文章编号：2096-1472(2016)-06-01-03

# 主成分分析方法综述

赵 蔷
（咸阳师范学院计算机学院，陕西 咸阳 712000）

**摘 要**：主成分分析是一种非常有效的数据分析处理的技术，具有非常广泛的应用前景。本文首先概述了主成分分析方法，然后介绍了PCA的定义、模型、算法及选取主成分个数的标准，对PCA技术的优势和缺陷分别进行了剖析和总结，对PCA在评价排序、特征提取、模式识别、图像处理、图像分类和图像压缩等领域的实际应用进行了讨论，对主成分分析方法的发展趋势和应用前景做了展望。

**关键词**：主成分分析；PCA模型；特征提取；图像处理
**中图分类号**：TP391 **文献标识码**：A

# A Review of Principal Component Analysis

ZHAO Qiang
(*School of Computing Science,Xianyang Normal University,Xianyang* 712000,*China*)

**Abstract:**PCA(Principal Component Analysis)is an effective data analysis technique with a bright future of extensive application.The paper summarizes PCA in the first place,and then introduces its definition,data model,algorithm and the standards to determine the number of selected principal components.Moreover,the paper analyzes and summarizes the advantages and disadvantages of the PCA technique,and discusses its practical application in different fields,like evaluation and sorting,feature extraction,patter recognition,image processing,image classification and image compression.Finally,the paper makes expectation about the development trend and application prospect.

**Keywords:**PCA;PCA model;feature extraction;image processing

## 1 引言(Introduction)

PCA(Principal Component Analysis)，主成分分析，是一种数据分析的技术，主要思想是将高维数据投影到较低维空间，提取多元事物的主要因素，揭示其本质特征。主成分分析的应用范围非常广泛，经常和分类、聚类，以及与其他方法连用进行数据处理。它可以高效地找出数据中的主要部分，将原有的复杂数据降维，去除整个数据中的噪音和冗余。

PCA是一种统计分析方法，它将原来众多具有一定相关性的多个指标，重新组合成一组新的互相无关的综合指标[1]。它是一种最小均方意义上的最优变换，目的是去除输入随机向量之间的相关性，突出原始数据中的隐含特性[2]。PCA方法的优势在于数据压缩以及对多维数据进行降维，它操作简单，且没有参数限制，可以方便的应用于各个场合。它经常被用于人脸识别和图像压缩、特征提取等领域，是在高维数据中寻找模式的一种技术[3]。

## 2 主成分分析方法(Principal component analysis )

### 2.1 主成分分析方法的目标

由于原始数据的变量基数比较复杂，难以描述其特征，主成分分析提出了一种简单解决问题的思想，从事物的主要方面进行重点分析。该方法认为某个事物的特征集中在几个主要变量上，只需要将这几个变量分离出来，对这几个变量进行重点分析，用它们的线性组合表示事物的主要特征。因此，主成分分析方法的目标就是寻找$x(x<n)$个新变量，用这$x$个变量反映事物的主要特征，实现对原始数据矩阵规模的压缩。这$x$个新变量就是“主成分”，用它们反映原来$n$个变量的特征，并且这$x$个变量之间是互不相关的。通过主成分分析实现数据维数的压缩，将多元数据特征在低维空间中反映出来[3]。

### 2.2 PCA模型

PCA是一种正交变换，利用二阶的统计信息进行计算。它强调数据之间的相似和不同，是一种在高维数据中寻找模式的技术[2]。对于原始数据，我们可以通过一些变换来提取数据间的内在特征，其中一种方法就是通过线性变换去实现[4]。这个过程可以表示为：

$$Y = wX$$

这里$w$是一个变换值，可以把它当作基本的变换矩阵，通过此变换来提取原始数据的特征。令$x$为表示环境的$m$维随机向量。假设$x$均值为零，即：

$$E[x] = 0$$

令$w$表示为$m$维单位向量，$x$在其上投影。这个投影被定义为向量$x$和$w$的内积，表示为：

$$Y = \sum_{k=1}^{n} w_k \cdot x_k = w^T \cdot x$$

在上式中，需满足以下约束条件：

基金项目：陕西省教育厅专项科研计划项目(15JK1803)，咸阳师范学院专项科研基金资助项目(13XSYK055).

$$\|w\| = \left(w^T w\right)^{1/2} = 1$$

主成分分析方法就是寻找一个权值向量$E[y^2]$，它能够使表达式 取最大值[4]。

### 2.3 特征值求解

PCA特征根求解的步骤如下：

(1)将原始数据表示为$m*n$的矩阵。$n$为原始数据的个数，$m$为变量个数。

(2)计算原始数据的均值。

(3)用原始数据减去均值，得到矩阵$X$。

(4)对$XX^T$进行特征根分解，求特征向量及其对应的特征值。

(5)选取最大的若干个特征值对应的特征向量，即为求得的主成分。

PCA方法用线形代数可以描述为：寻找一组正交基组成的矩阵$P$，定义$Y=PX$，使得$C_Y=MYY^T$是对角阵。$P$的行向量，就是数据$X$的主成分，也就是$XX^T$的特征向量，矩阵$C^Y$对角线上第$i$个元素是数据$X$在方向$P_i$的方差[4]。

### 2.4 主成分数量的选取

主成分是$n$个原始变量的线性组合，各主成分之间互不相关。每个主成分对应一个方差，该方差为协方差阵对应的特征值，各主成分特征值之和为1。将主成分按照其对应的方差值从大到小依次排列，则最大的方差对应第一主成分，以此类推。

选择主成分的数量取决于保留部分的累积方差在总方差中所占的百分比。由于所有主成分的总方差值是确定的，前面变量的方差较大，则后面的变量方差就较小。只有前几个综合变量才称得上是主成分，后几个综合变量为次成分。一般情况下，可根据问题的实际需要，主观地确定一个百分比值，当前$x$项的方差之和大于此百分比值时，就可以决定保留前$x$个主成分，而忽略后面的次成分[5]。

## 3 主成分分析的特点(Characteristic of PCA)

综上所述，主成分分析方法有很多优点，可将其归纳如下：

(1)在数据处理时，舍弃了一部分主成分，只取前几个方差较大的几个主成分来表示原始数据，可减少计算量。

(2)主成分之间是互不相关的，消除了原始数据之间的相关影响。在选取评价指标时，消除了指标之间的相关影响，因此更容易选择指标。而且实践证明指标之间相关程度越高，主成分分析效果越好。

(3)在综合评价函数中，主成分的权数为各个主成分的贡献率，反映了该主成分包含原始数据的信息量占全部信息量的比重，这样确定地权数比较客观、合理，克服了某些评价方法中人为确定权数的缺陷。

(4)主成分分析的计算方法比较规范，便于在计算机上实现。

主成分分析方法的不足主要体现在两个方面：

(1)所得到的主成分实际含义模糊，没有原始数据的含义确切、清楚。

(2)主成分分析方法只考虑了数据的二阶统计量(自相关)，这对于高斯分布是足够的，但对于非高斯分布，由于高级统计量中含有附加的信息，因此PCA对其表示不够充分。

PCA方法算法比较简单，且具有一定的局限性。因此，越来越多的研究都集中在PCA和其它方法如LDA、K−means方法、核函数、SVM、粗糙集、专家模型、GMM等方法相结合的应用，并取得了很好的效果。

## 4 主成分分析的应用(Application of PCA)

主成分分析主要应用是评价排序、特征提取、图像处理、图像分类、模式识别、图像压缩等方面。下面将综述PCA的主要应用范例。

(1)评价排序

现实生活中人们经常要对事物进行评价和排序，但事物本身往往是由多元数据构成，且数据之间具有某些内在的联系。使用PCA进行数据处理，可以去除数据之间的相关性，又减少了工作量。在文献[6]中，作者介绍了一种基于PCA的教学质量评价方法，该方法消除了文中所确定的16个教学质量评价指标之间的相关性，将原来的16个指标简化为5个主成分，对这5个主成分的载荷进行分析，进而评价课堂教学质量。实验结果证明，主成分分析所得的结果与实际资料反映的情况相符，详细内容见参考文献[6]。

现在，越来越多的基于PCA的评价方法正不断地应用于各个领域，比如文化符号归因分析、软件质量评价、性能评估、可持续发展评价、城市交通拥挤评价、风险评价等。这些方法中，PCA都被用于降低维数并去除数据之间的相关性。评价指标的选取、层次权重如何分配是此类问题的研究重点。随着大数据时代到来，PCA在该领域将会有更加广泛的应用。

(2)特征提取

在特征提取领域应用最为广泛、提取特征效果较好的就是PCA方法。该方法提取了事物的主要特征元素，同时达到了降维的目的，简化了复杂模型。在文献[7]中，作者提出了一种基于多重组合特征提取算法(PCA−CFEA)的文本分类方法，首先用正交变换将文本空间降维，再通过多重组合特征提取算法在降维后的特征空间快速提取代表性强的特征项，过滤掉那些代表性弱的特征项，随后使用SVM分类器对文本进行分类，详细内容见参考文献[7]。

使用PCA进行特征提取已经成为该领域的热点问题之一。目前，很多基于PCA的特征提方法仅提取了事物的某些特征，根据提取的特征对事物进行检索和分类，则是该领域的应用要深入研究的内容。

(3)模式识别

模式识别是PCA的一个重要应用。由于高维数据对模式识别是不利的，解决这一问题的方法就是降维。以人脸识别为例，数据源是M幅不同的人脸图像，可使用PCA方法提取出人脸的内部结构特征，即所谓“模式”[8]。当有新的图像需要识别，只需要在主成分空间对该图像进行分析，就可得到新图像与原人脸图像集的相似度差异，从而实现人脸识别。

在文献[8]中，作者提出一种方法，将PCA和LDA方法相结合进行性别鉴别。

随着“互联网+”的到来和可穿戴电子设备的普及，越来越多的人习惯于快捷交易、快捷支付，而这都有赖于身份认证，比如人脸识别、指纹识别。在此应用领域，PCA因为其降维的特点，将会有更加广泛的应用。但由于同一人脸会因为光照、表情和姿态不同而有较大的差异，因此，如何克服光照的影响以及由于表情和姿态不同造成的差异，是该类方法需要深入研究的问题。

(4)图像处理

PCA在图像边缘检测、图像融合等方面被广泛应用。在文献[9]中，作者提出一种基于多专家的PCA边缘检测模型，该方法将一个边缘检测法视为一个专家，首先采用Sobel算子、Canny算子等五种算子建立统计模型，然后利用PCA方法对五个专家的检测结果进行分析，最后利用提出的多个专家的检测模型融合多个专家的检测信息，得到综合的检测结果，实验结果证明，该方法可以获得很好的边缘检测结果，详细内容见参考文献[9]。在文献[10]中，作者提出一种基于PCA和总方差模型的图像融合框架，首先用PCA对源图像处理，根据提取的前几个主成分重建图像，再经下采样过程得到近似图像，然后通过上采样得到细节图像，最后将近似图像和各个细节图像累加，完成图像重构，将该框架纳入总方差模型后形成一种新的框架。实验结果证明，该方法不仅可以获得较好的融合效果，还可去噪，而且能够保持全色图像和多光谱图像的光谱信息和空间信息，详细内容见参考文献[10]。

PCA是图像处理领域中很有研究价值的成果之一。由于PCA具有降维的特性，而图像本身又是多维数据，因此PCA在图像处理领域有着非常广泛的应用。鉴于PCA本身的局限性，在这些应用中，需要将PCA和其他的方法相互结合。

(5)图像分类

图像的颜色使图像的主要特征之一。我们曾采用PCA和LDA进行了对卫星遥感图像进行分类，首先获得图像的PCA颜色特征子空间，计算图像的LDA颜色特征子空间，将PCA算法和LDA算法的特征空间相融合，将原始卫星图像投影到PCA-LDA算法的融合颜色特征空间中，进行图像分类。该方法去除了图像的R、G、B间的相关性，去掉了原始图像中大量冗余信息，改善了光照敏感性，在该方法中了采用基于区域分类的空间一致性原则来合并空间信息。实验表明，该方法能高效的描述卫星图像的颜色特征，分类准确度高，详细内容见参考文献[11]。在文献[12]中，作者提出一种基于PCA和GMM的图像分类算法。

图像分类会使用分类器，通常使用的分类器有Euclid Distance法、Maximum Likelihood法和K均值聚类算法(K-means)，前两个属于监督分类法，后者属于非监督分类法，是一种动态聚类方法算法。如何选取合适的分类器，并在分类算法中结合图像的其他信息比如的纹理信息来提高分类效果，对图像分类方法深入研究近一步提出基于图像特征的图像检索方法，是基于PCA的图像分类算法需要深入探讨的问题。

(6)图像压缩

PCA的另一个广泛应用是图像压缩。假设有20幅图像，使用PCA方法处理该图像集，将得到20个特征向量，提取其中15个主成分。使用这15个特征向量进行图像复原变换，就得到一个只有15维的数据。数据维数从原来的20降到了15，图像压缩了四分之一。该方法是一种有损压缩，但保持了原始图像中最“重要”的信息，是一种非常重要且有效的方法[5]。

## 6 结论(Conclusion)

主成分分析的应用范围非常广泛，经常和分类、聚类算法及与其他方法连用进行数据处理。其最大优势就是对原有数据进行简化，去除了噪音和冗余，对数据进行降维处理，揭示隐藏在复杂数据背后的简单结构。当然PCA方法也存在一定的不足，为了得到更好的效果，需要对PCA进一步深入研究，结合其他算法对PCA进行改进。相信随着PCA为越来越多的人所认知，该方法会得到更加广泛的应用。

## 参考文献(References)

[1] M.P.Dubuisson-Jolly & A.Gupta.Color and Texture Fusion: Application to Image Segmentation and GIS Updating[J].Image and Vision Computing,2012(18):823-831.

[2] Kwitt.R.Meerwald & Uhl.A.Lightweight detection of additive watermarking in the DWT domain[J].IEEE Transactions on Image Processing,2013(2):474-484.

[3] Luo.M & Bors.A.G.Principal Component Analysis of spectral coefficients for mesh watermarking[J].Signal Processing,2011(9):625-634.

[4] 鲁晨.主元分析(PCA)理论分析及应用[J].图像处理学报,2007(8):38-41.

[5] 高琪.主成分分析方法在图像处理中的应用[J].计算机学报,2011(3):98-100.

[6] 李宏明.基于多元统计分析的地方高校课堂教学质量评价[J].台州学院学报,2010,32(3):77-80.

[7] 李建林.一种基于PCA的组合特性提取文本分类方法[J].计算机应用研究,2011(8):2399-2401.

[8] 何国辉,甘俊英.PCA-LDA算法在性别鉴别中的作用[J].计算机工程,2006,32(19):208-210.

[9] 李建军,韦志辉,张正军.多专家的PCA边缘检测模型[J].哈尔滨工业大学学报,2012,44(12):92-95.

[10] 潘瑜,等.基于PCA和总方差模型的图像融合框架[J].计算机辅助设计与图形学学报,2011,23(7):1200-1210.

[11] 赵蕾,等.一种基于PCA-LDA的卫星遥感图像的分类方法[J].计算机应用与软件,2013,30(2):198-204.

[12] 肖政宏,等.基于PCA和GMM的图像分类算法[J].计算机工程与设计,2009(5):93-107.

## 作者简介：

赵 蕾(1971-)，女，硕士，副教授.研究领域：软件理论，图形图像处理.