高维数据的维数约简方法及其应用

王建中　张宝学　著

科学出版社

高维数据的维数约简方法及其应用

王建中　张宝学　著

科学出版社
北京

内 容 简 介

高维数据的维数约简技术是当今计算机科学、机器学习等领域的热门研究问题之一，具有广泛的发展前景。本书在对已有维数约简方法进行分析和总结的基础上，从特征提取和特征选择两个方面提出五种新的维数约简方法，并以人脸图像和微阵列数据分析等问题为例，通过与目前较流行的维数约简方法对比，验证了所提出方法的性能。

书中包含很多实例、不同方法的介绍与对比及详细的图表与实验结果分析，不仅适用于计算机科学与技术专业的本科生和研究生阅读，也可为从事机器学习、模式识别等相关领域研究的工程技术人员提供参考。

图书在版编目(CIP)数据

高维数据的维数约简方法及其应用/王建中，张宝学著. —北京：科学出版社，2016. 10

ISBN 978-7-03-050063-2

Ⅰ. ①高… Ⅱ. ①王… ②张… Ⅲ. ①数据处理 Ⅳ. ①TP311. 13

中国版本图书馆 CIP 数据核字(2016)第 233906 号

责任编辑：张艳芬 纪四稳/ 责任校对：郭瑞芝

责任印制：张 倩 / 封面设计：蓝 正

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

新科印刷有限公司印刷

科学出版社发行 各地新华书店经销

*

2016 年 10 月第 一 版 开本：720×1000 1/16

2016 年 10 月第一次印刷 印张：8 1/4 插页：2

字数：156 000

定价：88.00 元

(如有印装质量问题，我社负责调换)

前　言

近年来，随着科学技术的发展，人们对于各种数据的获取较之以往更为方便和普遍。然而，在很多实际应用问题中，人们所采集到的数据往往具有高维数、非线性等特征。这些特征一方面导致了“维数灾难”现象的出现，另一方面不利于人们直接理解及发现数据集所蕴含的内在规律。因此，利用维数约简技术对高维数据进行处理就显得尤为重要。本书对维数约简相关研究背景进行概述，对目前高维数据维数约简技术的研究现状进行总结，并对目前应用比较广泛的维数约简方法基本原理进行介绍和分析。以人脸数据、微阵列基因表达数据等为例，从特征提取和特征选择两个方面对维数约简技术进行研究，主要研究内容如下。

(1) 为了解决传统主成分分析算法无法应用于非线性结构数据的缺点，提出一种基于局部主成分分析和低维坐标排列的流形学习算法。在本方法中，首先利用测地线距离约束和最小集覆盖方法将数据所在的整体非线性流形划分成若干个互相重叠的最大线性贴片。由于得到的每个最大线性贴片所包含的数据具有线性结构，因此可以利用传统的主成分分析方法对每个最大线性贴片中的数据进行维数约简，得到其局部低维坐标。最后，将坐标排列技术和最大间隔准则结合，对所有最大线性贴片的局部低维坐标进行排列，得到整体数据集的全局维数约简结果。由于本方法在维数约简的过程中同时考虑了数据的流形结构和类别信息，因此在人脸识别的实验中取得了较好的结果。

(2) 提出一种自适应加权的子模式局部保持投影算法。与传统的局部保持投影算法不同，该算法首先将输入的高维原始数据划分成若干个子模式，然后利用局部保持投影算法对得到的每个子模式集合分别进行维数约简，得到可以保持各个子模式集合局部结构的低维特征。此外，为了增强算法的鲁棒性，采用一种自适应的方法计算每个子模式对于识别的权重。通过将该算法应用于人脸识别问题可以看出，其不仅能够提高传统局部保持投影算法的计算效率，在识别的准确率方面也要优于其他子模式算法。

(3) 提出一种结构保持投影算法。在该方法中，同样将原始高维数据划分成若干个子模式。但与前面提到的自适应加权的子模式局部保持投影算法和其他基于子模式的方法不同，结构保持投影算法在对每个子模式进行维数约简的过程中，不仅考虑了其所在子模式集合的流形结构，还考虑了其与来自同一样本的其他子模式之间的关系。通过结构保持投影算法，可以在保持各个子模式集合的非线性流形结构的同时，保持每个输入样本的内在结构。与前面提到的两种基于流

形学习的维数约简算法相同，将结构保持投影算法应用于人脸识别问题并在标准人脸数据库中验证算法的有效性。由实验结果可以看出，结构保持投影算法要优于其他全局和局部识别方法。

(4) 提出一种基于改进有效范围的特征选择方法，并将其应用于微阵列基因表达数据中的特征选择。在本方法中引入了一个新的概念，即包含面积，以此来描述每个特征对每个类别有效范围之间存在的包含关系。除此之外，本方法还考虑了每个特征对每个类别在重叠面积和包含面积内的样本比例。基于改进有效范围的特征选择方法通过综合考虑重叠面积、包含面积和样本比例三方面的因素来评价基因的重要程度，从而更加有利于选出最优的特征子集。

(5) 提出一个新的过滤式特征选择框架，并将其应用于微阵列基因表达数据的特征选择。该框架基于最大权重最小冗余准则选择最优特征子集。与基于排序的过滤式特征选择算法相比，该算法考虑特征重要性的同时考虑了特征间的冗余度；与其他基于空间搜索的过滤式特征选择方法相比，该算法不依赖于特征重要性与冗余度，这使其应用范围更加广泛。

高维数据的维数约简已经成为目前机器学习领域的一个重要研究方向，希望本书所涉及的内容可以为相关学者提供一定的参考与启发，为进一步促进维数约简技术的发展贡献一点微薄之力。

在撰写本书过程中，得到了东北师范大学计算机科学与信息技术学院领导的指导，特别感谢孔俊教授的帮助和支持。另外，还要感谢郭羽婷、武丽珊、周爽等对本书的贡献。

本书得到了国家自然科学基金项目(No. 11271064，61403078，61672150)及吉林省重点科技攻关项目(No. 20160204047GX)的资助。

限于作者水平，书中难免存在疏漏或不足之处，敬请广大读者批评指正并提出宝贵意见。

目　录

第1章 绪 论

1.1 研究背景及意义

随着科学技术的发展和进步，人们可以越来越方便地获取大规模的数据，特别是在科学研究领域，为了对客观现象或事物进行丰富、细致的描述，科学家往往需要大量的数据信息。这些海量数据在带给人们便利的同时，也产生了新的问题：在很多实际应用中，如生物信息学、计算机图像处理、信息检索、文本分析、生物信息认证等，人们所采集到的大量数据通常是繁杂的、高维的及非结构化的。这就导致了隐藏在数据背后的内在关系和规律难以被直观地发现，在一些文献中，把这种无法有效地挖掘隐含在海量数据中的知识的现象称为“数据爆炸但知识匮乏”现象[1]。

如何从数据中发现有价值的信息和内在规律，并根据现有数据对事物未来的发展趋势进行预测，是统计机器学习及其相关领域所关注的主要目标之一。但当面对高维数据时，传统的数据分析方法往往遇到以下问题。首先，高维数据会导致“维数灾难”(curse of dimensionality)[2]现象的出现。所谓“维数灾难”，就是指在缺乏简化数据的前提下，要在给定的精度下准确地对某些变量的函数进行估计，人们所需要的样本数量会随着样本维数的增加而呈指数形式增长。其次，与“维数灾难”相对应的一个几何现象称为“空空间现象”(empty space phenomenon)[3]，该现象指出，因为人们获得的样本数量往往是有限的，所以高维空间本质上是稀疏的。例如，随着数据维数的显著增加，高斯分布中的三法则将不再适用，即在高维空间中，大多数数据不再处于高斯函数的中间，而是集中在函数的边界。此外，有研究者已经证明，在高维空间中，数据更多地分布在超球的表面而不是球的中心[4]。在这种现象下，相对低密度的区域包含了大部分的数据，而高密度区域反而数据缺乏，这就造成了很难利用一般的多元数据分析方法对高维空间的密度和几何性质进行直接分析。最后，近期的研究又发现了高维数据处理中的另一个问题，即高维空间中测度的“集中现象”(concentration phenomenon)[5,6]。该现象表明，样本点之间距离测度的可区分性会随着维数的增加而减弱。这导致一些基于距离测度的分析算法(如最近邻分类算法等)难以在高维空间取得较好的效果。

由于存在以上种种问题，对高维数据进行维数约简处理就成为数据分析中非常重要的一个步骤和组成部分。维数约简主要是指将样本从高维观测空间通过

线性或非线性方法投影到一个低维特征空间，从而发现隐藏在高维数据中的有意义的低维结构。利用维数约简技术对高维数据进行处理主要有以下好处[7]：①可以对数据进行有效压缩，以节省存储空间；②在一定程度上消除数据中存在的噪声；③便于从数据中提取特征，以完成后续的分类或识别任务；④通过将数据投影到低维空间（二维或三维），可以实现数据集的可视化。

1.2 研究现状

目前，维数约简技术主要有两种：特征提取和特征选择。其中，特征提取指通过某种变换将数据变换到另外一个更能揭示数据本质特征的空间，并在此空间中选择一部分具有代表性的特征表示数据[8]；特征选择指在原数据中选择一部分满足特定指标的特征，并用它们表示数据[9]。两者的共同特点是：可以简化分类器的结构，降低学习模型的训练时间；在一定程度上避免“维数灾难”现象并提高分类器的推广能力。两者的区别是：特征提取得到的特征不是数据的原始特征，而特征选择得到的特征是原始特征。

对于特征提取技术的研究是一个既古老又年轻的课题。说其古老，是因为早在 20 世纪初，Hotelling 就提出了主成分分析（principal component analysis，PCA）方法[10]，并将其应用于心理学数据的统计分析。随后，一些基于不同优化准则的特征提取方法相继提出，如多维尺度（multidimensional scaling，MDS）变换[11]、独立成分分析（independent component analysis，ICA）[12]、Fisher 线性判别分析（linear discriminat analysis，LDA）[13]等。这些算法具有实现简单、容易计算、解释性强等特点，并且已经成功应用于很多领域。然而，随着信息时代的到来，人们逐渐发现在很多实际应用问题中，所采集到的样本在高维空间中往往呈现高度的非线性结构。因此，传统的基于线性模型的方法将不再适用。为了有效地处理这一问题，非线性特征提取方法应运而生。

在非线性特征提取方法中，比较具有代表性的是基于核的方法和基于流形学习的方法。其中，基于核的维数约简方法利用核方法将原始数据隐式地映射到更高维的特征空间，并期望在原始空间中呈非线性结构的数据集可以在核空间利用线性方法进行处理。由于这类方法不用创建复杂的假设空间（hypothesis space），因此很多已有的线性特征提取算法都可以扩展为对应的基于核的非线性方法，如核主成分分析（kernel principal component analysis，KPCA）[14]、核线性判别分析（kernel discriminat analysis，KDA）[15]、核独立成分分析（kernel independent component analysis，KICA）[16]等。虽然基于核的方法在处理非线性数据时较之线性方法具有一定的优势，但也存在一定的局限性。首先，由于在算法中引入了核方法，如何选择合适的核函数并对函数中的参数进行适当的设置，就成为一个难点。

恰当的核函数会给数据分析带来便利，但并不是对所有数据集均适用。因此，在实际应用中，人们经常需要根据某些经验和先验知识对核函数的类型进行选择。其次，由于在算法中采用的映射是隐式的，人们并不知道数据集在核特征空间中的具体表达，因此基于核的特征提取算法往往解释性不强。基于流形学习的特征提取方法是最近几年发展起来的一项技术，这类方法的主要特点是假设分布在高维空间中的样本点处于或者近似地处于非线性流形上，而流形学习的目标就是发现数据集中的非线性流形结构并在特征提取的同时尽可能地保持这些结构信息。目前，比较有代表性的流形学习算法主要包括等距映射(isometric mapping，ISOMAP)[17]、局部线性嵌入(locally linear embedding，LLE)[18]、拉普拉斯特征映射(Laplacian eigenmap，LE)[19]、局部切空间排列(local tangent space alignment，LTSA)[20]等。

在特征选择方面，早在20世纪60年代就有统计学领域的学者对其进行了研究。近些年，特征选择方法凭借其巨大的潜力在多个领域受到很大的关注，其应用领域包括生物信息学[21]、计算机视觉[22]、目标识别[23]及医学数据处理[24]。特征选择作为一种维数约简技术，其目标是确定出最能代表观测数据特点的特征以提高机器学习模型的性能。与PCA、LDA等特征提取技术相比，特征选择方法不改变特征的原始表达，仅从中选择一个最优的特征子集，能很好地保持特征的含义，更利于人们的理解和判断。经过半个世纪的发展，目前的特征选择技术已经形成了自己的体系结构，其按照与后续学习模型的结合方式可以分为三类：过滤式、封装式和嵌入式[9,21]。过滤式方法仅通过分析特征的固有特性来判断特征的重要程度，整个特征选择过程完全独立于后续的学习模型，因此，过滤式方法计算简单快捷且容易应用于高维度数据集，典型方法有信息增益(information gain，IG)[25]、Fisher得分(Fisher score)[26]等。封装式方法将特征选择与学习模型相结合，为每一个可能的特征子集训练一个学习模型，并使用该子集的分类识别性能作为特征选择的评估标准。这一类方法因为在特征选择过程中使用了学习模型而增加了计算量，但是同时也提高了分类准确率。典型方法有序列前向选择[27]、序列后向选择[27]和定向搜索[28]等。嵌入式方法将特征选择作为一个组成部分嵌入学习模型中，模型建立的过程也就是特征选择的过程。例如，决策树算法，该算法选择分类能力最强的特征作为节点，然后根据该特征划分子空间建立分支，重复此过程直到所有子集仅包含同一类别的数据，可见决策树的生成过程也就是特征选择的过程。嵌入式方法不需要将样本分为训练集和测试集，这在一定程度上减少了计算复杂度。实际应用中，封装式方法和嵌入式方法的准确率往往高于过滤式方法，但在以微阵列基因表达数据等为代表的高维数据特征选择问题中，过滤式方法使用更为广泛，主要原因有两点：首先，与封装式和嵌入式方法相比，过滤式方法计算更简单、快速；其次，过滤式方法的特征选择结果独立于学习模型，这使其能灵活地应用于多种学习器。

目前所研究的过滤式特征选择方法主要分为两类：排序法和空间搜索法[29]。基于排序的过滤式方法研究较早，其将特征选择看做是一个排序问题，为每一个特征都计算一个权重或得分，然后选择权重大或得分高的特征作为最后的特征子集。对于此类方法，如何计算特征权重（得分）是整个算法的关键。在早期的研究中，学者们引入了很多其他学科的方法和理论来计算特征权重，其中包括 Pearson 相关系数（Pearson correlation coefficient，PCC）[30]、Kendall 等级相关系数[31]、互信息[32]和信息增益[25]等。Pearson 相关系数和 Kendall 等级相关系数是基于统计学的方法。Pearson 相关系数可以度量两个变量之间的线性关系，但其只适用于连续变量，对于离散变量可以使用 Kendall 等级相关系数度量其相似度。互信息和信息增益是基于信息论的方法，这些方法主要用于度量变量之间的非线性关系。1992 年，Kira 等提出了一种名为 Relief 的特征权重计算方法[33]，Relief 算法如今已经成为特征选择领域的经典算法，其在多年中被众多学者学习研究，并形成了 Relief 系列算法。这一系列算法通过分析在某特征下样本与其来自多个类的近邻样本点之间的关系判断该特征的重要性。1995 年，Bishop 等根据 Fisher 准则提出 Fisher 得分算法[26]计算特征权重，其基本思想是根据特征的判别能力为特征计算权重，特征的类别区分能力越强，权重越大，反之权重越小。2006 年，He 等提出了 Laplacian 得分（Laplacian score）[34]，Laplacian 得分是无监督算法，其主要思想与特征提取中的 LE[19]算法相似，两个算法都是希望最后的特征能够更好地保持原始数据流形结构。2008 年，Zhang 等提出了 Constraint 得分[22]，Constraint 得分能在有监督和半监督两种情况下计算特征权重，这一特点使其应用更为广泛。基于排序的过滤式特征选择方法已经成功应用于很多实际问题，但这类方法有一个不可忽视的缺点：所选特征子集中往往含有冗余[21,29]。冗余严重降低了特征子集携带的信息量，为解决这个问题，学者提出了基于空间搜索的过滤式方法。该方法不再像基于排序的方法那样为每个特征打分，它们通常通过优化一个评价函数来确定所选特征，评价函数中既要包括对特征重要性的度量，又要包括对子集冗余度的度量[29]。经典的基于空间搜索的过滤式方法是 1996 年 Koller 等提出的 Markov blanket 过滤式算法[35]，该算法使用条件熵计算特征的冗余度。2000 年，Hall 和 Holmes 提出了基于相关性的特征选择（correlation based feature selection，CFS）算法[36]，其评价函数定义为特征预测能力与特征间关联程度的商。在 CFS 算法提出后不久就有学者提出了快速的基于相关性的特征选择（fast correlation based feature selection，FCFS）算法[37]。FCFS 算法既保持了 CFS 算法能够去除特征子集冗余的特点，又降低了算法复杂度，使运算更快速。2005 年，Peng 等提出了最小冗余最大相关（minimal redundancy maximal relevance，MRMR）算法[38,39]，MRMR 算法使用互信息度量特征的相关性与冗余度，并且其使用信息差和信息熵两个代价函数来评价特征子集。

第 2 章　维数约简技术简介

本章对目前应用比较广泛的维数约简方法进行介绍。在特征提取方面，本章的内容既涉及传统的线性特征提取方法（如 PCA[10]、MDS[11]、LDA[13]、非负矩阵分解（NMF）[40,41]），也包括基于核的非线性特征提取方法（如 KPCA[14]、KDA[15]）和基于流形学习的非线性特征提取方法（如 ISOMAP[17]、LLE[18]、LE[19]、LTSA[20]、最大方差展开（MVU）、局部敏感判别分析（locality sensitive discriminant analysis，LSDA）[42]）。在特征选择方面，本章主要介绍基于排序的过滤式特征选择算法（如 Laplacian 得分[34]、Constraint 得分[22]、Fisher 得分[26]、PCC[30]、Relief-F[43]、信息增益（IG）[25]、基于有效范围的基因选择（effective range based gene selection，ERGS）方法[44]）和基于空间搜索的过滤式特征选择算法（如 CFS[36] MRMR 算法[38,39]）。

2.1　特征提取方法

本节主要介绍线性特征提取和包括流形学习在内的非线性特征提取方法。

2.1.1　线性特征提取方法

1. 主成分分析

PCA 是多元统计分析和数据挖掘中最广为人知的数据维数约简方法。早在 1901 年，Pearson 就对 PCA 的主要思想进行了阐述[45]，此后，Hotelling又进一步发展了这一算法并将其应用于心理学数据的研究[10]。此外，Karhunen 和 Loeve 也在随机过程的框架下提出了 PCA 算法[46,47]。因此，在一些文献中，PCA 算法又称 Karhunen-Loeve 变换。

假设存在一组高维数据 $X=\{x_i, i=1,\cdots,N\}\in\mathbf{R}^D$，且均值为 0，PCA 的目的就是通过线性变换的方式找到数据集的低维描述 $Y=\{y_i, i=1,\cdots,N\}\in\mathbf{R}^d$ $(d\ll D)$，即 $Y=A^{\mathrm{T}}X$。对于 PCA 中矩阵 A 的求解，主要有最小化重构误差和最大方差变化两种方式，虽然这两种方式从不同的角度出发，但最终得到的结果却是一致的。

（1）从最小重构误差的角度，PCA 的目标函数可表示为

$$\min_A \|E\| = \min_A \|X-AA^{\mathrm{T}}X\|^2 \quad \text{s.t.} \quad A^{\mathrm{T}}A=I \tag{2.1}$$

式中，A 为 X 在高维空间中所张成的子空间的部分正交基，E 为重构误差。换句话说，PCA 的目标就是在最小二乘的意义下，寻找一组使重构误差达到最小的 A。通过简单的数学推导，式(2.1)可转化为

$$\begin{aligned}\min_A \| E \| &= \min_A \| X - AA^{\mathrm{T}}X \|^2 \\ &= \min_A \mathrm{tr}\{(X - AA^{\mathrm{T}}X)^{\mathrm{T}}(X - AA^{\mathrm{T}}X)\} \\ &= \min_A \mathrm{tr}\{X^{\mathrm{T}}X - 2X^{\mathrm{T}}AA^{\mathrm{T}}X + X^{\mathrm{T}}AA^{\mathrm{T}}X\} \\ &= \min_A \mathrm{tr}\{X^{\mathrm{T}}X - X^{\mathrm{T}}AA^{\mathrm{T}}X\} \end{aligned} \tag{2.2}$$

式中，$X^{\mathrm{T}}X$ 为常量，因此，最小化式(2.2)就变为最大化 $X^{\mathrm{T}}AA^{\mathrm{T}}X$ 的问题。通过推导，可以将式(2.2)转化成关于 A 的二次型最大化问题，因此可以通过特征值分解的方式求解，即求样本协方差矩阵 XX^{T} 前 d 个最大特征值所对应的特征向量。

(2) 从另一个角度，PCA 算法的目的是找到高维数据集中彼此正交且数据方差变化最大的几个方向。设高维数据的协方差矩阵为

$$C_{xx} = E\{XX^{\mathrm{T}}\} \tag{2.3}$$

则根据低维数据与高维数据之间的线性变换关系，低维空间中样本的协方差矩阵可表示为

$$C_{yy} = E\{YY^{\mathrm{T}}\} = E\{A^{\mathrm{T}}XX^{\mathrm{T}}A\} = A^{\mathrm{T}}C_{xx}A \tag{2.4}$$

通过式(2.4)可以看出，在这里对矩阵 A 的求解同样可以转化为对高维空间样本的协方差矩阵的特征值分解。

PCA 算法将数据方差的大小作为对信息衡量的标准，认为方差越大，它所能够提供的信息就越多，反之，提供的信息就越少。因此，PCA 的维数约简实际上就是一个坐标变换过程，即将数据从高维观测坐标系投影至由一组彼此正交且数据方差变化最大方向所组成的坐标系(图 2.1)。虽然 PCA 具有计算简单、解释性强等优点，但还存在以下不足[1]：

(1) 当样本数据集具有非线性结构时，PCA 所得到的维数约简结果不能有效地反映数据的本质特征。

(2) PCA 能够找到数据方差变化最大的方向，但这些方向对于分类和识别问题不一定是最有利的。

(3) 对于 PCA 中所要保持的主分量的个数比较难估计。虽然在某些情况下，可以通过数据协方差矩阵相邻特征值之间的比值对主分量的个数进行估计，但当特征值变化比较平缓时，难以对主分量进行取舍。

(4) 在某些问题中，难以对 PCA 所求得的主分量进行解释。例如，在图像处理问题中，图像中的像素值都是非负的，当利用 PCA 对一组图像进行维数约简处理后，主分量中的负值难以进行语义上的解释。

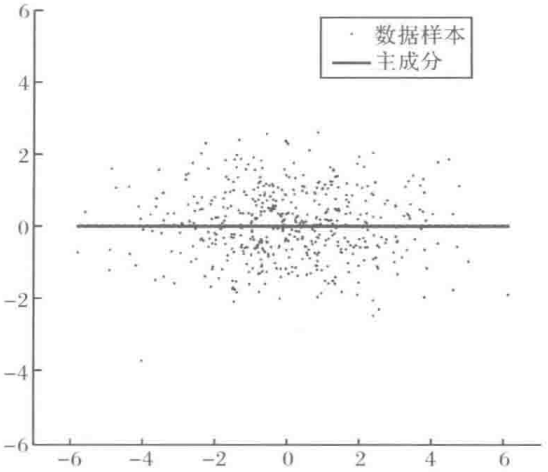

图 2.1　PCA 算法的几何直观：寻找方差最大的方向

2. 多维尺度变换

MDS 源于对心理学的研究，其基本思想是：给定高维数据点之间的不相似性矩阵，寻找低维空间中与其相对应的一组数据，使其点间距离与高维数据点之间的不相似性相一致[48]。如果给定的不相似性度量是欧氏距离，那么 MDS 所得到的低维结果与 PCA 相同。

设高维数据集 X 中各元素的不相似性矩阵为 $\Delta=(\delta_{ij})_{N\times N}$，需要寻找一组低维数据点 $Y=\{y_i,i=1,\cdots,N\}\in\mathbf{R}^d$，使它们之间的距离 s_{ij} 尽可能地接近 δ_{ij}。如果采用欧氏距离作为相似性度量，那么

$$\delta_{ij}^2 = s_{ij}^2 = \| x_i - x_j \|^2 = x_i^{\mathrm{T}}x_i - 2x_i^{\mathrm{T}}x_j + x_j^{\mathrm{T}}x_j \tag{2.5}$$

记 N 维向量 $\varphi=(x_1^{\mathrm{T}}x_1,\cdots,x_N^{\mathrm{T}}x_N)^{\mathrm{T}}$，$e=(1,1,\cdots,1)^{\mathrm{T}}$，距离矩阵可表示为

$$D = \varphi e^{\mathrm{T}} - 2X^{\mathrm{T}}X + e\varphi^{\mathrm{T}} \tag{2.6}$$

令 $H=I-ee^{\mathrm{T}}/N$ 为中心化矩阵，则有

$$B = \frac{-HDH}{2} = X^{\mathrm{T}}X \tag{2.7}$$

记 B 的特征值分解为

$$B = U\ \mathrm{diag}(\lambda_1,\cdots,\lambda_N)U^{\mathrm{T}} \tag{2.8}$$

式中，U 为正交矩阵且特征值为降序排列 $\lambda_1\geqslant\cdots\geqslant\lambda_N$，则维数约简结果为

$$Y = \mathrm{diag}(\lambda_1^{1/2},\cdots,\lambda_d^{1/2})U_d \tag{2.9}$$

式中，U_d 为式(2.8)中前 d 个最大特征值所对应的特征向量组成的矩阵。值得注意的是，在式(2.8)中是对 $X^{\mathrm{T}}X$ 进行特征分解，而在 PCA 算法中，是对 XX^{T} 进行特征值分解，显然，采用欧氏距离的 MDS 算法与 PCA 算法互为对偶形式。在实

际问题中，当高维数据是以向量的形式给出时，MDS 算法和 PCA 算法同样适用，而当数据是以距离的形式给出时，PCA 算法将不再适用，只能利用 MDS 算法对数据进行处理。此外，在 MDS 算法中，低维数据点是由式(2.9)直接构造的，因此高维数据与低维数据之间的映射是隐式的，这一特点使 MDS 算法主要应用于数据的可视化。

目前，基于传统 MDS 算法的扩展算法主要包括 Sammon Map[49]、Metric Embedding[50]、Interative MDS[51]、Distribution MDS[52]、Robust MDS[53]、Landmark MDS[54]和 Block MDS[55]等。

3. Fisher 线性判别分析

虽然 PCA 算法可以有效地对高维数据进行维数约简处理，但通过 PCA 算法所找到的主成分却无法区分不同类别的数据。因此，在维数约简过程中引入监督机制的 Fisher 线性判别分析方法在分类和识别问题中得到了广泛应用。在 LDA 算法中，所要找到的投影方向是最能区分各个类别数据的方向，即样本沿着该方向，在 Fisher 准则下能够得到最好的区分。

在 LDA 算法中，首先需要对数据集的类内散度矩阵(S_w)和类间散度矩阵(S_b)进行定义，假设数据集中共包含 C 个类别，则 S_w 和 S_b 的计算公式为

$$S_b = \frac{1}{N}\sum_{i=1}^{C} N_i (m_i - m)(m_i - m)^{\mathrm{T}} \tag{2.10}$$

$$S_w = \sum_{i=1}^{C} S_i \tag{2.11}$$

式中，N_i 为类别 i 中包含的样本数；m_i 和 m 分别为类别 i 和总体样本的均值；S_i 为属于类别 i 中的样本的协方差矩阵。LDA 算法的准则函数定义为

$$J(A) = \frac{\mathrm{tr}(A^{\mathrm{T}} S_b A)}{\mathrm{tr}(A^{\mathrm{T}} S_w A)} \tag{2.12}$$

求解可以使该准则函数最大化的 A 的过程需要一些技巧，但最后的解的形式却比较简单，最优矩阵 A 的列向量是下面等式中最大特征值所对应的特征向量，即

$$S_b a_i = \lambda_i S_w a_i \tag{2.13}$$

作为一种有监督的数据维数约简算法，LDA 算法由于在维数约简过程中考虑到高维数据的类别信息，相对于前面提到的 PCA 算法和 MDS 算法，它所得到的低维数据点的区分性更强。然而，LDA 算法同样是线性方法，因此不适用于非线性结构的数据，同时，由于类间散度矩阵 S_b 中至多包含 $C-1$ 个非零特征值，因此在 LDA 算法中，得到的低维空间维数也至多是 $C-1$ 维，这就限制了 LDA 在一些实际问题中的应用。目前，基于 LDA 的改进算法主要有 KDA(kernel discriminat analysis)[15]、SDA(Sub-class discriminat analysis)[56]、BODA(Bayes optimized

discriminant analysis)[57]等。

4. 非负矩阵分解

关于 NMF 的研究源于认知科学领域，一直以来，从事认知学研究的科学家所热衷的问题是：对于事物整体的认知是否源于基于局部分量的认知？一些来自心理学和神经生理学的证据表明，某些目标的识别依赖以上的假设。因此，如何利用数学方法对人类的这一认知行为进行模拟，就成为研究的热点。在认知方式的研究中，Lee 和 Seung 认为，当认知更关注局部信息时，可以通过某种算法对其内在部件进行学习，并根据这一论断提出了 NMF 算法来学习文本的语义特征和人脸图像的分量或部件(parts)[40,41]。与前面提到的 PCA 算法等不同，NMF 算法所学习到的是基于部件的表示，而不是全局分量。为了达到这一目的，NMF 算法对矩阵引入了非负的限制。图 2.2 展示了 PCA 算法和 NMF 算法在人脸图像处理中的对比，从图中可以看出，在 NMF 算法中，只允许图像由一系列基图像或部件的加法来构成，而不允许减法或加减法组合的出现。因此，在理论上保证了基于部件的认知行为。同时，NMF 算法的这一特点，使其所得到的基图像的解释性也要优于 PCA 算法。

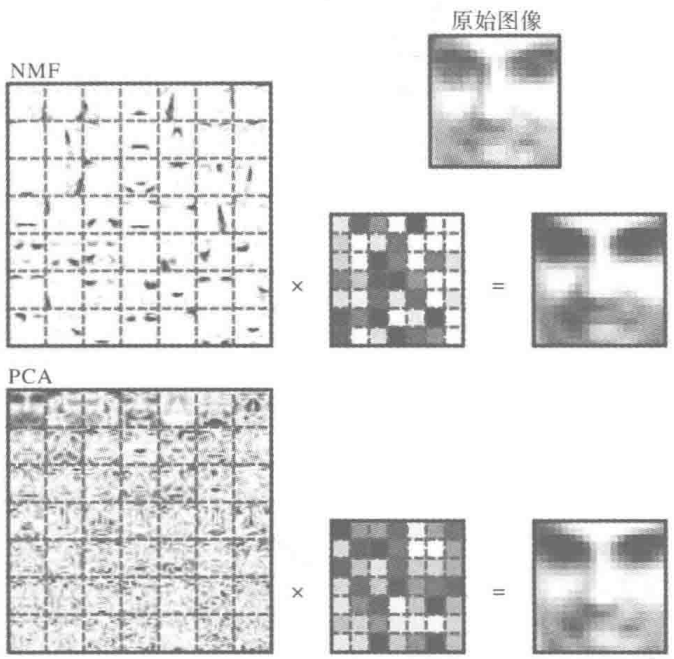

图 2.2　NMF 算法与 PCA 算法在人脸图像处理中的对比
(黑色代表正值，红色代表负值[40])(见彩图)

实际上，在统计机器学习中，包括 PCA、SVD、NMF 等算法都可以表示为矩阵