Statistics and Application 统计学与应用, 2020, 9(1), 47-52
Published Online February 2020 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2020.91006

Hans汉斯

Dimension Reduction Comparison between PCA and LDA

Lihong Bao

Yunnan University of Finance and Economics, Kunming Yunnan
Email: 2041467626@qq.com

Received: Dec. 27th, 2019; accepted: Jan. 12th, 2020; published: Jan. 19th, 2020

Abstract

Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are commonly used in machine learning. In this paper, we extend PCA and LDA to 2DPCA and 2DLDA, 2DPCA and 2DLDA are directly to reduce the dimension on the original matrix data structure, overcoming the problem of the curse of dimensionality. Empirical research suggests that 2DLDA is a better dimensionality reduction classification method than 2DPCA compared with the effect of dimension reduction and the error rate of classification.

Keywords

PCA, LDA, Matrix Data

主成分分析与线性判别分析降维比较

保丽红

云南财经大学，云南 昆明
Email: 2041467626@qq.com

收稿日期：2019年12月27日；录用日期：2020年1月12日；发布日期：2020年1月19日

摘 要

主成分分析(PCA)和线性判别分析(LDA)是机器学习领域中常用的降维方法。本文针对矩阵型数据结构，将一维的降维方法PCA和LDA推广为二维PCA和二维LDA，2DPCA和2DLDA对矩阵型数据进行降维处理时，克服了维数灾难的问题。实验研究表明，对比降维效果和分类错误率，2DLDA相比2DPCA是一种更为出色的降维分类方法。

文章引用: 保丽红. 主成分分析与线性判别分析降维比较[J]. 统计学与应用, 2020, 9(1): 47-52.
DOI: 10.12677/sa.2020.91006

关键词

主成分分析，线性判别分析，矩阵型数据

1. 引言

在数据降维处理中[1]，比较常用的方法为线性变换技术。即通过线性投影将原样本数据投影到低维子空间中，其中典型的有 PCA 与 LDA。PCA 的意义是在重构时其平方误差最小；LDA 是一种有监督的线性降维算法[2]，其基本思想是选择使得 Fisher 准则函数达到极值的向量作为最佳投影方向，从而使得样本在该方向上投影后，达到最大的类间散布距离和最小的类内散布距离。PCA 和 LDA 的不同之处在于，无监督的 PCA 能保持数据信息，而 LDA 是使降维后的数据点尽可能地容易被区分。

针对矩阵型的数据结构，我们将一维的降维方法 PCA 和 LDA 在矩阵模式上推广为二维 PCA 和二维 LDA。采用二维降维方法 2DPCA 和 2DLDA，其最大的优势于是不需要将高数据转化向量，克服了维数灾难[3]。

本文将在真实的两个数据集上，验证在不同的多元时间序列数据集下，2DPCA 和 2DLDA 两种数据降维方法效果的优劣。

2. 研究方法概述

本节分别介绍二维主成分分析(2DPCA)算法，二维线性判别分析(2DLDA)算法。

2.1. 二维主成分分析(2DPCA)算法

令 $X=\{x_i\}_{i=1}^N$ 是一组样本，$x_i\in\mathbb{R}^{r\times c}$，则样本平均值为：

$$\bar{X}=\frac{1}{N}\sum_{i=1}^{N}x_i \tag{1}$$

定义协方差矩阵为 G：

$$G=\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{X})(x_i-\bar{X})^T \tag{2}$$

其中 G 为 $r\times r$ 的非负定矩阵。对 G 进行特征值分解，最大的 d 个特征值所对应的标准正交的特征向量构成投影向量组 $U=(u_1,\cdots,u_d)$。

2.2. 二维线性判别分析(2DLDA)算法

令 $X=\{x_i\}_{i=1}^N$ 是一组样本，$x_i\in\mathbb{R}^{r\times c}$，其中 X 分为 $\pi_i(i=1,2,\cdots,k)$ 类，n_i 为第 i 类样本的个数，则样本均值为：

$$\bar{X}=\frac{1}{N}\sum_{i=1}^{N}x_i \tag{3}$$

第 i 类样本的类内平均：

$$\bar{X}_{\pi_i} = \frac{1}{n_i} \sum_{x_j \in \pi_i}^{n_i} x_j \tag{4}$$

定义如下类间散步矩阵：

$$S_b = \frac{1}{N} \sum_{i=1}^{k} n_i \left(\bar{X}_{\pi_i} - \bar{X} \right) \left(\bar{X}_{\pi_i} - \bar{X} \right)' \tag{5}$$

类内散布矩阵：

$$S_w = \frac{1}{N} \sum_{i=1}^{k} \sum_{x \in \pi_i} \left(x - \bar{X}_{\pi_i} \right) \left(x - \bar{X}_{\pi_i} \right)' \tag{6}$$

2DLDA 寻找的最佳投影矩阵 U：

$$J(U) = \frac{U'S_bU}{U'S_wU}。 \tag{7}$$

3. 实证分析

本节将在 Wafer、Ausla 这两个真实的数据集，验证在不同的数据集下，2DPCA 和 2DLDA 两种数据降维方法效果的优劣。

3.1. 数据集介绍

3.1.1. Wafer 数据集

Wafer 数据集是由一个真空传感器应用一个硅晶片在刻画中记录下来的。该数据集记录的晶片类型分为两个类型：“正常”与“不正常”。其中“正常”类型的样本数为 1067 个，“不正常”类型的样本数为 127 个。

3.1.2. AUSLAN 数据集

AUSLAN 数据集有 2565 个数据样本，包含 95 个语音信号，每个信号由 27 个样本组成，其中每个样本的长度在 47 到 95 之间。每一个样本有 22 个变量，记录了这 25 个由当地人发音的语音信息。

2 个数据集描述见表 1 所示。

Table 1. Data set description summary
表 1. 数据集描述汇总

	Wafer	AUSLAN
变量数	6	22
最大样本长度	198	95
最小样本长度	104	47
分类个数	2	25
样本量	327	675

3.2. 降维效果

为了降低实验结果的变化性，我们在每次数据集上重复 5 次实验，下面分别给出了各个数据集 2DPCA 和 2DLDA 的实验结果。

3.2.1. Wafer 数据集上的实验效果

Table 2. Error rate results and dimension reduction results of 2DPCA and 2DLDA on Wafer dataset
表 2. 2DPCA 和 2DLDA 在 Wafer 数据集上的错误率结果和降维结果

实验次数	初始维数	2DPCA		2DLDA	
		错误率	维数	错误率	维数
1	[104, 6]	0.1523	[8, 4]	0.0457	[12, 5]
2		0.1421	[9, 5]	0.0660	[27, 5]
3		0.1624	[8, 4]	0.0660	[28, 5]
4		0.1777	[9, 5]	0.0508	[1, 6]
5		0.1066	[14, 6]	0.0457	[4, 5]

表 2 给出了 2DPCA 和 2DLDA 在 Wafer 数据集上的 5 次实验的错误率结果和降维结果。由表 2 可知，相比较，2DPCA 的降维效果比 2DLDA 的降维效果好，但 2DPCA 的分类错误率比 2DLDA 的高。图 1 直观呈现出两种方法实验的分类错误率。

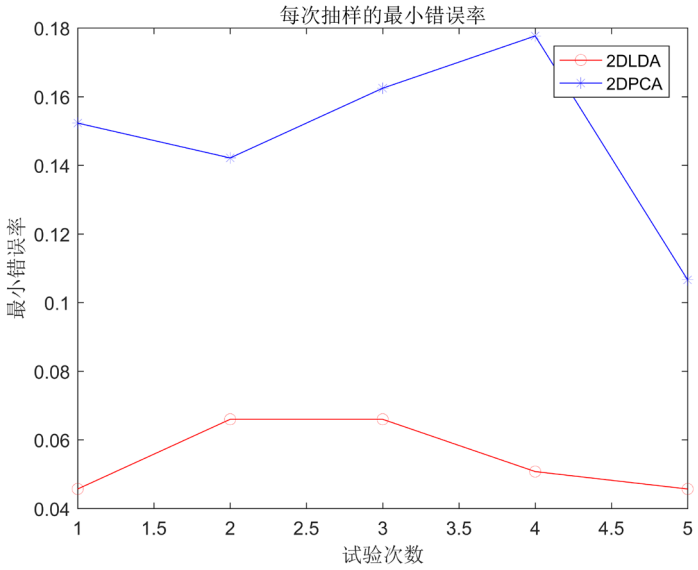

Figure 1. Error rates of 2DPCA and 2DLDA on wafer datasets
图 1. 2DPCA 和 2DLDA 在 wafer 数据集上的错误率

3.2.2. AUSLAN 数据集上的实验效果

Table 3. Error rate results and dimensionality reduction results of 2DPCA and 2DLDA on the AUSLAN dataset
表 3. 2DPCA 和 2DLDA 在 AUSLAN 数据集上的错误率结果和降维结果

实验次数	初始维数	2DPCA		2DLDA	
		错误率	维数	错误率	维数
1	[47, 22]	0.0500	[1, 21]	0.0300	[1, 21]
2		0.0733	[1, 20]	0.0233	[1, 22]
3		0.0500	[1, 20]	0.0300	[2, 20]
4		0.0500	[1, 20]	0.0233	[1, 20]
5		0.0633	[1, 21]	0.0233	[1, 21]

表 3 给出了 2DPCA 和 2DLDA 在 AUSLAN 数据集上的 5 次实验的错误率结果和降维结果。由表 3 可知，相比较，2DPCA 的降维效果与 2DLDA 的降维效果差别不大，但 2DPCA 的分类错误率明显比 2DLDA 的高。图 2 直观呈现出两种方法实验的分类错误率。

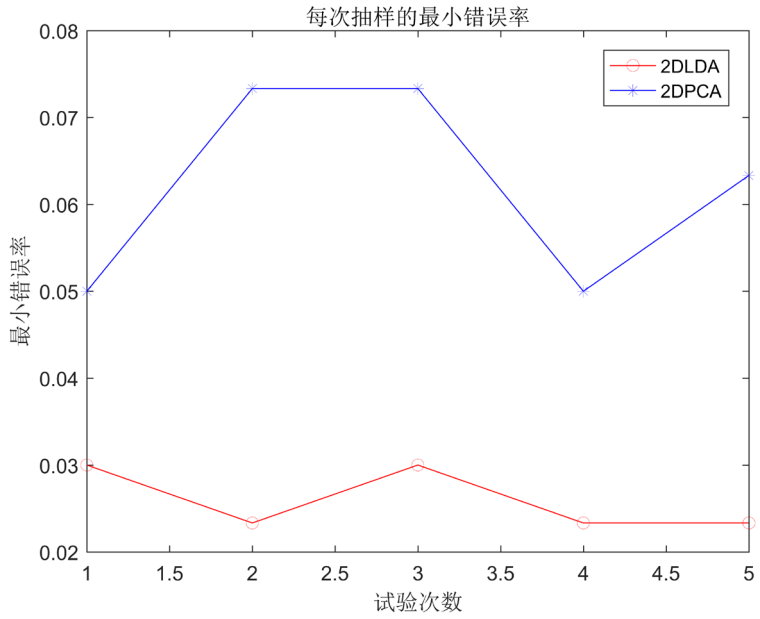

Figure 2. Error rates of 2DPCA and 2DLDA on the AUSLAN dataset
图 2. 2DPCA 和 2DLDA 在 AUSLAN 数据集上的错误率

4. 结论

通过提取 2DPCA 和 2DLDA 分别在 wafer、Auslan 这两个真实的数据集上的最小错误率，我们确定了用不同方法进行降维的最佳维数，通过每种方法降维的最佳维数提取了与之相对应的错误率，最后我们求出相应的平均错误率，如表 4 所示。

Table 4. Average classification error rate results and dimensionality reduction results of 2DPCA and 2DLDA on 3 datasets
表 4. 2DPCA 和 2DLDA 在 3 个数据集上的平均分类错误率结果和降维结果

数据集	2DPCA		2DLDA	
	错误率	维数	错误率	维数
wafer	0.1534	[9, 6]	0.0670	[2, 5]
Auslan	0.0653	[1, 20]	0.0320	[2, 20]

由表 4 可知：2DLDA 的平均分类错误率均小于 2DPCA。这说明判别分析的分类效果要优于主成分分析的分类效果。因此，综合降维效果和分类错误率这两个因素，对比实验证实了 2DLDA 相比 2DPCA 是一种更为出色的分类方法。

参考文献

[1] Zhang, K. and Kwok, J.T. (2010) Clustered Nyström Method for Large Scale Manifold Learning and Dimension Reduction. *IEEE Transactions on Neural Networks*, **21**, 1576-1587. https://doi.org/10.1109/TNN.2010.2064786

[2] Zuo, W.M., Zhang, D., Yang, J. and Wang, K.Q. (2006) BDPCA Plus LDA: A Novel Fast Feature Extraction Tech-

nique for Face Recognition. *IEEE Transactions on Systems*, *Man*, *and Cybernetics*, *Part B*: *Cybernetics*, 36, 946-953. https://doi.org/10.1109/TSMCB.2005.863377

[3] Xie, X.C., Yan, S.C., Kwok, J.T. and Huang, T.S. (2008) Matrix-Variate Factor Analysis and Its Applications. *IEEE Transactions on Neural Networks*, **19**, 1821-1826. https://doi.org/10.1109/TNN.2008.2004963