包含 4.53%的原始变量所含信息；前 3 个主成分的累计贡献率达到 92.10%。所以，根据主成分选取原则，可以选取前 3 个主成分代替原始的 12 个指标，达到数据降维目的，此时信息的损失率仅为 7.90%。

表 3　主成分特征值、贡献率与累计贡献率

主成分编号	特征值	贡献率(%)	累计贡献率(%)
1	8.949 04	74.58	74.58
2	1.559 19	12.99	87.57
3	0.544 019	4.53	92.10
4	0.217 576	1.81	93.91
5	0.183 369	1.53	95.44
6	0.159 443	1.33	96.77
7	0.119 883	1.00	97.77
8	0.107 274	0.89	98.66
9	0.053 134 4	0.44	99.10
10	0.046 966 1	0.39	99.49
11	0.035 266 3	0.29	99.78
12	0.024 836 7	0.21	100.00

3.2.3　计算主成分特征向量

将标准化的样本矩阵(60×12 维)与 3.2.2 节中提取到的前 3 个主成分的特征向量矩阵(12×3 维)相乘，可以建立 PCA-BPNN 预测模型的重构的样本集(60×3 维)，如表 4 所示。

表 4　重构的样本集

样本编号	f_1	f_2	f_3	评价结果
1	2.721 7	0.552 1	−0.124 4	8.72
2	1.809 3	−1.100 9	−0.603 4	8.44
3	3.420 2	0.024 6	−0.305 7	8.94
4	1.152 3	−0.991 0	0.093 2	8.24
5	−1.190 0	−2.715 6	1.086 9	7.17
⋮	⋮	⋮	⋮	⋮
56	1.480 8	−0.918 4	0.785 2	8.16
57	0.368 3	−1.102 1	1.274 7	7.70
58	−0.536 3	−0.733 4	1.646 6	7.45
59	−1.802 5	−2.208 2	−0.863 0	7.23
60	−1.995 7	−1.756 7	−0.949 3	7.27

3.2.4　PCA-BPNN 预测模型的训练

为了对比说明 PCA-BPNN 预测模型的有效性，本文同时建立了动量 BP 算法改进的 BPNN。以表 1 中的 1～40 号样本作为训练样本集，经过反复试验，BPNN 的隐含层神经元数为 15 时，模型的收敛速度最快，训练误差最小，所以，最终确定 BPNN 拓补结构为 12－15－1。由3.2.2节可知，由于提取了 3 个主成分，故将表 4 中的 1～40 号样本作为训练样本集输入到 BPNN，因为输入样本向量维数较之前明显降低，故 BPNN 隐含层神经元数目需要重新确定，结合经验公式并经过反复试验，当隐含层神经元数为 12 时，网络性能最佳，所以 PCA-BPNN 的拓补结构为 3－12－1。两种模型的转换函数均采用 Sigmoid 型函数，隐含层传递函数采用 tansig 函数，输出层传递函数采用 purelin 函数；利用 traingdm 函数训练两种网络。目标精度均设为 0.001、最大训练次数设为 20 000、学习速率为 0.1。

BPNN 和 PCA-BPNN 模型输出与目标值相关性曲线如图 1 和图 2 所示。线性相关系数 R 表明拟合直线对样本点的拟合程度。R 值越大，越接近于 1，表明拟合程度越高，训练样本集的各个样本点聚集在拟合直线周围越紧密，越接近于理想拟合直线，网络的训练输出结果与目标值越接近；R 值越小，越接近于 0，表明拟合程度越低，训练样本集的各个样本点在拟合直线周围越分散，实际拟合直线越偏离理想拟合直线，网络的训练输出结果与目标值相差越大。由图 1 和图 2 可以看出，BPNN 模型 R=0.958 94，PCA-BPNN 模型 R=0.992 17，这表明与 BPNN 相比，PCA-BPNN 的学习精度显著提高。仿真实验中同时发现，在相同目标精度情况下，BPNN 需要 3 897个迭代周期才能收敛，而 PCA-BPNN 只需要 231 个迭代周期就可以达到目标精度，在收敛速度上较单一的 BPNN 也具有明显优势。

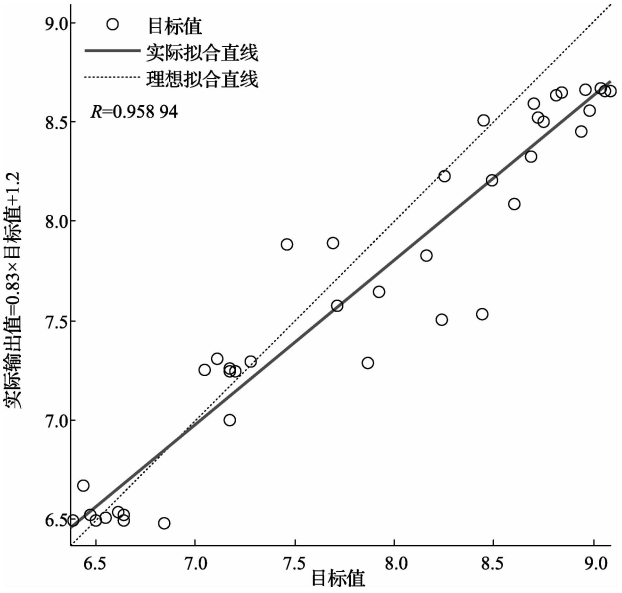

图 1　BPNN 模型输出与目标值相关性曲线

4　PCA-BPNN 预测模型的检验

将表 4 中 41～60 号样本作为 PCA-BPNN 模型的测试样本集，将表 1 中 41～60 号样本作为单一的 BPNN 模型的测试样本集，对两种模型作泛化能力测试。两种模型对于测试样本集的相对误差曲线如图 3 所示。

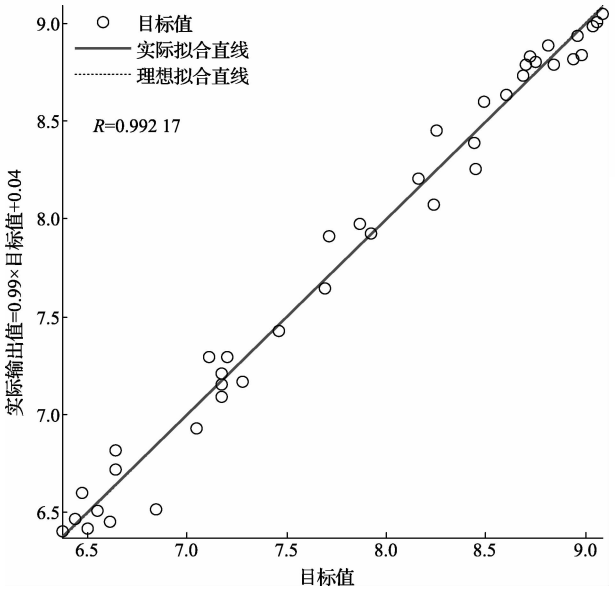

图 2 PCA-BPNN 模型输出与目标值相关性曲线

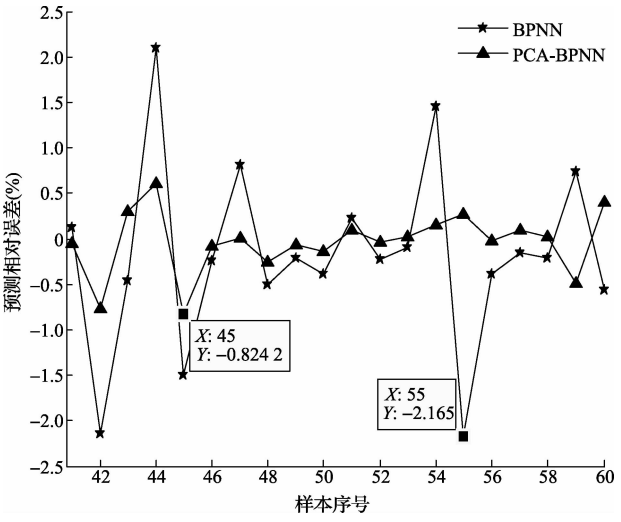

图 3 两种模型对于测试样本集的相对误差

可以看出，PCA-BPNN 预测的最大相对误差仅为−0.824 2%，单一 BPNN 预测模型的最大相对误差已经达到−2.165%；测试样本集范围内 PCA-BPNN 误差波动范围明显较 BPNN 小；PCA-BPNN 在测试集各样本观测点预测值相对误差也明显小于 BPNN，这说明所建立的 PCA-BPNN 的逼近能力比单一 BPNN 更强。这是因为 PCA-BPNN 拓补结构更为简单，模型复杂度远远低于单一的 BPNN，这也最终决定了 PCA-BPNN 具有更快的收敛速度。

5 结 论

将 PCA 与 BPNN 结合在一起，提出了一种基于 PCA-BPNN 的学生写作成绩预测模型。利用 PCA 选取前 3 个主成分，对原始复杂的指标体系作降维处理，构造了新的样本矩阵，减少了输入样本之间的信息重叠，尽可能地消除了样本之间的相互干扰，简化了 BPNN 结构，提高了 BPNN 的训练速率、预测精度和泛化能力。本文所建立的模型具有较强通用性，只需要对于后台运行程序稍作修改即可用于其他领域的包含多维指标的预测分析、分类以及评价等。

参 考 文 献

[1] 黄建明. 贝叶斯网络在学生成绩预测中的应用[J]. 计算机科学，2012，39(11)：280-282.

[2] 陈益均，殷莉. 基于数据挖掘的学生成绩影响模型的研究[J]. 现代教育技术，2013，23(1)：94-96.

[3] DING S，WU Q H. Research on inverse model based on ANN and analytic method for induction motor[J]. Automation and Control，2011，5(4)：356-370.

[4] 丁硕，巫庆辉. 基于改进 BP 神经网络的函数逼近性能对比研究[J]. 计算机与现代化，2012，1(11)：10-13.

[5] 庄育锋，胡晓瑾，翟宇，等. 基于 BP 神经网络的微量药品动态称重系统非线性补偿[J]. 仪器仪表学报，2014，35(8)：1914-1920.

[6] 韩春玉，王玉影，张帆，等. 基于神经网络的 NTC 热敏电阻的校正模型[J]. 电子测量技术，2013，36(9)：5-8，22.

[7] 郑永，陈艳. 基于 BP 神经网络的高校教师教学质量评价模型[J]. 重庆理工大学学报：自然科学版，2015，29(1)：85-90.

[8] 丁硕，常晓恒，巫庆辉，等. 基于 GRNN 与 BPNN 的二维向量模式分类对比研究[J]. 国外电子测量技术，2014，33(5)：56-59.

[9] 刘征宇，杨俊斌，张庆，等. 基于 QPSO-BP 神经网络的锂电池 SOC 预测[J]. 电子测量与仪器学报，2013，27(3)：224-228.

[10] 邹丽娜，丁茜. 基于 BP 算法的成绩预测模型[J]. 沈阳师范大学学报：自然科学版，2011，29(2)：226-229.

[11] 张宇，袁晓曦，弓小倩，等. 基于 BP 神经网络算法的体育成绩预测研究[J]. 科技通报，2013，29(6)：149-151.

[12] 罗永国. 基于改进的遗传算法的学生成绩预测模型[J]. 科技通报，2012，28(10)：223-225.

[13] 孙健，王成华，洪峰，等. 基于 PCA-LVQ 的模拟电路故障诊断[J]. 电路与系统学报，2013，18(2)：188-192.

[14] 张宁，任茂文，刘萍. 基于主成分分析和 BP 神经网络的煤岩界面识别[J]. 工矿自动化. 2013. 39(4)：55-58.

[15] 孙健，王成华，闫之烨，等. 基于 PCA 和 PNN 的模拟电路故障诊断[J]. 微电子学，2014，44(1)：123-126.

[16] 魏正元，颜克胜，苏盈盈. 基于最大熵 NN 的教学质量评价模型及仿真[J]. 计算机仿真，2013，30(5)：284-287.

作 者 简 介

胡帅，1980 年出生，硕士，讲师。主要研究方向为语料库语言学、神经网络理论及其应用。
E-mail：hushuai6@163.com