ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 5,Sep. 2022,pp. 1012－1017
DOI:10. 16337/j. 1004-9037. 2022. 05. 006

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

主成分分析阈值选择差异性分析研究

张　婧[1]，刘　倩[2]

(1. 兰州石化职业技术大学,兰州 730070;2. 国家电网兰州供电公司,兰州 730050)

摘　要：主成分分析是特征提取和数据降维中常用的方法,在很多应用中一般选择平均特征值作为主成分选择的标准。但是主成分的多少与应用结果之间的关系目前还没有具体的分析结果。因此,提出一种主成分阈值选择差异性的实验分析方法,为不同应用中主成分分析阈值的选择提供依据。将本文分析方法应用于手写数字样本集MNIST进行降维处理,根据不同的阈值构建不同的神经网络进行分类,分析不同阈值下分类准确率的变化情况。实验结果表明主成分阈值选择在79%～81%之间(维度为41～50)时,分类准确率最高;低于或高于该区间,准确率随之下降。实验结果证明了主成分分析阈值的选择与应用结果之间不为正相关关系,且平均特征值不是一个硬性的选择标准。

关键词：主成分分析;阈值选择;神经网络;手写体数字分类

中图分类号：TP183　　**文献标志码**：A

Difference Analysis Research of Threshold Selection in Principal Component Analysis

ZHANG Jing[1], LIU Qian[2]

(1. Lanzhou Petrochemical University of Vocational Technology, Lanzhou 730070, China;2. State Grid Lanzhou Electric Power Supply Company, Lanzhou 730050, China)

Abstract: Principal component analysis (PCA) is a commonly used method for feature extraction and data dimension reduction. In many applications, the components whose eigenvalues are greater than the average value are retained. However, there is no specific analysis result for the relationship between the number of principal components and the application results. Therefore, an experimental analysis of the difference in selection of PCA threshold is carried out to provide basis for the PCA threshold selection in different applications. The experiment analysis is used to reduce the dimension of handwritten digital sample set MNIST, and different neural networks are constructed according to different thresholds for classification. Furthermore, the change of classification accuracy under different thresholds is analyzed. The experimental results show that when the threshold of PCA is between 79%—81% (dimension is 41—50), the classification accuracy is the highest, and the accuracy decreases accordingly when the threshold is lower or higher than that region. It is proved that there is no positive correlation between application results and threshold selection of PCA, and the average of the eigenvalues is not a mandatory criterion.

Key words: principal component analysis(PCA); threshold selection; neural network; handwritten digital sample classification

收稿日期:2020-04-28;**修订日期**:2021-01-30

引　　言

主成分分析(Principal component analysis, PCA)是特征提取和降维的一种方法,其目的是将一系列具有相关关系的多个指标或影响因子转化为一组新的相互独立的综合指标,在转化的同时尽可能多地保留原始变量的信息,以实现对多变量的降维,从而降低问题的复杂度[1]。关于主成分阈值选择的标准,文献[2]明确给出了选择依据,即:“保留特征值大于平均特征值的主成分,对于一个相关矩阵,对应的平均特征值为1”。文献[3]对核心主成分代表的含义进行了解释和分析。具体地,文献[1]选择95%作为主成分阈值;文献[4]选择90%作为阈值;文献[5]以99%作为阈值;文献[6]同样选择99%作为阈值;文献[7]中没有说明具体阈值选择多少,而是将维度从784维降到441维;文献[8]分别选择95%、97%和99%作为阈值,进行了数字图像压缩;而文献[9]选择76%的主成分,将原有11维数据降至4维;文献[10]中对于分布鲁棒的有条件风险值和风险规避的生产-运输问题,提出的PCA仅使用50%的主成分,得到近似解(在1%以内),计算时间减少了1~2个数量级。

综合以上分析,很多应用以平均特征值作为选择主成分的标准,但是阈值选择低于该标准应用结果是否会降低,并没有明确的分析。因此本文提出一种PCA阈值选择差异性的实验分析方法,并选择PCA-BP神经网络对手写体数字进行分类,研究主成分阈值与分类结果之间的变化关系。为了使结果具有可比较性,选择来自美国国家标准与技术研究所的手写体数据集MNIST作为样本进行实验。结果表明本文得出的阈值选择差异性分析结果可以为不同应用中主成分分析阈值的选择提供依据。

1　基本算法

1.1　主成分分析

(1)主成分分析原理

主成分分析就是将原来的p个变量,通过线性组合方式生成p个包含原有变量信息的综合变量,组合方式表示为

$$\begin{cases} F_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p \\ F_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p \\ \vdots \qquad\qquad \vdots \\ F_p = a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pp}x_p \end{cases} \tag{1}$$

将上述线性组合方式用向量方式重写为$F_j = \alpha_{j1}x_1 + \alpha_{j2}x_2 + \cdots + \alpha_{jp}x_p$, $j = 1, 2, \cdots, p$。同时应该具有以下3个条件:

① 两者互不相关,即$\mathrm{Cov}(F_i, F_j) = 0$。

② F_1的$\mathrm{Var}(F_1)$最大,F_2的$\mathrm{Var}(F_2)$次之,以此类推,F_p的$\mathrm{Var}(F_p)$最小。

③ $a_{k1}^2 + a_{k2}^2 + \cdots + a_{kp}^2 = 1, \quad k = 1, 2, \cdots, p$。

因此,本文将$\mathrm{Var}(F)$最大的称为第1主成分,即F_1;将$\mathrm{Var}(F)$次之的称为第2主成分,即F_2;以此类推,$\mathrm{Var}(F_1)$最小的为第p主成分,即F_p。其中a_{ij}称为主成分系数。

(2)主成分选择

在主成分分析中可以获得与原变量维度相同的p综合变量,且对应的方差依次递减,相应的所含原变量数据的信息也递减。在实际应用中,如果将这p个主成分全部使用,则达不到降维的效果。而且根据方差,前面几个主要的成分基本上包含了绝大多数原始数据信息,因此p个主成分不用全选。主成分的选择方法是以每个主成分的累计贡献率作为依据,根据其大小选择前k个。其中贡献率指该主成分的方差占所有主成分方差和的比重,用贡献率β_i来表示,表达式为

$$\beta_i=\frac{\lambda_i}{\sum_{i=1}^{p}\lambda_i}\times 100\% \tag{2}$$

贡献率β_i的大小反应了该主成分包含原变量信息的多少，其中选择多少个主成分一般以平均特征值作为依据。

(3)主成分分析

对MNIST数据集进行主成分分析，根据单个主成分贡献率以及累计主成分贡献率，绘制主成分贡献率直方图，如图1所示。

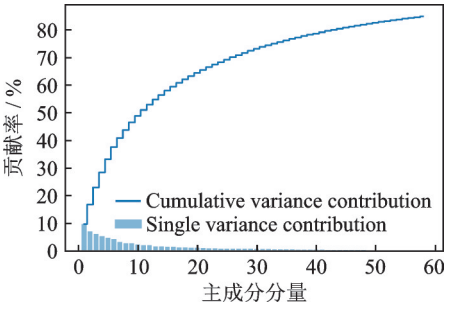

图1 主成分贡献率

Fig.1 Contribution rate of principal components

为了直观感受PCA分析降维以后与原图的区别，这里选择将维度下降到86维、96维、99维的数据，进行反向还原处理，然后对图像进行重构，由于样本数量较大，选择训练数据的前100个样本进行还原对比，如图2所示。

根据分析，保留主成分的个数根据$\bar{\lambda}=\sum_{i=1}^{p}\lambda_i/p$进行选择，MNIST主成分平均特征值为0.914 56，大于该平均值的特征值个数为179个，因为保留原始数据有效成分的维度为179维(原维度为784)，其对应的累计方差贡献率为95.993%。但是根据图2可知，99维基本可以还原原数据，其对应的累计方差贡献率为91.36%，因此还有80个维度的压缩空间。

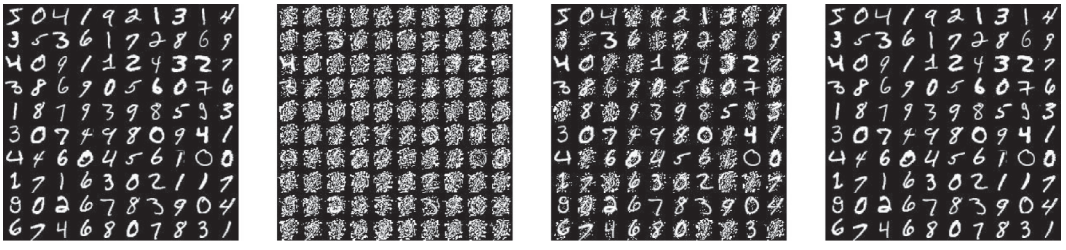

图2 降维后数据重构对比

Fig.2 Comparison of data reconstruction after dimensionality reduction

1.2 PCA-BP神经网络分类模型

(1)BP神经网络

BP神经网络有输入层、隐含层和输出层3层构成，典型的网络结构如图3所示[11]。网络激活函数选择Sigmoid函数，通过反向传播输出层误差，调整权重值与偏置使得代价函数$C=\frac{1}{2n}\sum_x\|y(x)-a^L(x)\|^2$最小。

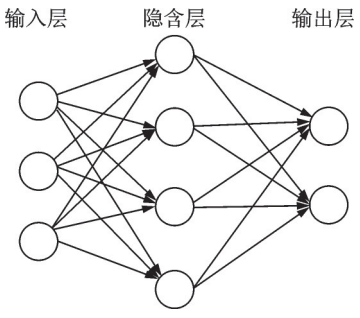

图3 神经网络结构

Fig.3 Neural network structure

在实际使用过程中，将反向传播算法与随机梯度下降等算法进行结合使用，从而能计算许多训练样本所对应的梯度。例如给定一个大小为m的小批量数据，下面对小批量数据应用梯度下降学习算法进行描述：

① 输入训练样本的集合。

② 对每个训练样本x:设置对应的输入激活$\boldsymbol{a}^{x,1}$,并执行下面的步骤:

前向传播:对每一层$l=2,3,\cdots,L$,计算$\boldsymbol{z}^{x,l}=\boldsymbol{w}^l\boldsymbol{a}^{x,l-1}+\boldsymbol{b}^l$和$\boldsymbol{a}^{x,l}=\boldsymbol{\sigma}(\boldsymbol{z}^{x,l})$。

输出误差$\boldsymbol{\delta}^{x,L}$:计算误差向量$\boldsymbol{\delta}^{x,L}=\nabla_a C_x\odot\boldsymbol{\sigma}'(\boldsymbol{z}^{x,L})$。

反向传播误差:对每一层$l=L-1,L-2,\cdots,2$(除了输出层和输入层),计算$\boldsymbol{\delta}^{x,l}=((\boldsymbol{w}^{l+1})^{\mathrm{T}}\cdot\boldsymbol{\delta}^{x,l+1})\odot\boldsymbol{\sigma}'(\boldsymbol{z}^{x,l})$。

③ 梯度下降:对每一个$l=L-1,L-2,\cdots,2$根据$\boldsymbol{w}^{l\prime}=\boldsymbol{w}^l-\frac{\eta}{m}\cdot\sum_x\boldsymbol{\delta}^{x,l}(a^{x,l-1})^{\mathrm{T}}$,$\boldsymbol{b}^{l\prime}=\boldsymbol{b}^l-\frac{\eta}{m}\cdot\sum_x\delta^{x,l}$更新权重和偏置。

(2)分类模型

PCA-BP神经网络分类模型如图4所示。基于PCA神经网络的分类识别步骤为:

① 获取数据且进行预处理并归一化。

② 进行样本的主成分分析,对预处理后的样本集进行主成分分析,选择特征值大于平均特征值的m个主成分作为新的训练和测试样本。

③ 根据不同主成分,创建不同结构的神经网络,进行训练,最后进行测试。

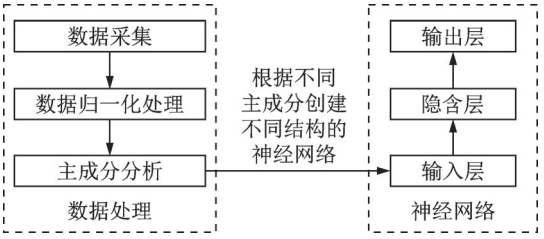

图4　PCA-BP神经网络分类模型

Fig.4　PCA-BP neural network classification model

④ 将第3步的测试结果与没有主成分分析的测试结果进行比较,观测效果是否降低,训练效率是否提升。

2　仿真结果与分析

应用PCA分析,将MNIST数据的784维降至10、20、30、40、41、42、43、44、45、46、47、48、49、50、58、86、153和330维,然后分别选择隐含层神经元个数为30和100,与原784维分类结果进行比较。由于维数过多,因此选择部分对比结果,结果如图5所示,其中第1行为隐含层神经元个数为30的对比结果,第2行为隐含层神经元个数为100的对比结果。

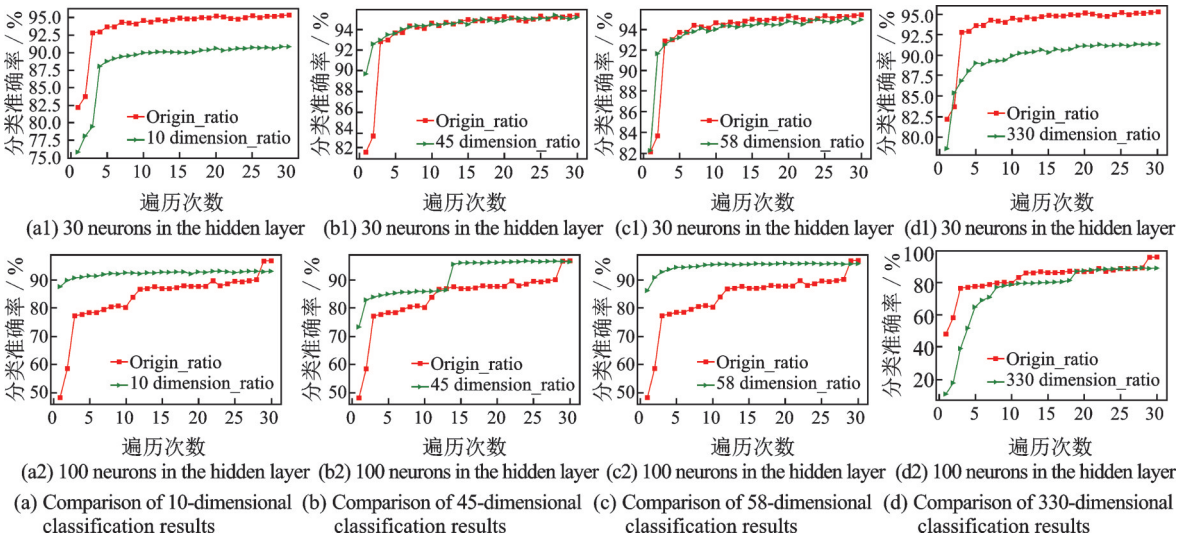

图5　降维后分类结果对比

Fig.5　Comparison of classification results after dimensionality reduction

通过图5的分析可知，增加隐含层神经元个数，对于分类结果有明显的提升。而维度从10～330分类结果明显是从低到高再到低的一种变化规律。主成分分析在降维的同时，提高了网络学习的效率，不同维度学习时间对比如图6所示。从图6分析可知，降维大大提高了网络的学习效率，所用时间与维度的关系近似为线性。

为了更好地说明分类准确率、主成分阈值（累计方差贡献率）和维度之间的关系，分别对隐含层神经元个数为30和100时，每个维度的最高分类准确率和主成分阈值进行仿真，结果如图7所示。

通过仿真结果可以看出方差累计贡献率与维度之间是正相关的。随着维度的增加，分类准确率也随之增加，当维度到50维以后，分类准确率明显减小。维度最密集区也是分类准确率最高区域，对应维度是41～50维之间，对应方差累计贡献率在79%～81%之间。58、153和330维对应方差累计贡献率分别为85%、90%、99%。因此，依据平均特征值作为阈值，保证了原有数据信息最大程度不丢失，但不能作为应用的硬性标准。

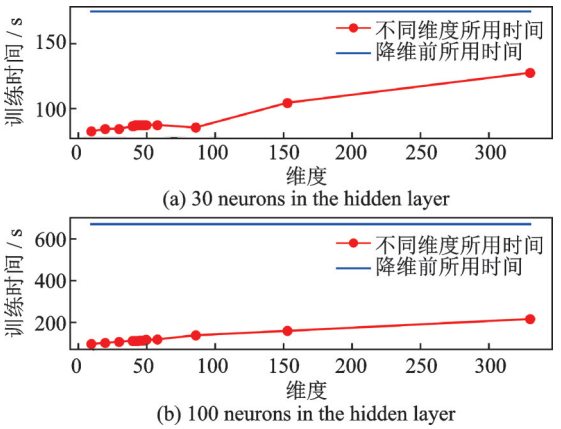

图6 不同维度学习时间

Fig.6 Learning time in different dimensions

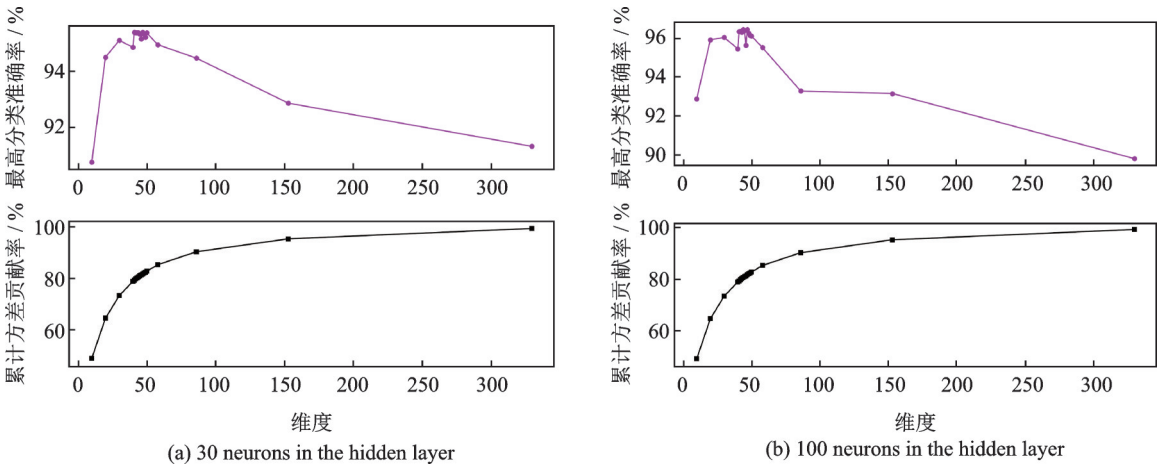

图7 不同维度分类准确率与主成分阈值对比

Fig.7 Comparison of classification accuracy and principal component thresholds in different dimensions

3 结束语

本文提出一种主成分阈值选择差异性的实验分析方法，通过研究得出降维大大提高了网络的学习效率，所用时间与维度的关系近似为线性。维度与累计方差贡献率正相关。随着维度的增加，分类准确率也随之增加，当但维度增加到50维以后，分类准确率明显减小。维度最密集区也是分类准确率最高区域，对应维度在41～50维之间，对应方差累计贡献率为79%～81%。因此，平均特征值作为阈值并不能作为其他应用的一个阈值，且主成分的提高并不会使分类准确率提高。但是根据降维后数据重构对比图（图2）发现，维度在小于86的范围内，信息识别难度加大，分析其原因是由于其数据主要结构特征在79%～81%范围形成，随着结构特征增加反而影响应用结果。

参考文献:

[1] 张义宏.基于PCA的BP神经网络优化的研究与应用[D].沈阳:东北大学,2014.
ZHANG Yihong. Research and application of BP neural network optimization based on the PCA[D]. Shenyang: Northeastern University, 2014.

[2] RENCHER A C, CHRISTENSEN W F. Methods of multivariate analysis[M]. 3rd ed. [S.l.]: Wiley, 2012.

[3] YANINA G, GUIDO G. Searching for the core variables in principal components analysis[J]. Stats, 2015, 32(4): 730-754.

[4] 赵秀红.基于主成分分析的特征提取的研究[D].西安: 西安电子科技大学,2016.
ZHAO Xiuhong. Research on feature extraction based on principal component analysis[D]. Xi'an: Xidian University, 2016.

[5] 陈善学,张燕琪.基于自适应波段聚类主成分分析和反向传播神经网络的高光谱图像压缩[J].电子与信息学报,2018,40(10): 2478-2483.
CHEN Shanxue, ZHANG Yanqi. Hyperspectral image compression based on adaptive band clustering principal component analysis and back propagation neural network[J]. Journal of Electronics & Information Technology, 2018, 40(10): 2478-2483.

[6] 许永强,刘万康.基于主成分-BP神经网络的我国农村居民用电量的预测研究[J].电力学报,2016,31(2): 162-166, 170.
XU Yongqiang, LIU Wankang. The prediction research of Chinese rural residents consumption based on the principal component-BP neural network[J]. Journal of Electric Power, 2016, 31(2): 162-166, 170.

[7] 杨浩.深度学习与主成分分析融合的研究与应用[D].成都:成都理工大学,2016.
YANG Hao. Research and application of the combination of deep learning and principal component analysis[D]. Chengdu: Chengdu University of Technology, 2016.

[8] NG S C. Principal component analysis to reduce dimension on digital image[J]. Procedia Computer Science, 2017, 111: 113-119.

[9] FRANCIS P J, WILLS B J. Introduction to principal components analysis[J]. Journal of Injury Function & Rehabilitation, 2014, 6(3): 275-278.

[10] CHENG J, CHEN R L Y, NAJM H N, et al. Distributionally robust optimization with principal component analysis[J]. SIAM Journal on Optimization, 2018, 28(2): 1817-1841.

[11] 杨丽丽.基于人工神经网络的手写数字模式识别和分类[D].太原: 中北大学,2012.
YANG Lili. Handwritten numeral pattern recognition and classification based on artificial neural network[D]. Taiyuan: North Central University, 2012.

作者简介:

张婧(1989-),**通信作者**,女,讲师,研究方向:现代数字通信理论、数据处理,E-mail:526510423@qq.com。

刘倩(1990-),女,工程师,研究方向:现代数字信号处理,E-mail:450299406@qq.com。

(编辑:张黄群)