PCA 和 Elman 网络在移动学习策略分类中的应用*

胡 帅 程迎新 顾 艳
(渤海大学大学外语教研部 锦州 121013)

摘 要：针对传统的大学生英语移动学习策略分类方法准确率较低的情况，提出了一种主成分分析(PCA)和 Elman 神经网络相结合的分类模型。首先，用 PCA 对所获得的移动学习策略原始数据作数据降维处理，提取前 5 个主成分，建立新的特征样本矩阵，再对 Elman 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 分类准确率为 70.0%，单一的 Elman 网络分类准确率为 80.0%，PCA-Elman 网络分类准确率为 100.0%，PCA-Elman 网络模型简化了单一 Elman 网络的结构，提高了网络的训练速率、分类准确率和泛化能力，验证了所提出的模型的有效性。

关键词：主成分分析；Elman 神经网络；BP 神经网络；移动学习策略；分类
中图分类号：TP183 **文献标识码**：A **国家标准学科分类代码**：520.2060

Application of PCA and Elman network in mobile learning strategy classification

Hu Shuai Cheng Yingxin Gu Yan
(Teaching and Research Institute of Foreign Languages, Bohai University, Jinzhou 121013, China)

Abstract: To overcome the problem of low accuracy of traditional methods in the area of college student mobile learning strategy classification, a new classification model based on principal component analysis (PCA) and Elman neural network is proposed. First, dimensionality reduction was done to the obtained original data of student mobile learning strategies using PCA and 5 principal components were extracted to create a new feature sample matrix. Then the Elman neural network was trained and its generalization performance was tested. The simulation results indicate that: the classification accuracy of the single BPNN is 70.0%, the one of the single Elman model is 80.0% and the one of the PCA-Elman model is 100.0%; the PCA-Elman model can simplify the structure of the single Elman network, improve the training speed, classification accuracy and generalization performance; the effectiveness of the recommended model is proved.

Keywords: principal component analysis; Elman neural network; BP neural network; mobile learning strategy; classification

1 引 言

对大学生移动学习策略分类的研究是对不同学生有针对性地进行移动学习策略培养的前提和基础。由于影响大学生英语移动学习策略的因素众多，使其分类呈现高维、非线性特性，近年来，随着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大学生英语移动学习策略的分类提供了新的方法。在众多的人工神经网络类型中，BP 神经网络(back propagation neural networks, BPNN)是应用最为广泛的一类网络。但 BPNN 存在收敛速度慢、易陷入局部极小值等缺点[1-4]，文献[5]利用问卷调查和多元回归分析对大学生英语移动学习策略进行分类，但由于未考虑各评价指标之间存在的信息重叠现象，从而导致分类准确度不高。文献[6-7]都通过对标准 BP 算法进行了改进，实现了 BPNN 分类准确度的提高，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。Elman 神经网络是一种从输出到输入具有反馈连接的神经网络，逼近能力优于一般的静态网络，收敛速度快，能较好地克服 BP 网络训练时间长及计算复杂度高等缺点。传统的分类方法多数都只关注于网络自身算法的改进，而对于作为主体的影响分类结果的各个评价指标之间的关联度关注不够，所以，本文尝试利用主成分分析方

收稿日期：2015-10
* 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)、辽宁省社会科学基金重点项目(L15AYY001)资助

法(principal component analysis,PCA)对原始评价体系中的 22 项指标作数据降维处理,提取出前 5 个主成分,构建新的特征样本矩阵,并将其输入到 Elman 神经网络,建立了 PCA-Elman 移动学习策略分类模型,并与单一的 BPNN 模型和单一的 Elman 网络分类模型的分类结果作对比,以验证 PCA-Elman 分类模型的有效性。

2　PCA 算法原理

主成分分析(principal component analysis,PCA)的实质是在保证原有信息含量前提下,对包含有大量相关信息的原始数据集作数据降维处理,以较少量的指标取代原始数据集,这些互不相关的指标被称为主成分[8]。

设输入样本矩阵 $\boldsymbol{X}=(x_1,x_2,\cdots,x_n)$ 的样本容量为 n,每个样本有 m 个特征指标,则 $\boldsymbol{X}$ 的协方差矩阵如下:

$$Cor(\boldsymbol{X})=\frac{\sum_{k=1}^{n}[x_k-E(\boldsymbol{X})][x_k-E(\boldsymbol{X})]^{\mathrm{T}}}{n}=\boldsymbol{M}\cdot\boldsymbol{M}^{\mathrm{T}} \tag{1}$$

式中:$E(\boldsymbol{X})=\frac{\sum_{k=1}^{n}x_k}{n}$ 表示输入样本矩阵的均值,求得 $Cor(\boldsymbol{X})$的特征值为 $t_1,t_2,\cdots,t_m$,且 $t_1\geqslant t_2\geqslant\cdots\geqslant t_m\geqslant 0$,相应的特征向量为 $\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_m$[9]。

求前 p 个主成分的累计贡献率,如式(2)所示,当前 p 个主成分的累计贡献率大于 85%时,即可用前 p 个主成分作为新的样本特征,新的样本特征矩阵计算方法如式(3)所示,式(3)中 $\boldsymbol{V}(\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_p)$,$\boldsymbol{S}(\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_p)$[10-12]。

$$C=\sum_{l=1}^{p}t_l/\sum_{j=1}^{m}t_j\quad(p<m) \tag{2}$$

$$\boldsymbol{S}=\boldsymbol{V}^{\mathrm{T}}\cdot\boldsymbol{X} \tag{3}$$

3　Elman 网络的结构与算法

Elman 神经网络由输入层、隐含层、承接层、输出层 4 层组成,比 BPNN 多一个承接层。输入层可以实现信号传输作用,输出层可以实现输入向量线性加权和功能,隐含层神经元采用线性或非线性的传递函数,承接层神经元具有记忆功能,即可以将隐含层单元上一过程的输出值反馈至隐含层,正是因为增加了这样一个反馈支路,使得 Elman 神经网络具有动态记忆特性,其逼近能力较传统的静态网络更强,所以 Elman 网络可以较好地完成时域和空域的模式识别[13-14]。Elman 网络的数学模型可以表示为:

$$\boldsymbol{Y}(n)=T(\boldsymbol{W}^3X(n)) \tag{4}$$

$$X(n)=f(\boldsymbol{W}^1X_c(n)+W^2(u(n-1))) \tag{5}$$

$$X_c(n)=\beta\cdot X(n-1)+X(n-1) \tag{6}$$

式(4)中:$\boldsymbol{Y}(n)$ 为输出向量,$T(\cdot)$ 为输出层的传递函数,$\boldsymbol{W}^3$ 为隐含层与输出层之间的连接权值矩阵,$x(n)$表示隐含层单元的输出。式(5)中,$f(\cdot)$ 为非线性作用函数,一般取 Sigmoid 函数,W^2 为输入层与隐含层单元之间的连接权值矩阵,$\boldsymbol{W}^1$ 为承接层与隐含层单元之间的连接权值矩阵。$u(\cdot)$ 为阈值函数。式(6)中,β 为自连接反馈增益,$x_c(n)$表示承接层单元的输出。Elman 网络采用如式(7)所示的误差平方和函数作为学习的目标函数,如式(7)所示,式(7)中,$\hat{Y}_k(w)$ 为目标输出量[15]。Elman 网络采用自适应学习速率的动量梯度下降算法,在提高网络的训练速率同时,又能有效抑制标准 BPNN 陷入局部极小值的缺点。

$$Err(w)=\sum_{k=1}^{n}[Y_k(w)-\hat{Y}_k(w)]^2 \tag{7}$$

4　PCA-Elman 分类模型的建立

4.1　评价指标体系的构建

本文对某综合性大学大二年级中随机选取的 2 个自然班,共 60 名学生进行了外语移动学习策略问卷调查。问卷内容以 Oxford 语言学习策略调查问卷为参照,该问卷主要考察学习者的外语学习策略,问卷编制同时借鉴 LASSI 在线学习版中反应学习者移动学习策略的内容。编制的问卷含有认知、元认知、情感、资源管理和信息处理 5 个一级指标。其中认知策略含有 7 个二级指标:注意策略、组织策略、复述策略、注意策略、精细加工策略、问题解决策略、自主学习策略;元认知策略含有 3 个二级指标:计划策略、监控策略、调节策略;情感策略含有 4 个二级指标:情绪控制策略、合作策略、焦虑控制策略、激励策略;信息处理策略含有 4 个二级指标:信息检索策略、信息甄别策略、信息整理策略、信息重构策略;资源管理策略含有 4 个二级指标:信息渠道管理策略、时间管理策略、心境管理策略、社会性人力资源管理策略。共含有二级指标 22 个,每个二级指标下设 5 个题目,每题得分为 0～2 分。回收问卷后将不合格问卷剔除,最终获得有效问卷 60 份。对问卷得分进行统计并交由 7 位专家分别对 60 份样本进行分类,拟分为 5 类。之后专家对分类结果进行讨论,最终反馈给 2 个班的英语教师进行进一步调整与确认,最终得到用于分类的大学生外语移动学习策略原始数据(标准化处理后),如表 1 所示。

表 1　学生外语移动学习策略评价数据表(标准化处理后)

样本编号	X_1	X_2	…	X_{21}	X_{22}	评价结果
1	1.471 3	1.894 4	…	1.659 9	1.715 2	1
2	1.777 2	1.883 2	…	1.808 2	1.714 7	1
3	−2.319 6	−1.099 0	…	−0.417 3	−1.039 6	2
4	−1.391 0	−1.648 3	…	−0.343 1	−0.658 2	2
5	−1.664 1	−1.984 7	…	−1.174 0	−2.031 3	2
⋮	⋮	⋮	…	⋮	⋮	⋮
56	0.225 9	0.706 0	…	−0.417 3	−0.463 2	4
57	−0.407 8	0.537 9	…	−0.306 0	−0.403 9	4
58	−0.156 5	0.706 0	…	−0.825 3	−0.632 8	4
59	−0.134 6	−0.213 3	…	−0.454 4	−1.039 6	5
60	−0.243 9	−0.303 0	…	−0.951 4	−0.514 1	5

4.2　主成分的提取与新的样本集矩阵的建立

4.2.1　计算原始样本数据的相关系数矩阵

为了获得较快的收敛速度，先对获得的原始数据作标准化处理，按照式(1)方法计算得到经过标准化处理后的原始数据的主成分相关系数矩阵 $\boldsymbol{R}$，如表 2 所示。可以看出，原始数据的诸多评价指标之间的相关系数较大，这表明原始指标之间存在着较为严重的信息相互干扰现象，如果直接将表 1 中的 22 个原始指标作为分类的特征数据，则会由于原始指标之间的存在较大的相关性而导致 Elman 网络的分类算法性能下降，并且会占用大量的机器处理时间和存储空间，最终导致 Elman 网络的收敛速度下降、分类准确率降低。

表 2　相关系数矩阵 $\boldsymbol{R}$

指标	X_1	X_2	X_3	…	X_{21}	X_{22}
X_1	1.000 0	0.721 0	0.736 9	…	0.724 2	0.734 5
X_2	0.721 0	1.000 0	0.708 5	…	0.708 1	0.765 5
X_3	0.736 9	0.708 5	1.000 0	…	0.604 3	0.637 9
X_4	0.657 8	0.661 2	0.720 6	…	0.538 6	0.519 2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
X_{19}	0.734 4	0.680 8	0.684 0	…	0.796 7	0.765 7
X_{20}	0.697 1	0.689 1	0.578 3	…	0.816 6	0.817 5
X_{21}	0.724 2	0.708 1	0.604 3	…	1.000 0	0.783 6
X_{22}	0.734 5	0.765 5	0.637 9	…	0.783 6	1.000 0

4.2.2　计算 R 的特征值、贡献率与主成分提取

根据式(2)和(3)计算得到 R 的主成分特征值、贡献率和累计贡献率，如表 3 所示。从表 3 可以发现：当主成分提取到第 5 个时，前 5 个主成分的累计贡献率达到 85.99%，涵盖了原始 22 个特征数据所反映的 85%以上的信息，表明此时符合主成分选取的原则。因此，本文提取了原始数据的前 5 个主成分作为最终特征数据，建立了新的样本集矩阵，经过归一化处理后新的样本集如表 4 所示。

表 3　主成分特征根、贡献率和累计贡献率

主成分序号	特征值	贡献率/(%)	累计贡献率/(%)
1	14.027 7	63.76	63.76
2	2.837 17	12.90	76.66
3	0.952 569	4.33	80.99
4	0.562 703	2.56	83.55
5	0.493 282	2.24	85.99
6	0.445 232	2.02	88.01
⋮	⋮	⋮	⋮
21	0.048 760 4	0.22	99.88
22	0.026 518 1	0.12	100.00

4.2.3　计算主成分特征向量

将表 1 中经过标准化处理后的样本矩阵(60×22)与 4.2.2 节中提取到的前 5 个主成分的特征向量矩阵(22×5)相乘，可以建立 PCA-Elman 分类模型的新的样本集(60×5)，如表 4 所示。

表 4　新的样本集(归一化处理后)

样本编号	P_1	P_2	P_3	P_4	P_5	评价结果
1	0.470 0	0.468 6	0.497 7	0.497 7	0.947 5	1
2	0.467 5	0.470 3	0.442 9	0.500 5	0.983 2	1
3	0.465 1	0.468 3	0.463 2	0.483 5	0.199 2	2
4	0.466 4	0.471 3	0.462 1	0.478 4	0.225 3	2
5	0.458 6	0.457 9	0.443 0	0.470 4	0.060 4	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
56	0.458 1	0.454 3	0.476 7	0.537 2	0.529 4	4
57	0.460 3	0.450 4	0.485 0	0.559 7	0.474 2	4
58	0.457 1	0.454 8	0.474 5	0.544 1	0.502 2	4
59	0.452 6	0.452 8	0.429 1	0.493 4	0.414 4	5
60	0.459 7	0.445 8	0.431 6	0.480 6	0.389 7	5

4.2.4　PCA-Elman 模型的参数设置

因为之前通过主成分分析提取了 5 个主成分，故后端 Elman 网络的输入层神经元数为 5；输出层神经元个数与目标类别数目相等，本文用 5 位代码(1 0 0 0 0)表示第 1 类、(0 1 0 0 0)表示第 2 类、(0 0 1 0 0)表示第 3 类、(0 0 0 1 0)表示第 4 类、(0 0 0 0 1) 表示第 5 类；隐含层对 Elman 网络的逼近能力及收敛速率影响较大，本文经过反复试验发现隐含层与承接层神经元数分别为 22 和 5 时，Elman 网络的分类能力最强且收敛速率最快。所以，最终确定 Elman 网络的拓补结构为 5-22-5-5；隐含层神经元的传递函数采用 *tansig*，承接层神经元的传递函数采用 *logsig*，目标误差设为 0.001，学习速率设为 0.1。

5　仿真实验

5.1　PCA-Elman 分类模型的训练

为了对比说明 PCA-Elman 分类模型的有效性，本文同时建立了单一的 Elman 网络分类模型和单一 BPNN 分类模型。

在建立单一 BPNN 时，将表 1 中的样本数据分为两部分：1～40 号样本作为 BPNN 的训练样本集，41～60 号样本作为 BPNN 的测试样本集。BPNN 采用典型的单隐含层结构，输入层节点数等于每个样本的特征向量维数，即 22 个评价指标；输出层节点数与分类结果数量一致，即等于 5；隐含层节点数根据隐含层节点计算的经验公式，并经反复多次试验最终确定隐含层节点为 16。所以，可以确定 BPNN 的拓补结构为 22-16-5；采用标准梯度下降算法训练

函数 *traingd* 对网络的进行训练;隐含层传递函数采用 *tansig* 函数;输出层传递函数采用 *purelin* 函数。

将表 4 中新的样本集的 1-40 号样本作为 PCA-Elman 分类模型的训练样本集,41～60 号样本作为 PCA-Elman 作为测试样本集。将表 1 中的未经 PCA 处理的原始数据直接输入到单一的 Elman 网络分类模型和单一 BPNN 分类模型,表 1 中的前 40 个样本作为训练样本集,后 20 个样本作为测试样本集。在目标精度为 0.001、最大训练次数为 20 000、学习速率为 0.1 的前提下,分别对所建立的 3 种网络模型进行训练。为了能够更为直观地显示分类结果,利用函数 *vec2ind*()将输出层神经元输出得到的 5 位代码转换为单值向量进行输出。图 1 和 2 分别为单一 Elman 网络和 PCA-Elman 网络对训练样本的分类结果。

由图 1 可以看出,1、2、38、39、40 号训练样本被错误地划分为第 4 类,训练样本分类的正确率仅为 87.5%。这是因为第 1 类样本、第 5 类样本都与第 4 类样本的评价数据较为接近,彼此之间存在着较为严重的信息重叠现象,所以,将表 1 中的数据未做任何处理直接输入到 Elman 网络,这势必会降低网络的收敛速度,导致网络的学习效率降低和分类错误。

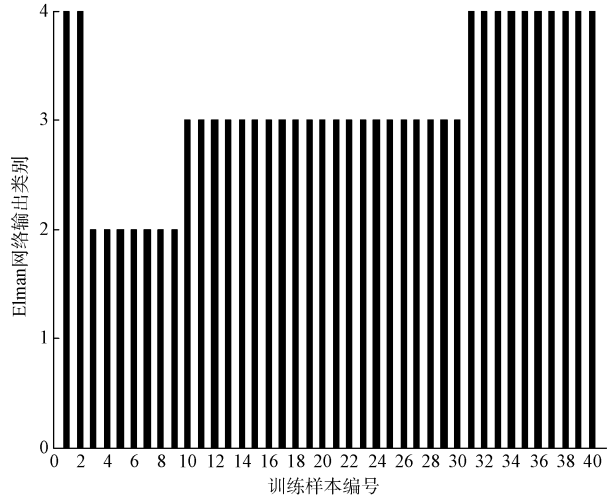

图 1　Elman 网络对于训练样本的分类结果

由图 2 可以看出,PCA-Elman 网络对训练样本分类的正确率达到 100%。可以看出,将经过 PCA 处理后的数据输入到 Elman 网络,使得 Elman 神经网络对于训练样本的分类准确率大幅度提高,这是因为经过 PCA 处理后的重构样本集在很大程度上避免了原始数据中的相互干扰,新的样本集的数据特征更为明显,从而提高了 Elman 神经网络的学习效率和准确率。

训练过程中可以发现,BPNN 需经过 13314 次迭代可以收敛;单一 Elman 网络需经过 6 147 次迭代可以收敛;PCA-Elman 网络则只需经过 21 次迭代就可以收敛;单一 BPNN 对训练样本集分类的正确率仅为 75.0%。这表明

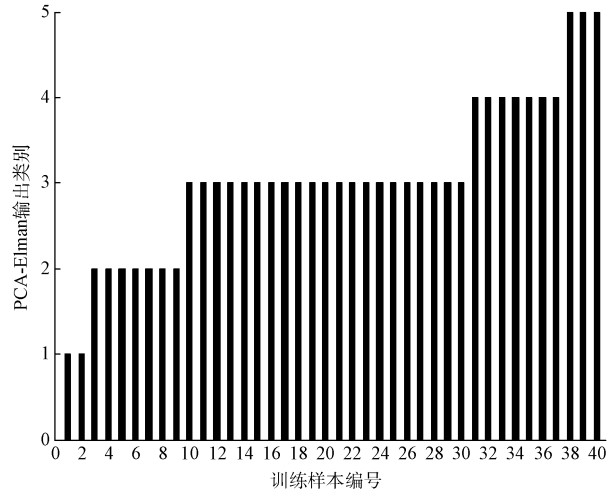

图 2　PCA-Elman 对于训练样本的分类结果

在 3 种模型的训练过程中,PCA-Elman 网络的收敛速度最快、分类的正确率最高,单一 Elman 网络次之、单一 BPNN 最差,即 PCA-Elman 网络分类模型的学习效率得到显著提高。

5.2　PCA-Elman 分类模型的泛化能力测试

图 3 和 4 分别为单一 Elman 网络和 PCA-Elman 网络对测试样本的分类结果。由图 3 可以看出,单一的 Elman 网络仅将 20 个测试样本划分为 4 类,其中第 41、42 号测试样本均被错误地划分为第 4 类,第 59、60 号测试样本被错误地划分为第 3 类,单一的 Elman 网络的分类准确率仅为 80.0%。由图 4 可以看出,PCA-Elman 网络将 20 个测试样本划分为 5 类,与表 1 中的实际评价结果完全一致,PCA-Elman 网络的分类准确率达到了 100.0%,仿真测试得到单一 BPNN 对测试样本集的分类准确率仅为 70.0%,这与另外 2 种网络模型的泛化能力测试结果相差很大。

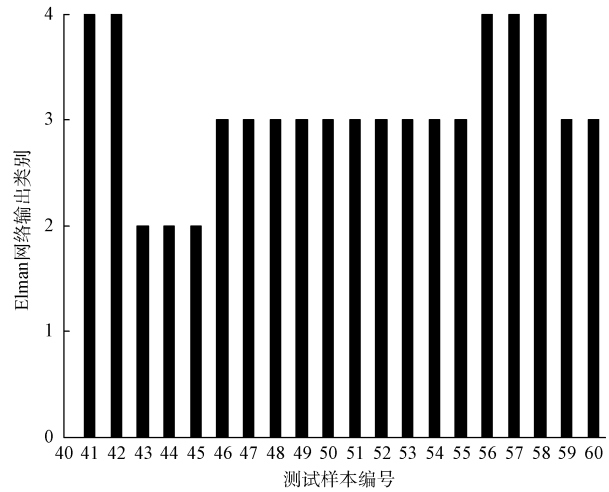

图 3　Elman 网络对于测试样本的分类结果

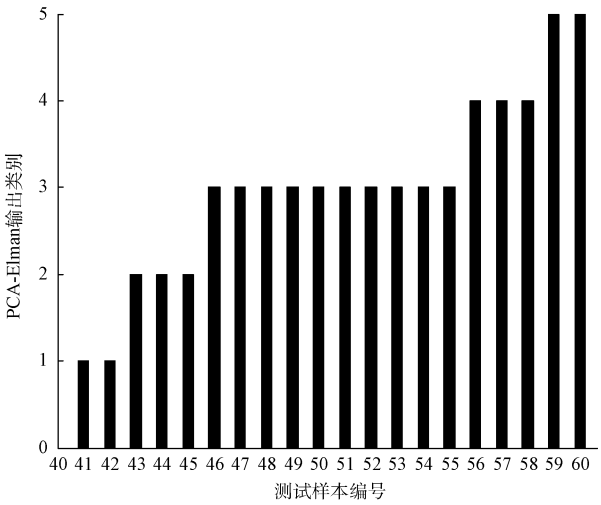

图 4　PCA-Elman 对于测试样本的分类结果

这是因为用 PCA 对表 1 中原始数据处理后提取了 5 个主成分，较原始的 22 个评价指标在数量上大为缩减，这 5 个主成分充分降低了原始数据集的信息冗余量，从而使得 Elman 网络的结构更为优化，充分降低了分类算法的复杂性，提高了 Elman 网络的分类识别能力。

6　结　　论

本文提出了一种 PCA 和 Elman 神经网络相结合的大学生的外语移动学习策略分类模型。该模型首先利用 PCA 方法将原始高维特征空间压缩到低维特征空间，再将得到的最终特征输入到 Elman 神经网络进行分类。仿真结果表明：采用 PCA 方法有效降低了初始特征空间信息间的相互干扰，相对于单一 Elman 神经网络和 BPNN，所提出的 PCA-Elman 模型在保证高准确率的同时提高了分类速度。

参考文献

[1] 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.

[2] 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.

[3] 丁硕，常晓恒，巫庆辉，等. DGA 与 GRNN 的联合变压器故障诊断研究[J]. 电子测量技术，2014，37(5)：142-145.

[4] 刘春，马颖. 遗传算法和神经网络结合的 PSD 非线性校正[J]. 电子测量与仪器学报，2015，29(8)：1157-1163.

[5] 王飞. 大学英语移动学习平台的构建[J]. 沈阳师范大学学报：自然科学版，2013，32(4)：561-564.

[6] 许玲，郑勤华. 大学生接受移动学习的影响因素实证分析[J]. 现代远程教育研究，2013，124(4)：61-66.

[7] 薛建强. 大学英语移动学习模式的构建与发展研究[J]. 实验技术与管理，2014，31(3)：176-179.

[8] 巫茜，蔡海尼，黄丽丰. 基于主成分分析的多源特征融合故障诊断方法[J]. 计算机科学，2011，38(1)：268-270.

[9] 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.

[10] 胡帅，顾艳，曲巍巍. 基于 PCA-LVQ 神经网络的教学质量评价模型研究[J]. 河南科学，2015，33(7)：1247-1252.

[11] 唐宏宾，吴运新，滑广军，等. 基于 PCA 和 BP 网络的液压油缸内泄漏故障诊断[J]. 中南大学学报：自然科学版，2011，42(12)：3709-3714.

[12] 胡帅，姜华，曲巍巍. 多元统计分析在外语教学质量评价中的应用[J]. 现代电子技术，2015，38(15)：126-129.

[13] 丁硕，常晓恒，巫庆辉，等. 基于 Elman 神经网络的传感器故障诊断研究[J]. 国外电子测量技术，2014，33(4)：72-75.

[14] 吕飞，沈振中. 基于 Elman 神经网络的面板堆石坝沉降预测模型[J]. 水电能源科学，2011，29(12)：56-59.

[15] 丁硕，常晓恒，巫庆辉，等. Elman 和 BP 神经网络在模式分类领域内的对比研究[J]. 现代电子技术，2014，37(8)：12-15.

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研究。

E-mail：hushuai6@163.com