软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(4):826−838 [doi: 10.13328/j.cnki.jos.004571]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于边界判别投影的数据降维*

何进荣, 丁立新, 李照奎, 胡庆辉

(软件工程国家重点实验室(武汉大学 计算机学院),湖北 武汉 430072)

通信作者: 何进荣, E-mail: hejinrong@whu.edu.cn, http://www.whu.edu.cn

摘 要: 为了提取具有较好判别性能的低维特征,提出了一种新的有监督的线性降维算法——边界判别投影,即,最小化同类样本间的最大距离,最大化异类样本间的最小距离,同时保持数据流形的几何形状.与经典的基于边界定义的算法相比,边界判别投影可以较好地保持数据流形的几何结构和判别结构等全局特性,可避免小样本问题,具有较低的计算复杂度,可应用于超高维的大数据降维.人脸数据集上的实验结果表明,边界判别分析是一种有效的降维算法,可应用于大数据上的特征提取.

关键词: 边界判别投影;数据降维;特征提取;边界样本点;人脸识别

中图法分类号: TP181 **文献标识码**: A

中文引用格式: 何进荣,丁立新,李照奎,胡庆辉.基于边界判别投影的数据降维.软件学报,2014,25(4):826−838. http://www.jos.org.cn/1000-9825/4571.htm

英文引用格式: He JR, Ding LX, Li ZK, Hu QH. Margin discriminant projection for dimensionality reduction. Ruan Jian Xue Bao/Journal of Software, 2014,25(4):826−838 (in Chinese). http://www.jos.org.cn/1000-9825/4571.htm

Margin Discriminant Projection for Dimensionality Reduction

HE Jin-Rong, DING Li-Xin, LI Zhao-Kui, HU Qing-Hui

(State Key Laboratory of Software Engineering (Computer School, Wuhan University), Wuhan 430072, China)

Corresponding author: HE Jin-Rong, E-mail: hejinrong@whu.edu.cn, http://www.whu.edu.cn

Abstract: A novel supervised linear dimensionality reduction algorithm called margin discriminant projection (MDP) is proposed to extract low-dimensional features with good performance of discriminant. MDP aims to minimize maximum distance of samples belong to the same class and maximize minimum distance of samples belong to different classes, and at the sametime preserve the geometrical structure of data manifold. Compared with classical algorithms based on the definition of margin, MDP is good at preserving the global properties, such as geometrical and discriminant structure of data manifold, and can overcome small size sample problem. Due to its low cost of computation, MDP can be directly applied on ultra-high dimensional big data dimensionality reduction. Experimental results on five face data sets show its effectiveness for feature extraction on big data.

Key words: margin discriminant projection; dimensionality reduction; feature extraction; margin sample; face recognition

随着计算机通信技术和存储技术的发展,数据的获取越来越便利.如何处理海量数据,成为现代科学研究面临的一个巨大挑战.传统的数据挖掘技术在数据维数和规模扩大时,所需资源呈指数级增加,研究既有效又简易的数据表示方法,从而消除数据冗余,建立高效率低成本的数据处理、存储和通信新技术,是大数据分析的一大难题[1].因此,需要研究大数据降维算法,通过寻求其低维特征表示,发现数据潜在的有价值信息.

数据降维算法可以分为线性和非线性两类,其中,非线性算法主要是流形学习方法,经典算法包括等度规特征映射(isometric feature mapping,简称 ISOMAP)[2]、局部线性嵌入(local linear embedding,简称 LLE)[3]、

* 基金项目: 中央高校基本科研业务费专项资金(2012211020209); 广东省省部产学研结合专项资金(2011B090400477); 珠海市产学研合作专项资金(2011A050101005, 2012D0501990016); 珠海市重点实验室科技攻关项目(2012D0501990026)

收稿时间: 2013-10-15; 修改时间: 2013-12-18; 定稿时间: 2014-01-27

Laplacian 特征映射(Laplacian eigenmap,简称 LE)[4]、局部切空间重排列(local tangent space alignment,简称 LTSA)[5]、Hessian 特征映射(Hessian eigenmaps,简称 HLLE)[6]、最大方差展开(maximum variance unfolding,简称 MVU)[7]等.近年来,流形学习已被成功应用于人脸识别[8]、图像检索[9,29]、文本分类[10]和视频分析[11,12]等领域.然而,在实际应用中仍然存在一些困难:一是所谓的“外样本(out-of-sample)”问题[13],即,高维数据和低维表示之间的映射关系是隐式的,需要通过对所有训练样本进行学习来获得,对于新的测试样本,无法直接求得其低维表示;二是“局部过学习(overlearning of locality)”问题[14],流形学习方法通常是考虑了数据的局部结构,并在局部信息保持的前提下进行特征提取,然而这与用户所要提取的信息并无直接关联;三是计算复杂度高,尤其是在图像和文本等超高维数据的分析处理中,变得难以施行.由于线性算法可以克服以上难点,从而引起了许多研究者的关注.

传统的线性降维方法有主成分分析(principle component analysis,简称 PCA)[15]和线性判别分析(linear discriminant analysis,简称 LDA)[16,17]两种.PCA 是一种无监督算法,没有考虑样本的类别信息,从而不能有效地发现数据潜在的判别结构;而 LDA 算法可直接应用于分类任务,通过最大化类间散度和最小化类内散度,从而获取其低维表示.但是 LDA 要求类内散度矩阵非奇异,这在实际应用中难以保证.比如,当样本维数高于样本个数时,类内散度矩阵是奇异的,这种情形通常称为小样本(small size sample,简称 SSS)问题[18],是机器学习研究领域的一个开放性难题.另外,LDA 最多只能提取 C–1 维的特征(这里,C 表示样本的类别个数),这限制了 LDA 的应用范围.

为了克服 LDA 应用中的小样本问题,李海峰等人[19]提出了最大边界准则(maximum margin criterion,简称 MMC);之后,有学者通过正则化技术对 MMC 进行了一系列的改进[20−22].邱锡鹏等人[23]提出了一种非参数边界最大化准则,通过最大化每个样本点与其不同类的最近样本点距离和最小化每个样本点与其同类的最远样本点距离,来获得低维表示.颜水成等人[24,25]在图嵌入的框架下提出了边际 Fisher 分析,其目标在于最大化局部类间散度和最小化局部类内散度,通过数据建图,MFA 使得局部边界增大,从而获得判别性能更好的低维表示.MFA 可以看作是 LDA 的局部化扩展.由于线性降维算法等价于学习一个合适的马氏距离度量,Weinberger 等人[26]从度量学习的角度提出了基于 K 最近邻分类器的大边界准则,并将其归结为半定规划问题.王飞、张长水等人[27]提出了平均邻域边界最大化(average neighborhood margin maximization,简称 ANMM)算法,该算法将属于同一类的近邻点尽可能地“拉近”,而将属于不同类的近邻点尽可能地“推开”.与 MMC 考虑数据的全局欧氏结构不同,王海贤等人[28]提出了局部和加权的最大边界判别分析(local and weighted maximum margin discriminant analysis,简称 LWMMDA),考虑了数据流形的局部性质,通过权重参数来调整不同类别之间的平均边界,可以有效发现非线性数据流形中的判别结构.蔡登等人[29]根据局部判别分析的思想,构建近邻图以对数据流形的几何结构进行建模,提出了一种半监督的线性降维方法,称为最大边界投影(maximum margin projection,简称 MMP),并将其应用于图像检索领域.这些基于边界思想的算法在应用于高维大数据集时,都具有较高的计算代价,或者容易导致数值不稳定.

本文重新定义了样本边界,提出了一种新的有监督的线性降维算法,将其称为边界判别投影(margin discriminant projection,简称 MDP).MDP 通过最大化属于不同类别的样本之间的最小距离和最小化属于同一类别的样本之间的最大距离,来获得具有最佳判别性能的低维表示.与经典的基于边界思想的数据降维算法相比,MDP 算法的优点概述如下:

(1) 边界的定义具有几何直观性.引入边界概念的目的是描述属于不同类别的样本之间的线性可分性,最大化属于不同类别的样本之间的最小距离,这意味着任意两个属于不同类别的样本之间的距离都在增大;而最小化属于不同类别的样本之间的最大距离,就意味着任意两个属于同一类别的样本之间的距离都在减小;

(2) 计算复杂性低.MDP 算法只考虑了边界样本点(具体定义见第 2.1 节),即,属于同一类的距离最大的两个样本点以及属于不同类的距离最小的两个样本点,且不涉及参数调整.在数据建图中,表征样本间相似关系的权值矩阵严重稀疏,给大规模数据矩阵的快速特征分解提供了便利;

(3) 可直接应用于小样本问题.与 MMC 类似,MDP 算法可归结为迹差准则优化问题,其求解过程中不涉及逆矩阵的计算,从而避免了奇异性问题,使得数值求解更加稳定;

(4) 在保持数据流形的几何结构的同时,发现其判别结构.MDP 得到的投影矩阵的列向量是 R^r 一组规范正交基,相当于将原始高维数据在这组规范正交基下进行正交分解,由此得到的新的低维坐标表示不会改变原始数据的度量结构和分布形状(证明见第 2.3.1 节中的定理 3),从而保持了数据的全局几何结构.

1 相关工作

给定数据矩阵 $X=\{x_1,x_2,\ldots,x_n\}\in R^{d\times n}$,即 n 个 d 维的样本数据,每个样本对应的类别为 $label(x_i)\in C$,其中,$C=\{c_1,c_2,\ldots,c_m\}$表示类别集合.线性降维算法通常假设高维数据 x_i 与其低维表示 y_i 之间具有显式的映射关系,即:

$$y_i=V^Tx_i \tag{1}$$

其中,$V=\{v_1,v_2,\ldots,v_r\}\in R^{d\times r}$ 称为投影矩阵(其中,$r<d$).

人们对数据感兴趣的特征不同,从而导致了不同的降维算法.通常,用于降维准则设计的思路有数据流形的全局结构(如距离保持等)、数据流形的局部结构(如近邻关系保持等)和数据类别的判别结构.其中,基于边界思想的经典算法有线性判别分析(LDA)、最大边界准则(MMC)和边界 Fisher 分析(MFA).

1.1 线性判别分析(LDA)

LDA 通过最大化类间散度,同时最小化类内散度,来获得判别性较强的低维表示.其中,类间散度矩阵 S_b 和类内散度矩阵 S_w 定义如下:

$$S_b=\frac{1}{n}\sum_{i=1}^{C}n_i(\mu_i-\mu)(\mu_i-\mu)^T \tag{2}$$

$$S_w=\frac{1}{n}\sum_{i=1}^{C}\sum_{x_j\in c_i}(x_j-\mu_i)(x_j-\mu_i)^T \tag{3}$$

其中,c_i 表示第 i 类样本集合,μ_i 表示第 i 类样本均值,μ 是所有样本均值.于是,LDA 算法可归结为下面的优化问题:

$$\max\ \frac{\sum_{i=1}^{C}n_i\|V^T\mu_i-V^T\mu\|^2}{\sum_{i=1}^{C}\sum_{x_j\in c_i}\|V^Tx_j-V^T\mu_i\|^2} \tag{4}$$

等价地,优化问题(4)可以改写为迹比优化形式:

$$\max\ \frac{tr(V^TS_bV)}{tr(V^TS_wV)} \tag{5}$$

由于目标函数(5)是非凸的,没有解析形式的全局最优解,通常将其近似转化为如下的比迹优化问题:

$$\max\ tr((V^TS_wV)^{-1}V^TS_bV) \tag{6}$$

当类内散度矩阵 S_w 非奇异时,由公式(6)得到的最优投影矩阵 V 由 $S_w^{-1}S_b$ 的 r 个最大特征值所对应的特征向量构成[30].

1.2 最大边界准则(MMC)

与 LDA 类似,MMC 的优化目标也是希望投影之后的低维空间中样本的类间散度最大,类内散度最小.不同之处在于,MMC 采用了迹差优化模型:

$$J(V)=tr(V^T(S_b-S_w)V^T) \tag{7}$$

限定 V 的每一列都是单位向量,于是,V 可以通过下面的特征分解来求得:

$$(S_b-S_w)v_i=\lambda_iv_i \tag{8}$$

1.3 边界Fisher分析(MFA)

训练样本数据之间的内在关系可以通过无向加权图 $G=\{G,W\}$来表示,其中,图的节点集合为训练样本数据 X,权重矩阵 W 表示数据节点之间的某种相似性度量.图 G 的 Laplacian 矩阵 L 定义为

$$L=D-W \tag{9}$$

这里,D 是一个对角矩阵,且 $D_{ii}=\sum_{j=1}^{n}W_{ij}$.

为了利用数据的局部判别信息,MFA 算法采用本征图 $G^+=\{G,W^+\}$和惩罚图 $G^-=\{G,W^-\}$表示数据之间的关系,其中,本征图刻画了属于同类的样本之间的近邻关系,而惩罚图刻画了属于不同类的样本之间的近邻关系.于是,权重矩阵可分别定义为

$$W_{ij}^{+}=\begin{cases}1, & i\in N_{k_1}^{+}(j)\vee j\in N_{k_1}^{+}(i)\\ 0, & \text{otherwise}\end{cases} \tag{10}$$

$$W_{ij}^{-}=\begin{cases}1, & i\in N_{k_2}^{-}(j)\vee j\in N_{k_2}^{-}(i)\\ 0, & \text{otherwise}\end{cases} \tag{11}$$

这里,$N_{k_1}^{+}(i)$ 表示与训练样本 x_i 同类的 k_1 个最近邻节点的下标集合,$N_{k_2}^{-}(i)$ 表示与训练样本 x_i 异类的 k_2 个最近邻节点的下标集合.

在投影之后的低维空间中,训练样本之间的类内紧密性和类间分离性可以分别用公式(12)和公式(13)表示:

$$\sum_{i}\sum_{i\in N_{k_1}^{+}(j)\vee j\in N_{k_1}^{+}(i)}\|V^Tx_i-V^Tx_j\|^2=\sum_{i,j}\|V^Tx_i-V^Tx_j\|^2W_{ij}^{+}=2tr(V^TX(D^{+}-W^{+})X^TV) \tag{12}$$

$$\sum_{i}\sum_{i\in N_{k_2}^{-}(j)\vee j\in N_{k_2}^{-}(i)}\|V^Tx_i-V^Tx_j\|^2=\sum_{i,j}\|V^Tx_i-V^Tx_j\|^2W_{ij}^{-}=2tr(V^TX(D^{-}-W^{-})X^TV) \tag{13}$$

为了使得样本的低维表示更有利于分类,类似于公式(4),MFA 最大化下面的目标函数:

$$J(V)=\frac{tr(V^TXL^{-}X^TV)}{tr(V^TXL^{+}X^TV)} \tag{14}$$

其中,$L^-=D^--W^-$和 $L^+=D^+-W^+$分别是惩罚图和本征图所对应的 Laplacian 矩阵.目标函数(14)的求解过程与目标函数(5)的求解过程类似.

2 边界判别投影

2.1 算法基本思想

为了增强原始样本低维表示的判别性能,我们希望经过投影之后,在低维子空间中,类间距离更大,类内距离更小.于是,不同类之间的样本边界更大,从而使得低维表示更加有利于分类.为此,需要重新定义类间距离、类内距离和边界等基本概念.

定义 1. 假设 x_i 与 x_j 为任意给定的两个样本,它们之间的距离定义为

$$d(x_i,x_j)=\|x_i-x_j\|_2.$$

定义 2. 假设 c_i 与 c_j 为任意给定的两个不同类别的样本集合,称其类间最近距离的样本点 x_j^i 与 x_i^j 为异类边界样本点,即:

$$\{x_j^i\in c_i,x_i^j\in c_j:d(x_j^i,x_i^j)\leqslant d(x_i,x_j),\forall x_i\in c_i,\forall x_j\in c_j\}.$$

定义 3. 对于任意给定的属于同一类的样本集合 c_i,称其类内最远距离的样本点 x_a^i 与 x_b^i 为同类边界样本点,即:

$$\{x_a^i,x_b^i\in c_i:d(x_a^i,x_b^i)\geqslant d(x_i,x_j),\forall x_i,x_j\in c_i\}.$$

注:异类边界样本点和同类边界样本点统称为边界样本点.

定义 4. 假设 c_i 与 c_j 为任意给定的两个不同类别的样本集合,其类间距离定义为

$$d(c_i,c_j)=d(x_j^i,x_i^j).$$

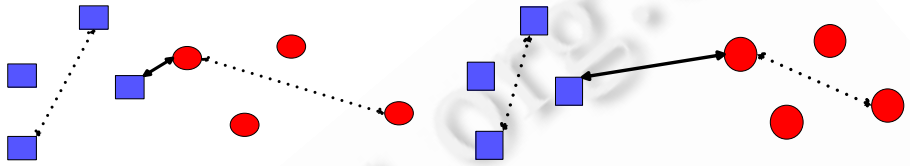
定义 5. 对于任意给定的某类样本集合 c_i,其类内距离定义为

$$d(c_i)=d(x_a^i,x_b^i).$$

定义 6. 假设样本数据 X 共有 C 类,即 $X=\{c_i\}_{i=1}^C$,其边界定义为

$$J=\sum_{i\neq j}d(c_i,c_j)-\sum_{i=1}^{C}d(c_i).$$

边界判别投影的基本思想是:将原始高维数据投影至低维空间,使其边界最大,即最大化类间距离,最小化类内距离.如图 1 所示,不同的几何形状代表不同的类,类间距离由实线表示,类内距离由虚线表示,通过实线或者虚线连接的样本点为边界样本点,经过变换之后,在保持数据的总体几何结构的前提下,边界得到了增强,样本的可分性更好.



(a) 变换前 (b) 变换后

Fig.1 Illustration of main idea behind our algorithm

图 1 MDP 算法基本思想

2.2 算法主要内容

2.2.1 数学模型

为了保持数据的整体几何结构,我们引入正交约束,即 $V^TV=I$,于是,边界判别投影的目标函数为

$$\max_{V^TV=I}\sum_{i\neq j}\delta(c_i,c_j)-\sum_{i=1}^{C}\delta(c_i) \tag{15}$$

其中,$\delta(c_i,c_j)$和$\delta(c_i)$分别为投影之后的低维数据的类间距离和类内距离,投影之后的低维数据之间的距离为

$$\delta(y_i,y_j)=\|V^Tx_i-V^Tx_j\|_2.$$

根据定义 4,有:

$$\delta(c_i,c_j)=\|V^Tx_j^i-V^Tx_i^j\|_2 \tag{16}$$

根据定义 5,有:

$$\delta(c_i)=\|V^Tx_a^i-V^Tx_b^i\|_2 \tag{17}$$

将公式(16)和公式(17)代入公式(15),则目标函数可改写为

$$\max_{V^TV=I}\sum_{i\neq j}\|V^Tx_j^i-V^Tx_i^j\|_2^2-\sum_{i=1}^{C}\|V^Tx_a^i-V^Tx_b^i\|_2^2 \tag{18}$$

显然,在目标函数(18)的计算中,用到的样本点个数至多为 $2C+2\binom{C}{2}=C(C+1)$. 当每类的样本个数大于 C+1 时,MDP 算法可以降低计算量.

为了获得 MDP 算法目标函数更为简洁的表示,我们将其归结到图嵌入框架下,为此引入类间相似性权重 $W^{(b)}$和类内相似性权重 $W^{(w)}$:

$$W_{ij}^{(b)}=\begin{cases}1, & \text{样本点为}x_j^i,x_i^j\\ 0, & \text{其他}\end{cases} \tag{19}$$

$$W_{ij}^{(w)}=\begin{cases}1, & \text{样本点为} x_a^c, x_b^c\\ 0, & \text{其他}\end{cases} \tag{20}$$

于是,目标函数(18)可表达为

$$\max_{V^TV=I}\sum_{i=1}^{n}\sum_{j=1}^{n}\|V^Tx_i-V^Tx_j\|^2W_{ij}^{(b)}-\sum_{i=1}^{n}\sum_{j=1}^{n}\|V^Tx_i-V^Tx_j\|^2W_{ij}^{(w)} \tag{21}$$

根据公式(12)、公式(13)的推导过程,优化问题(21)可以化简为

$$\max_{V^TV=I} tr(V^T(XL^{(b)}X^T-XL^{(w)}X^T)V) \tag{22}$$

其中,$L^{(b)}=D^{(b)}-W^{(b)}$,$L^{(w)}=D^{(w)}-W^{(w)}$,为 Laplacian 矩阵.

若令 $S^{(b)}=XL^{(b)}X^T$,$S^{(w)}=XL^{(w)}X^T$,则公式(22)又可表示为

$$\max_{V^TV=I} tr(V^T(S^{(b)}-S^{(w)})V) \tag{23}$$

从形式上来看,MDP 的目标函数与 MMC 相同.特别地,当训练集中每类只有 2 个样本时,MDP 等价于 MMC.

2.2.2 算法介绍

在人脸识别、文本分类等实际应用中,样本的维数 d 通常远远大于样本的个数 n,直接通过特征分解求解目标函数(23)的时间复杂度为 $O(d^2)$,空间复杂度为 $O(d^2)$,此时,计算耗时且可能导致数值不稳定.比如在人脸识别中,对于一张分辨率为 112×92 的人脸图像(d=10304),计算中时间代价和存储代价分别为 10^{12} 和 10^8,对于 PC 来说,代价过高.MDP 算法采用矩阵的 QR 分解技术[28,31]来避免这些问题,技术细节推导见第 2.3.1 节中的定理 2.

MDP 算法的步骤总结如下:

输入:数据矩阵 X,类别标号以及目标维数 r;

输出:低维表示 Y.

过程:

第 1 步:计算 Laplacian 矩阵 $L=L^{(b)}-L^{(w)}$.

第 2 步:QR 分解.采用不完全 Cholesky 分解技术[32],将数据矩阵 X 分解为 $X=QR$.

第 3 步:对 RLR^T 作特征分解,获得 $U=[u_1,u_2,\ldots,u_r]$,其中,u_i 是矩阵 RLR^T 的第 r 个最大特征值对应的特征向量.

第 4 步:计算投影矩阵 $V=QU$.

第 5 步:计算 $y_i=V^Tx_i$,获得低维表示 Y.特别地,对于训练数据 X,低维表示可以表示为

$$Y=V^TX=(QU)^TQR=U^TR \tag{24}$$

2.3 算法理论分析

2.3.1 相关证明

第 2.2.2 节中,MDP 算法第 3 步求得的投影矩阵的每一列是规范正交的,理论依据见下面的定理 1.

定理 1. 假设 L 是 $n\times n$ 对称矩阵,则存在规范正交矩阵 $V=[v_1,v_2,\ldots,v_d]\in R^{d\times d}$,使得:

$$XLX^Tv_i=\lambda_iv_i,1\leqslant i\leqslant d.$$

证明:下面采用数学归纳法证明此结论.令 k 为规范正交向量的个数,则:

当 k=1 时,显然成立.

假设当 $k=d-1$ 时结论成立.如果令 $XLX^Tv_1=\lambda_1v_1,v_1\in R^d$ 且$\|v_1\|=1$,对任意的 $u\in v_1^{\perp}$,有:

$$(XLX^Tu)^Tv_1=u^TXLX^Tv_1=u^T\lambda_1v_1=\lambda_1(u^Tv_1)=0.$$

因此, $XLX^Tu\in v_1^{\perp}$.这里, $\dim(v_1^{\perp})=d-1$.根据假设,当 $k=d-1$ 时,存在正交矩阵 $V'=[v_2,v_3,\ldots,v_d]\in R^{d\times(d-1)}$,使得:

$$XLX^Tv_i=\lambda_iv_i,2\leqslant i\leqslant d.$$

令 $V=v_1+V'\in R^{d\times d}$,故 $V=[v_1,v_2,\ldots,v_d]\in R^{d\times d}$ 为正交矩阵,且满足:

$$XLX^Tv_i=\lambda_iv_i,1\leqslant i\leqslant d.$$ □

定理 2. 假设 $X=QR$,其中,$Q\in R^{d\times t}$,$R\in R^{t\times n}$,$t=rank(X)$且 $Q^TQ=I$,L 是 $n\times n$ 对称矩阵,则 XLX^T 与 RLR^T 具有相同的特征值,且相应的特征向量具有下列关系:$V=QU$,其中,V 和 U 的列向量分别为 XLX^T 与 RLR^T 的特征向量.

证明:将 $X=QR$ 代入 XLX^T,得:

$$XLX^T=QRL(QR)^T=Q(RLR^T)Q^T \tag{25}$$

因为 V 和 U 分别由 XLX^T 与 RLR^T 的特征向量构成,即:

$$XLX^TV=V\Lambda_1, RLR^TU=U\Lambda_2 \tag{26}$$

又 XLX^T 与 RLR^T 是对称矩阵,则:

$$V^TV=I, U^TU=I.$$

于是,公式(26)可分别改写为

$$V^TXLX^TV=\Lambda_1 \tag{27}$$

$$U^TRLR^TU=\Lambda_2 \tag{28}$$

将公式(25)代入公式(27),得:

$$V^TQ(RLR^T)Q^TV=\Lambda_1 \tag{29}$$

对比公式(28)和公式(29),可知:

$$\Lambda_1=\Lambda_2 \text{ 且 } Q^TV=U,$$

即,$V=QU$. □

定理 3. MDP 可以保持数据流形的几何形状.

证明:对任意两个数据点的低维表示 y_i,y_j,其欧氏距离可以表示为

$$\| y_i-y_j \|_2=\sqrt{(y_i-y_j)^T(y_i-y_j)}=\sqrt{(V^Tx_i-V^Tx_j)^T(V^Tx_i-V^Tx_j)}=\sqrt{(x_i-x_j)^TVV^T(x_i-x_j)}.$$

由于 $V=(v_1,v_2,\ldots,v_r)$的列向量是规范正交的,则 $VV^T=\begin{pmatrix} I_r & O \\ O & O \end{pmatrix}_{d\times d}$,于是:

$$\| y_i-y_j \|_2=\sqrt{\sum_{k=1}^{r}(x_{ki}-x_{kj})^2}.$$

因此,低维表示之间的欧氏距离等价于原始样本在前 r 个坐标轴上做正交投影之后的欧氏距离,即,MDP 所得到的低维表示保持了原始样本的几何形状.

2.3.2 复杂度分析

第 1 步中,距离计算的时间复杂度为 $O(dn^2)$,寻找同类最远样本点和不同类最近样本点时可采用单轮的冒泡排序算法,其时间复杂度为 $O(n)$;

第 2 步中,计算 R 时,对 X 进行 QR 分解的时间复杂度为 $O(t^2n)$;

第 3 步中,对 RLR^T 进行特征分解的时间复杂度为 $O(t^3)$.

因此,MDP 算法的时间复杂度为 $O(t^3+t^2n+dn^2+n)$.

由于计算中矩阵的最大规模为 $d\times n$,因此,空间复杂度为 $O(dn)$.

3 实 验

图像数据是典型的高维数据,为了验证 MDP 算法的有效性,我们将其应用于人脸识别,并与 PCA,LDA,MFA,MMC 等经典算法进行比较.

3.1 数据集描述

实验中所选用的人脸数据集相关描述见表 1.为了便于数值处理,均采用灰度人脸图像,且经过裁剪和缩放至合适的大小,数据集中示例图像如图 2~图 7 所示.UMIST 数据集(http://www.sheffield.ac.uk/eee/research/iel/research/face)共有 20 个人的 564 张人脸图像,包括不同的种族、性别、外貌以及从侧面转向正面的不同姿态,具有典型的流形结构.YaleB 数据集(http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html)包含了 38 个人在不同的光照条件和面部表情下的人脸图像.FERET 数据集(http://www.datatang.com/data/44111)共有 14 051 张人脸图像,本文仅选取了该数据集的 200 个人共 1 400 张图像用于实验,每个人有 7 张图像,包含了表情、光照

和姿态等变化.GeorgiaTech 数据集(http://www.anefian.com/research/face_reco.htm)由 50 个人在不同的时间下拍摄,每个人有 15 张图像,包含不同的倾斜、表情和光照.Yale 数据集(http://cvc.yale.edu/projects/yalefaces/yalefaces.html)由耶鲁大学计算视觉与控制中心创建,每个人脸有 15 张图像,包括了光照(如正面、左侧和右侧)、是否戴眼镜、人脸表情(如正常、高兴、悲伤、困乏、惊讶和眨眼)等因素的变化.AR 数据集(http://rvl1.ecn.purdue.edu/~aleix/aleix_face_DB.html)是公认度比较高的一种数据库,实验中选取了 100 个人,每人 14 张图像,分别对应于不同表情和光照条件下的人脸.

Table 1 Data sets description

表 1 数据集说明

Name	Dimensionality	Number of samples	Number of classes
UMIST	56×46	575	20
YaleB	32×32	2 414	38
FERET	40×40	1 400	200
GeorgiaTech	50×36	750	50
Yale	32×32	165	15
AR	42×30	1 400	100

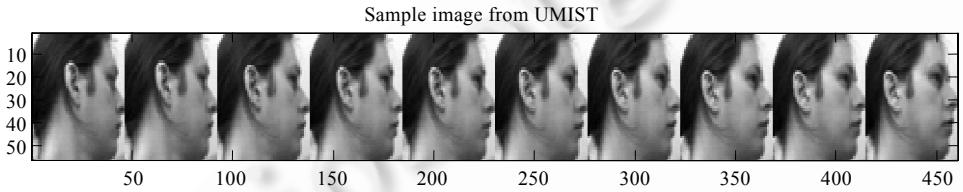


Fig.2 Sample image from UMIST data set

图 2 UMIST 数据集上的图像示例

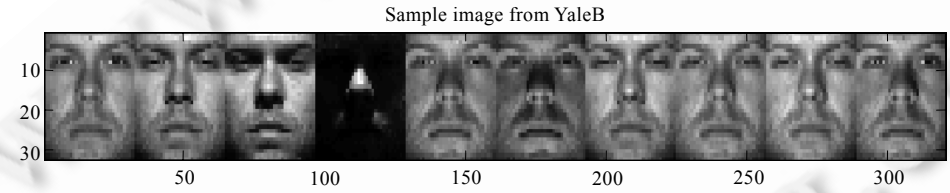


Fig.3 Sample image from YaleB data set

图 3 YaleB 数据集上的图像示例

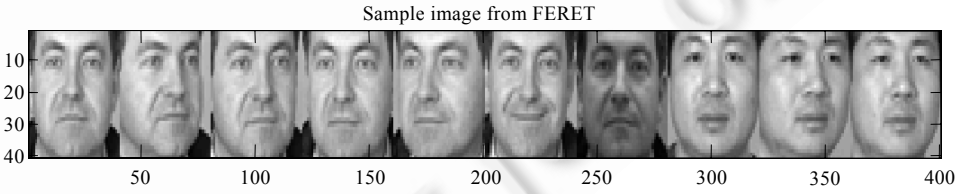


Fig.4 Sample image from FERET data set

图 4 FERET 数据集上的图像示例

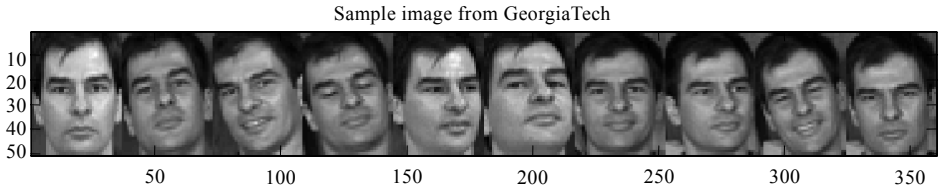


Fig.5 Sample image from GeorgiaTech data set

图 5 GeorgiaTech 数据集上的图像示例

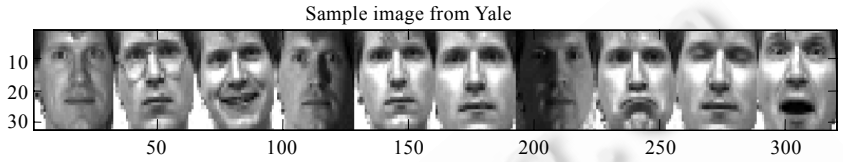


Fig.6 Sample image from Yale data set

图 6 Yale 数据集上的图像示例

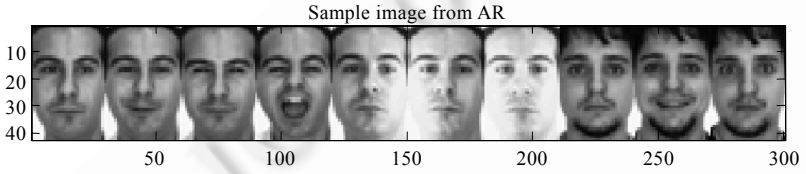


Fig.7 Sample image from AR data set

图 7 AR 数据集上的图像示例

3.2 实验参数设置

为了克服计算中的奇异性问题,其中,LDA 和 MFA 均采用 PCA 进行预处理,并设置主成分贡献率为 0.95.实验中所涉及的参数均由经验给出,MFA 算法 Gauss 热核参数 t 的设置参见局部判别嵌入[33]的源代码(http://www.cs.nthu.edu.tw/~htchen/),MFA 中的近邻参数 k 设置为 5,同类近邻参数 k_1 设置为 2,异类近邻参数 k_2 设置为 10.降维之后,统一采用 KNN 算法进行分类,并比较其分类的准确率(即识别率).每次实验中,在每个类别中随机选取 l(l=3,4,5)个样本作为训练集,剩余的样本作为测试集.重复进行 20 次随机划分,并计算其平均识别率和标准差(均采用百分比).

3.3 结果分析

表 2~表 7 分别报告了在不同人脸数据集选取不同数目的标记样本下的实验结果,包括最优平均识别率及其对应的标准差和目标维数.从中可以看出,除 Yale 数据集外,MDP 算法在其他人脸数据集上的识别率都高于 PCA,LDA,MFA 和 MMC.在 Yale 数据集上,当 l=3 时,MDP 和 MFA 最优识别率基本一致,高于其他算法,然而在 l=4 和 l=5 时,MDP 算法的识别率低于 MFA.与 MFA 相比,MDP 只考虑了数据的全局结构,而在某些复杂结构的数据集中,全局结构不足以较好地刻画不同类别数据之间的判别信息.图 8 显示了在 UMIST 数据集上,不同数目的标记样本情况下,随着目标维数的改变,这 5 种算法平均识别率的变化.随着训练样本个数的增加,识别率也在增加.MDP 算法不涉及参数的调整,而且在不同的数据集和不同训练样本情况下,均取得了最优的识别率.

PCA 预处理在克服小样本问题的同时,也降低了后续算法处理的计算量.为了比较各算法的时间损耗以及 PCA 预处理对 MDP 算法性能的影响,表 8 报告了 AR 人脸数据集(l=5,重复 20 次随机划分)上各算法在经过 PCA 预处理(主成分贡献率为 0.95)后的数据上的平均最优识别率及其对应的标准差(均采用百分比表示)和取得最优识别率的维数,以及各算法平均运行 1 次的时间消耗(时间单位以秒计).实验均采用 MATLAB 2010 编码实现,

计算设备为个人计算机,其中,处理器为英特尔双核 3.16GHz,内存为 4G,操作系统为 64 位 Windows 7.对比表 6 和表 7 的结果可以看出:MDP 算法的运行时间与 LDA 相当,但低于 MFA 和 MMC.同时,PCA 预处理会降低 MDP 算法的识别率[34].

另外,实验中注意到,MMC 算法并不稳定,在某些数据集上识别率很差.当样本数目小于样本维数(即小样本问题)时,基于迹差准则(即目标函数(7))的降维算法比基于迹比准则(即目标函数(5))的降维算法效果更好;当样本数目较多或者计算中的奇异性问题能够被有效避免时,基于迹比准则的降维算法比基于迹差准则的降维算法效果更好,因为迹比准则所生成的投影向量是相互统计无关的[35].

Table 2 Recognition results on UMIST data set

表 2 UMIST 数据集上的识别结果

Algorithm	3 labeled	4 labeled	5 labeled
PCA	70.6±3.1(49)	78.0±1.8(45)	82.5±3.4(48)
LDA	79.4±4.6(19)	85.6±2.7(19)	88.1±3.5(17)
MFA	82.6±3.1(46)	87.5±3.0(15)	89.6±3.2(21)
MMC	74.9±3.1(43)	81.6±3.9(56)	86.3±3.6(49)
MDP	**83.1±3.0(38)**	**89.0±2.4(14)**	**91.4±2.3(19)**

Table 3 Recognition results on YaleB data set

表 3 YableB 数据集上的识别结果

Algorithm	3 labeled	4 labeled	5 labeled
PCA	22.7±1.5(50)	26.2±1.3(50)	29.3±1.4(50)
LDA	53.4±2.0(37)	59.9±2.4(37)	65.5±1.6(37)
MFA	54.8±2.3(46)	60.4±2.2(49)	66.3±2.0(46)
MMC	22.3±1.4(60)	26.0±0.6(60)	29.3±1.6(60)
MDP	**61.4±2.3(60)**	**69.9±1.5(59)**	**75.1±1.5(59)**

Table 4 Recognition results on FERET data set

表 4 FERET 数据集上的识别结果

Algorithm	3 labeled	4 labeled	5 labeled
PCA	31.9±1.2(50)	36.3±1.6(50)	40.9±2.1(50)
LDA	37.4±1.8(50)	34.9±1.6(49)	31.5±1.4(50)
MFA	43.6±1.3(49)	48.2±2.0(50)	61.1±1.8(44)
MMC	32.6±2.3(13)	42.9±1.8(15)	50.6±2.6(16)
MDP	**80.6±1.4(36)**	**85.2±1.1(42)**	**88.0±0.9(36)**

Table 5 Recognition results on GeorgiaTech data set

表 5 GeorgiaTech 数据集上的识别结果

Algorithm	3 labeled	4 labeled	5 labeled
PCA	63.1±1.6(33)	68.0±1.7(38)	71.6±1.9(50)
LDA	48.6±2.4(47)	53.8±2.7(49)	54.9±1.7(49)
MFA	59.2±2.7(50)	60.7±2.1(50)	58.8±1.9(50)
MMC	65.6±1.5(20)	70.0±1.9(18)	73.2±2.0(19)
MDP	**69.3±1.8(49)**	**74.7±1.6(34)**	**78.2±1.5(46)**

Table 6 Recognition results on Yale data set

表 6 Yale 数据集上的识别结果

Algorithm	3 labeled	4 labeled	5 labeled
PCA	55.33±4.21(44)	55.76±3.11(28)	59.50±2.46(30)
LDA	62.17±5.00(14)	71.76±5.17(14)	76.50±3.89(14)
MFA	**66.04±4.00(16)**	**74.24±3.77(17)**	**77.06±3.25(17)**
MMC	60.38±2.91(16)	65.29±3.31(14)	67.17±3.74(20)
MDP	**66.00±4.03(26)**	72.10±4.27(21)	74.72±4.32(43)

Table 7 Recognition results on AR data set

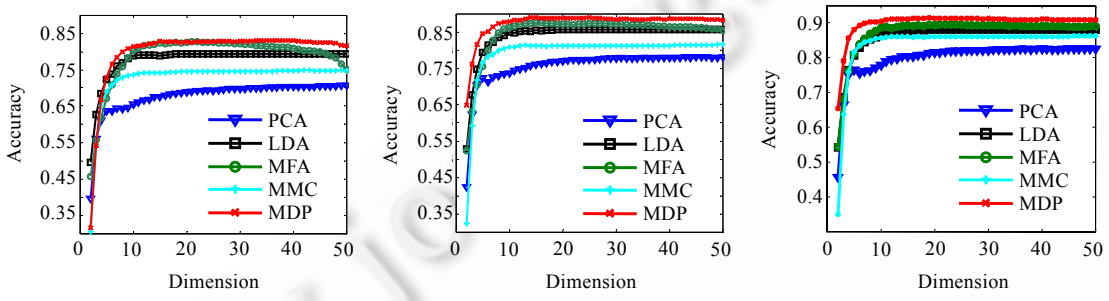
表 7 AR 数据集上的识别结果

Algorithm	3 labeled	4 labeled	5 labeled
PCA	43.2±1.7(50)	50.4±1.5(50)	55.8±1.6(50)
LDA	84.7±1.7(50)	88.0±1.3(50)	89.5±1.2(49)
MFA	85.4±1.6(50)	89.6±1.3(50)	92.1±1.1(50)
MMC	45.8±1.3(60)	53.0±1.4(60)	58.6±1.0(60)
MDP	**91.9±1.1(59)**	**95.2±0.9(58)**	**96.7±0.6(59)**

Table 8 Performance comparison on AR data set (l=5)

表 8 AR 数据集(l=5)上的性能比较

Algorithm	Accuracy (%)	Stand deviation (%)	Dimensionality	Time (s)
LDA	90.02	1.23	50	0.41
MFA	92.49	0.86	50	1.32
MMC	57.55	1.17	42	1.36
MDP	**93.48**	**1.13**	**37**	**0.44**



(a) 3 labeled (b) 4 labeled (c) 5 labeled

Fig.8 Comparisons of recognition results on UMIST data set

图 8 UMIST 数据集上的识别结果比较

4 结束语

作为一种新的有监督线性降维方法,MDP 通过最大化异类样本之间的最小距离,同时最小化同类样本之间的最大距离,在保持样本全局几何结构的基础上,增强数据的判别性能.与经典的基于边界思想的降维算法相比,MDP 无需额外的参数设置,而且求得的投影矩阵的列向量是规范正交的,保持了原始数据流形的全局几何结构.从计算的角度来讲,MDP 避免了小样本问题,并采用 QR 分解技术建立了高效而稳定的算法.人脸数据集上的实验结果表明了该算法的有效性.

致谢 在此,我们向对本文的工作给予支持和建议的同行,尤其是武汉大学软件工程国家重点实验室丁立新教授领导的讨论班上的同学和老师表示感谢.

References:

[1] Li GJ. The scientific value of big data. Research Communications of The CCF, 2012,8(9):8−15 (in Chinese with English abstract).

[2] Tenenbaum JB, De Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science, 2000, 290(5500):2319−2323. [doi: 10.1126/science.290.5500.2319]

[3] Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000,290(5500):2323−2326. [doi: 10.1126/science.290.5500.2323]

[4] Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Proc. of the NIPS, Vol.14. 2001. 585−591. http://machinelearning.wustl.edu/mlpapers/paper_files/nips02-AA42.pdf

[5] Zhang Z, Zha H. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. Journal of Shanghai University (English Edition), 2004,8(4):406−424. [doi: 10.1007/s11741-004-0051-1]

[6] Donoho DL, Grimes C. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. Proc. of the National Academy of Sciences, 2003,100(10):5591−5596. [doi: 10.1073/pnas.1031596100]

[7] Weinberger KQ, Sha F, Saul LK. Learning a kernel matrix for nonlinear dimensionality reduction. In: Proc. of the 21st Int'l Conf. on Machine Learning (ICML 2004). 2004. 839−846. [doi: 10.1145/1015330.1015345]

[8] He XF, Yan SC, Hu YX, Niyogi P, Zhang HJ. Face recognition using laplacianfaces. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005, 27(3):328−340. [doi: 10.1109/TPAMI.2005.55]

[9] He XF, Ma WY, Zhang HJ. Learning an image manifold for retrieval. In: Proc. of the 12th Annual ACM Int'l Conf. on Multimedia. ACM Press, 2004. 17−23. [doi: 10.1145/1027527.1027532]

[10] Cai D, He XF. Manifold adaptive experimental design for text categorization. IEEE Trans. on Knowledge and Data Engineering, 2012,24(4):707−719. [doi: 10.1109/TKDE.2011.104]

[11] Tang JH, Hua XS, Qi GJ, Wang M, Mei T, Wu XQ. Structure-Sensitive manifold ranking for video concept detection. In: Proc. of the ACM Conf. on Multimedia. New York: ACM Press, 2007. 852−861. [doi: 10.1145/1291233.1291430]

[12] Hoi SCH, Lyu MR. A multimodal and multilevel ranking scheme for large-scale video retrieval. IEEE Trans. on Multimedia, 2008, 10(4):607−619. [doi: 10.1109/TMM.2008.921735]

[13] Bengio Y, Paiement JF, Vincent P. Out-of-Sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In: Advances in Neural Information Processing Systems. MIT Press, 2003. 177−184. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.5.1709

[14] Vlassis N, Motomura Y, Krose B. Supervised dimension reduction of intrinsically low dimensional data. Neural Computation, 2002, 14(1):191−215. [doi: 10.1162/089976602753284491]

[15] Jolliffe IT. Principal Component Analysis. New York: Springer-Verlag, 1986.

[16] Fisher RA. The statistical utilization of multiple measurements. Annals of Eugenics, 1938,8(4):376−386. [doi: 10.1111/j.1469-1809.1938.tb02189.x]

[17] Rao CR. The utilization of multiple measurements in problems of biological classification. Journal of the Royal Statistical Society-Series B: Statistical Methodology, 1948,10(2):159−203.

[18] Belhumeur PN, Hespanda J, Kiregeman D. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Trans. on PAMI, 1997. [doi: 10.1109/34.598228]

[19] Li HF, Jiang T, Zhang KS. Efficient and robust feature extraction by maximum margin criterion. IEEE Trans. on Neural Networks, 2006,17(1):157−165. [doi: 10.1109/TNN.2005.860852]

[20] Zheng WM, Zou CR, Zhao L. Weighted maximum margin discriminant analysis with kernels. Neurocomputing, 2005,67:357−362. [doi: 10.1016/j.neucom.2004.12.008]

[21] Liu QS, Tang XO, Lu HQ, Ma SD. Face recognition using kernel scatter-difference-based discriminant analysis. IEEE Trans. on Neural Networks, 2006,17(4):1081−1085. [doi: 10.1109/TNN.2006.875970]

[22] Liu J, Chen SC, Tan XY, Zhang DQ. Comments on "efficient and robust feature extraction by maximum margin criterion". IEEE Trans. on Neural Networks, 2007,18(6):1862−1864. [doi: 10.1109/TNN.2007.900813]

[23] Qiu XP, Wu LD. Face recognition by stepwise nonparametric margin maximum criterion. In: Proc. of the 10th IEEE Int'l Conf. on Computer Vision (ICCV 2005), Vol.2. IEEE, 2005. 1567−1572. [doi: 10.1109/ICCV.2005.91]

[24] Yan SC, Xu D, Zhang BY, Zhang HJ. Graph embedding: A general framework for dimensionality reduction. In: Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR 2005), Vol.2. IEEE, 2005. 830−837. [doi: 10.1109/CVPR.2005.170]

[25] Yan SC, Xu D, Zhang BY, Zhang HJ, Yang Q. Graph embedding and extensions: A general framework for dimensionality reduction. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2007,29(1):40−51. [doi: 10.1109/TPAMI.2007.250598]

[26] Weinberger KQ, Blitzer J, Saul LK. Distance metric learning for large margin nearest neighbor classification. In: Advances in Neural Information Processing Systems, Vol.18. MIT Press, 2006. 1473. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.117.5831

[27] Wang F, Zhang CS. Feature extraction by maximizing the average neighborhood margin. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2007). IEEE, 2007. 1−8. [doi: 10.1109/CVPR.2007.383124]

[28] Wang HX, Zheng WM, Hu ZL, Chen SB. Local and weighted maximum margin discriminant analysis. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2007). IEEE, 2007. 1−8. [doi: 10.1109/CVPR.2007.383039]

[29] He XF, Cai D, Han JW. Learning a maximum margin subspace for image retrieval. IEEE Trans. on Knowledge and Data Engineering, 2008,20(2):189−201. [doi: 10.1109/TKDE.2007.190692]

[30] Fukunaga K. Introduction to Statistical Pattern Recognition. 2nd ed., Boston: Academic Press, 1990.

[31] Ye JP, Li Q, Xiong H, Park H, Janardan R, Kumar V. Idr/qr: An incremental dimension reduction algorithm via qr decomposition. In: Proc. of the 10th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD 2004). New York, 2004. 364−373. [doi: 10.1145/1014052.1014093]

[32] Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern Analysis. Cambridge University Press, 2004.

[33] Chen HT, Chang HW, Liu TL. Local discriminant embedding and its variants. In: Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, Vol.2. IEEE, 2005. 846−853. [doi: 10.1109/CVPR.2005.216]

[34] Zhang TP, Fang B, Tang YY, Shang ZW, Xu B. Generalized discriminant analysis: A matrix exponential approach. IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics, 2010,40(1):186−197. [doi: 10.1109/TSMCB.2009.2024759]

[35] Tao Y, Yang J. Quotient vs. difference: Comparison between the two discriminant criteria. Neurocomputing, 2010,73(10): 1808−1817. [doi: 10.1016/j.neucom.2009.10.026]

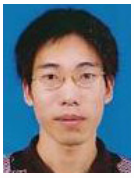
附中文参考文献:

[1] 李国杰.大数据研究的科学价值.中国计算机学会通讯,2012,8(9):8−15.

何进荣(1984－),男,甘肃民勤人,博士生,主要研究领域为机器学习,数据降维,特征提取.
E-mail: hejinrong@whu.edu.cn

李照奎(1976－),男,博士生,主要研究领域为机器学习,人脸识别.
E-mail: lmy52wy@163.com

丁立新(1967－),男,博士,教授,博士生导师,CCF 会员,主要研究领域为智能信息处理,云计算.
E-mail: lxding@whu.edu.cn



胡庆辉(1976－),男,博士生,主要研究领域为机器学习,演化计算.
E-mail: Huqinghui2004@126.com