ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.11, November 2008, pp.2833−2842
DOI: 10.3724/SP.J.1001.2008.02833

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

基于局部与全局保持的半监督维数约减方法*

韦 佳+, 彭 宏

(华南理工大学 计算机科学与工程学院,广东 广州 510641)

Local and Global Preserving Based Semi-Supervised Dimensionality Reduction Method

WEI Jia+, PENG Hong

(School of Computer Science and Engineering, South China University of Technology, Guangzhou 510641, China)

+ Corresponding author: E-mail: wei.jia@mail.scut.edu.cn

Wei J, Peng H. Local and global preserving based semi-supervised dimensionality reduction method. *Journal of Software*, 2008,19(11):2833−2842. http://www.jos.org.cn/1000-9825/19/2833.htm

Abstract: In many machine learning and data mining tasks, it can't achieve the best semi-supervised learning result if only use side-information. So, a local and global preserving based semi-supervised dimensionality reduction (LGSSDR) method is proposed in this paper. LGSSDR algorithm can not only preserve the positive and negative constraints but also preserve the local and global structure of the whole data manifold in the low dimensional embedding subspace. Besides, the algorithm can compute the transformation matrix and handle unseen samples easily. Experimental results on several datasets demonstrate the effectiveness of this method.

Key words: side-information; local and global preserving; semi-supervised learning; dimensionality reduction; graph embedding

摘 要: 在很多机器学习和数据挖掘任务中,仅仅利用边信息(side-information)并不能得到最好的半监督学习(semi-supervised learning)效果,因此,提出一种基于局部与全局保持的半监督维数约减(local and global preserving based semi-supervised dimensionality reduction,简称 LGSSDR)方法.该算法不仅能够保持正、负约束信息而且能够保持数据集所在低维流形的全局以及局部信息.另外,该算法能够计算出变换矩阵并较容易地处理未见样本.实验结果验证了该算法的有效性.

关键词: 边信息;局部与全局保持;半监督学习;维数约减;图嵌入

中图法分类号: TP181 **文献标识码**: A

1 半监督学习介绍

许多机器学习及数据挖掘算法的性能取决于对输入空间特征提取的有效性,在诸如人脸识别以及图像检索等应用中,由于所处理数据具有高维数的特点,如果不对其进行有效的特征提取,则很容易出现所谓的"维数

* Supported by the Natural Science Foundation of Guangdong Province of China under Grant No.07006474 (广东省自然科学基金); the Sci & Tech Research Project of Guangdong Province of China under Grant No.2007B010200044 (广东省科技攻关项目)

Received 2008-02-24; Accepted 2008-08-26

灾难”问题[1].因此,维数约减的有效性是许多机器学习及数据挖掘算法成败与否的关键.最著名的维数约减方法是人们所熟知的主成分分析(principal component analysis,简称 PCA)[2,3]以及线性判别分析(linear discriminant analysis,简称 LDA)[3],这两种算法分别是无监督式(unsupervised)以及监督式(supervised)算法.但是在许多学习任务中,人们常常面对着的是大量的未标记数据以及相对来说少得可怜的有标记数据.如果只使用少量的标记样本,那么所训练出的学习系统很难具有良好的泛化能力;另一方面,如果只使用未标记样本,则浪费了标记样本中所提供的有用信息.因此,在这种情况下,无论是监督式算法还是无监督式算法都不能获得令人满意的结果.近年来,如何从标记数据以及未标记数据中学习出有用的知识来改善学习性能吸引了越来越多研究者的关注[4].这种学习方式称为“半监督学习”.然而,在许多情况下,人们往往不能明确得知某一样本的具体类别标签,所知的是某两个样本是否属于同一类别(类别标签未知)的成对约束信息.这种成对约束信息称为边信息(side-information)[5].边信息包括两种,一种是正约束(positive constraint 或称 must-link constraint),另一种是负约束(negative constraint 或称 cannot-link constraint).正约束表示两个样本属于同一类,但并不知道其确切的类别标签;负约束表示两个样本不属于同一类别.边信息是一种比标签信息更一般的信息,因为边信息可以从标签信息中得到,反之则不可以[6].本文的研究重点就是基于边信息的半监督线性维数约减.

近年来,利用边信息进行线性特征提取受到了越来越多的关注.Shental 等人提出一种相关成分分析算法(RCA)[7],但该算法只能利用正约束信息并且忽略了隐藏在大量未标记数据中的潜在信息.Bar-Hiller 等人提出一种约束 Fisher 线性判别分析算法(cFLD)[8],但该算法存在着与 RCA 同样的问题.Xing 等人[9],Tang 等人[10]以及 Yeung 等人[11]提出的算法不仅能够利用正约束信息而且能够利用负约束信息,但是没有利用隐藏在未标记数据中的潜在信息.Wu 等人提出一种迭代自增强相关成分分析算法(ISERCA)[12],该算法能够同时利用边信息以及未标记数据中的潜在信息,但它的缺点是耗时并且可能陷入局部极小.Zhang 等人提出一种半监督维数约减算法(SSDR)[13],该算法不仅能够保持成对约束的结构并且能够保持未标记数据所在低维流形的结构,其缺点是只考虑了低维流形的全局协方差结构而没有考虑其局部结构.基于上述分析,本文提出一种基于局部与全局保持的半监督维数约减(local and global preserving based semi-supervised dimensionality reduction,简称 LGSSDR)方法.该算法能够很好地利用成对约束信息以及未标记数据中的潜在信息,并且考虑到了给定数据集的流形结构,能够保持数据集的局部以及全局结构.另外,由于该算法能够计算出变换矩阵,可以非常方便地对未见样本进行处理.我们将在后面的实验中验证该算法的有效性.

2 LGSSDR算法

基于边信息的半监督线性维数约减问题的基本设置如下所示:给定样本集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_n\}\subset\boldsymbol{R}^D$ 以及成对约束 M 和 C,其中 M 为正约束,C 为负约束,也就是说,如果 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 属于同一类,那么 $(\boldsymbol{x}_i,\boldsymbol{x}_j)\in M$;如果 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 分别属于两个不同的类,那么 $(\boldsymbol{x}_i,\boldsymbol{x}_j)\in C$.所期望的结果是从上述给定的条件中学习到线性变换矩阵 $\boldsymbol{W}=[\boldsymbol{w}_1,\boldsymbol{w}_2,...,\boldsymbol{w}_d]\in\boldsymbol{R}^{D\times d}(d\ll D)$,使得原数据经由变换矩阵所得的低维投影为 $\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,...,\boldsymbol{y}_n\}\subset\boldsymbol{R}^d$(其中 $\boldsymbol{y}_i=\boldsymbol{W}^T\boldsymbol{x}_i$),该低维投影不仅能够保持 M 和 C 中的边信息,即 M 中的点对在低维投影中相互靠近而 C 中的点对在低维投影中相互远离,而且能够保持原始数据集的内在低维流形结构.

为了方便讨论,我们仅考虑目标维数为 1 的情况(即 d=1,则变换向量为 $\boldsymbol{w}\in\boldsymbol{R}^D$),其他维数下的情况可以很容易地推广得知.对于正约束 M,从中可以得知刻画类内紧凑程度的标量值 Q_m,定义如下:

$$\begin{aligned}Q_m&=\sum_{\substack{ij\\ \text{where }(\boldsymbol{x}_i,\boldsymbol{x}_j)\in M\text{ or }(\boldsymbol{x}_j,\boldsymbol{x}_i)\in M}}(\boldsymbol{w}^T\boldsymbol{x}_i-\boldsymbol{w}^T\boldsymbol{x}_j)^2\\&=2\sum_i(\boldsymbol{w}^T\boldsymbol{x}_iD_{ii}^m\boldsymbol{x}_i^T\boldsymbol{w})-2\sum_{ij}(\boldsymbol{w}^T\boldsymbol{x}_iS_{ij}^m\boldsymbol{x}_j^T\boldsymbol{w})\\&=2\boldsymbol{w}^T\boldsymbol{X}(\boldsymbol{D}^m-\boldsymbol{S}^m)\boldsymbol{X}^T\boldsymbol{w}=2\boldsymbol{w}^T\boldsymbol{X}\boldsymbol{L}^m\boldsymbol{X}^T\boldsymbol{w}\end{aligned}\tag{1}$$

$$S_{ij}^m=\begin{cases}1, & \text{if }(\boldsymbol{x}_i,\boldsymbol{x}_j)\in M\text{ or }(\boldsymbol{x}_j,\boldsymbol{x}_i)\in M\\0, & \text{else}\end{cases}\tag{2}$$

其中,$\boldsymbol{D}^m$ 为对角矩阵,其对角线上的元素是矩阵 $\boldsymbol{S}^m$ 中相应的列和(或行和,因为 $\boldsymbol{S}^m$ 为对称矩阵),即 $D_{ii}^m=\sum_j S_{ij}^m$. $\boldsymbol{L}^m=\boldsymbol{D}^m-\boldsymbol{S}^m$ 称为 Laplacian 矩阵[14,15],是一个对称的半正定矩阵.

另一方面,通过负约束 C,可以得到描述类间离散程度的标量 Q_c,定义如下:

$$Q_c=\sum_{\substack{ij\\ \text{where } (\boldsymbol{x}_i,\boldsymbol{x}_j)\in C \text{ or } (\boldsymbol{x}_j,\boldsymbol{x}_i)\in C}}(\boldsymbol{w}^T\boldsymbol{x}_i-\boldsymbol{w}^T\boldsymbol{x}_j)^2=2\boldsymbol{w}^T\boldsymbol{X}(\boldsymbol{D}^c-\boldsymbol{S}^c)\boldsymbol{X}^T\boldsymbol{w}=2\boldsymbol{w}^T\boldsymbol{X}\boldsymbol{L}^c\boldsymbol{X}^T\boldsymbol{w} \tag{3}$$

$$S_{ij}^c=\begin{cases}1, & \text{if } (x_i,x_j)\in C \text{ or } (x_j,x_i)\in C\\ 0, & \text{else}\end{cases} \tag{4}$$

其中,$\boldsymbol{D}^c$ 为对角矩阵, $D_{ii}^c=\sum_j S_{ij}^c$, $\boldsymbol{L}^c=\boldsymbol{D}^c-\boldsymbol{S}^c$. 因此,目标变换向量可以定义如下:

$$\boldsymbol{w}^*=\arg\max_{\boldsymbol{w}}\frac{Q_c}{Q_m}=\arg\max_{\boldsymbol{w}}\frac{\boldsymbol{w}^T\boldsymbol{X}\boldsymbol{L}^c\boldsymbol{X}^T\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{X}\boldsymbol{L}^m\boldsymbol{X}^T\boldsymbol{w}} \tag{5}$$

公式(5)的意义在于,使得不属于同一类的数据离得越远越好,而属于同一类的数据靠得越近越好.但是,这仅仅考虑了边信息,而没有利用未标记数据中的潜在信息.为了能够利用大量未标记数据中的潜在信息,我们作出如下假设(邻域假设或流形假设):高维空间中互相靠近的点在低维投影空间中也是互相靠近的.举例来说,如果两点 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 距离很近,那么它们的低维投影点 $\boldsymbol{y}_i$ 与 $\boldsymbol{y}_j$ 之间的距离也应该很短.基于此,我们可以利用 k -最近邻图来表示这种邻接关系[16].具体来说,如果点 $\boldsymbol{x}_i$ 是离点 $\boldsymbol{x}_j$ 最近的 k 个点中的一个,就用一条边把节点 i 和节点 j 连接起来,这样形成的图就是 k-最近邻图,记为 G.有了上述定义,可以得到相应的权矩阵如下:

$$S_{ij}^n=\begin{cases}1, & \text{if } \boldsymbol{x}_i\in N_k(\boldsymbol{x}_j) \text{ or } \boldsymbol{x}_j\in N_k(\boldsymbol{x}_i)\\ 0, & \text{else}\end{cases} \tag{6}$$

其中,$N_k(\boldsymbol{x}_j)$表示点 $\boldsymbol{x}_j$ 的 k-最近邻点的集合.这样,我们可以定义如下项来刻画邻近点之间的紧密程度:

$$Q_n=\sum_{ij}(\boldsymbol{w}^T\boldsymbol{x}_i-\boldsymbol{w}^T\boldsymbol{x}_j)^2S_{ij}^n=2\boldsymbol{w}^T\boldsymbol{X}(\boldsymbol{D}^n-\boldsymbol{S}^n)\boldsymbol{X}^T\boldsymbol{w}=2\boldsymbol{w}^T\boldsymbol{X}\boldsymbol{L}^n\boldsymbol{X}^T\boldsymbol{w} \tag{7}$$

其中,$\boldsymbol{D}^n$ 为对角矩阵,$D_{ii}^n=\sum_j S_{ij}^n$, $\boldsymbol{L}^n=\boldsymbol{D}^n-\boldsymbol{S}^n$.

另一方面,我们希望非邻近点的低维投影能够尽量散开(非邻域假设).这种性质可以用下式来度量:

$$Q_f=\sum_{ij}(\boldsymbol{w}^T\boldsymbol{x}_i-\boldsymbol{w}^T\boldsymbol{x}_j)^2S_{ij}^f=2\boldsymbol{w}^T\boldsymbol{X}(\boldsymbol{D}^f-\boldsymbol{S}^f)\boldsymbol{X}^T\boldsymbol{w}=2\boldsymbol{w}^T\boldsymbol{X}\boldsymbol{L}^f\boldsymbol{X}^T\boldsymbol{w} \tag{8}$$

$$S_{ij}^f=\begin{cases}1, & \text{if } \boldsymbol{x}_i\notin N_k(\boldsymbol{x}_j) \text{ and } \boldsymbol{x}_j\notin N_k(\boldsymbol{x}_i)\\ 0, & \text{else}\end{cases} \tag{9}$$

其中,$\boldsymbol{D}^f$ 为对角矩阵,$D_{ii}^f=\sum_j S_{ij}^f$, $\boldsymbol{L}^f=\boldsymbol{D}^f-\boldsymbol{S}^f$.

有了上述准备,LGSSDR 的目标向量可以定义为

$$\boldsymbol{w}^*=\arg\max_{\boldsymbol{w}}\frac{Q_c+\alpha Q_f}{Q_m+\beta Q_n}=\arg\max_{\boldsymbol{w}}\frac{\boldsymbol{w}^T\boldsymbol{X}(L^c+\alpha\boldsymbol{L}^f)\boldsymbol{X}^T\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{X}(L^m+\beta\boldsymbol{L}^n)\boldsymbol{X}^T\boldsymbol{w}} \tag{10}$$

其中,参数α和β分别用来调节 Q_f 和 Q_n 的贡献度.

可以看出,式(10)为一个广义瑞利商问题,如果 $\boldsymbol{X}(\boldsymbol{L}^m+\beta\boldsymbol{L}^n)\boldsymbol{X}^T$ 是非奇异的,那么它的解为下式最大广义特征值所对应的特征向量:

$$\boldsymbol{X}(\boldsymbol{L}^c+\alpha\boldsymbol{L}^f)\boldsymbol{X}^T\boldsymbol{w}=\lambda\boldsymbol{X}(\boldsymbol{L}^m+\beta\boldsymbol{L}^n)\boldsymbol{X}^T\boldsymbol{w} \tag{11}$$

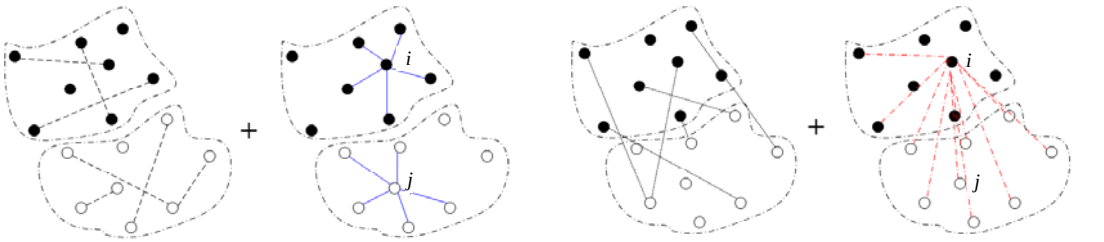
易知,如果 d>1,那么取前 d 个最大非零广义特征值所对应的特征向量即可组成变换矩阵 $\boldsymbol{W}$.

另外,当样本数较少而样本维数较大时, $\boldsymbol{X}(\boldsymbol{L}^m+\beta\boldsymbol{L}^n)\boldsymbol{X}^T$ 也可能是奇异的,即使是非奇异的,对如此高维的矩阵进行计算也是费时、费力的,这时可以先用 PCA 算法把原始样本降到一个低维子空间,然后再用 LGSSDR 算法在 PCA 的低维子空间中进行计算获得最终的结果,即 PCA+LGSSDR.这种方法与 Fisherface[17]中使用的 PCA+LDA 方法是类似的.

3 LGSSDR算法的图嵌入解释

在图嵌入理论中[18],无论是监督式的还是无监督式的维数约减算法都可以认为是描述数据集某种统计性质或几何性质的图的图嵌入过程.本节通过分析说明,图嵌入理论不仅适用于监督式和无监督式算法,而且适用于半监督算法.另外,如果没有约束信息,那么 LGSSDR 算法与 UDP[19]算法则是等价的.下面我们将利用图嵌入理论来解释 LGSSDR 算法.

根据图嵌入理论,在 LGSSDR 算法中定义一个内在图(intrinsic graph).该图与正约束以及邻域假设相对应,它刻画了类内数据的紧凑性.另外,再定义一个惩罚图(penalty graph).该图与负约束以及非邻域假设相对应,刻画了类间数据的分散性.图 1(a)所示的内在图相当于正约束图与 k-最近邻图之和,其中正约束图表示的是属于同一类的样本点对;k-最近邻图表示的是每一样本点与其 k 个最近点的连接关系.图 1(b)所示的惩罚图相当于负约束图与非 k-最近邻图之和,其中负约束图显示的是不属于同一类的样本点对;非 k-最近邻图显示的是每一样本点与不属于其 k 个最近点的样本点的连接关系(这里只画出了点"i"的非 k-最近邻图.易知,非 k-最近邻图是 k-最近邻图的补图).



(a) Intrinsic graph (a) 内在图 (b) Penalty graph (b) 惩罚图

Fig.1 Intrinsic graph and penalty graph

图 1 内在图与惩罚图

通过上述解释,可以总结 LGSSDR 步骤如下:

输入:样本集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_n\}\subset\boldsymbol{R}^D$, 正约束 M,负约束 C;

输出:变换矩阵 $\boldsymbol{W}\in\boldsymbol{R}^{D\times d}(d\ll D)$.

第 1 步,构造如图 1 所示的内在图及惩罚图.这些图的邻接矩阵参见式(2)、式(4)、式(6)和式(9).

第 2 步,确定权值 α 和 β,构造维数约减的优化准则.最优线性变换向量 $\boldsymbol{w}^*$如式(10)所示.

第 3 步,计算线性变换矩阵.令线性变换矩阵 $\boldsymbol{W}=[\boldsymbol{w}_1,\boldsymbol{w}_2,...,\boldsymbol{w}_d](d\ll D)$, 则 $\boldsymbol{w}_1,\boldsymbol{w}_2,...,\boldsymbol{w}_d$ 就是式(11)的最大的 d 个广义特征值所对应的广义特征向量.

求出变换矩阵 $\boldsymbol{W}$ 后,对于一个新来样本 $\boldsymbol{x}$ 可以直接得到其低维投影 $\boldsymbol{y}$,即 $\boldsymbol{x}\to\boldsymbol{y}=\boldsymbol{W}^T\boldsymbol{x}$. 另外,如果输入样本集的维数过高(如人脸图像),那么可以在不损失大部分信息的情况下(如保留其中 98%的主成分),通过 PCA 算法把输入数据投影到一个较低的维数,然后再利用 LGSSDR 算法处理.

4 KLGSSDR算法

许多线性维数约减算法都有其相对应的核版本,如与 PCA 算法相对应的是 KPCA(kernel principal component analysis)[20],与 LDA 算法相对应的是 KFD(kernel fisher discriminant analysis)[21]等等.与其他算法相似,LGSSDR 算法也可以推导出其相应的核版本,我们称其为 KLGSSDR.本节将给出其推导过程,该方法与 Cai[15]等人及 He[22]等人给出的方法类似.

设一个欧氏空间 $\boldsymbol{R}^D$ 经由一个非线性映射 $\phi:\boldsymbol{R}^D\mapsto\mathcal{H}$被映射到一个再生核 Hilbert 空间(reproducing kernel

Hilbert space,简称 RKHS)[23].令 $\kappa(\boldsymbol{x},\boldsymbol{y})=(\phi(\boldsymbol{x})\cdot\phi(\boldsymbol{y}))=\phi^T(\boldsymbol{x})\phi(\boldsymbol{y})$, 其中 $\kappa(\cdot,\cdot)$ 为一个半正定核函数,如多项式核、高斯核等.又令 $\phi(\boldsymbol{X})$ 表示 RKHS 中的数据矩阵,即 $\phi(\boldsymbol{X})=[\phi(\boldsymbol{x}_1),\phi(\boldsymbol{x}_2),...,\phi(\boldsymbol{x}_n)]$. 于是,原问题可以转化为 RKHS 中的特征向量求解问题:

$$\phi(\boldsymbol{X})(\boldsymbol{L}^c+\alpha\boldsymbol{L}^f)\phi^T(\boldsymbol{X})\boldsymbol{w}=\lambda\phi(\boldsymbol{X})(\boldsymbol{L}^m+\beta\boldsymbol{L}^n)\phi^T(\boldsymbol{X})\boldsymbol{w} \tag{12}$$

由于式(12)的特征向量 $\boldsymbol{w}$ 是 $\phi(\boldsymbol{x}_1),\phi(\boldsymbol{x}_2),...,\phi(\boldsymbol{x}_n)$ 的线性组合,那么必然存在系数 $\alpha_i,i=1,2,...,n$,使得 $\boldsymbol{w}=\sum_{i=1}^{n}\alpha_i\phi(\boldsymbol{x}_i)=\phi(\boldsymbol{X})\boldsymbol{\alpha}$, 其中 $\boldsymbol{\alpha}=[\alpha_1,\alpha_2,...,\alpha_n]^T\in\boldsymbol{R}^n$. 因此可得:

$$\begin{aligned}&\phi(\boldsymbol{X})(\boldsymbol{L}^c+\alpha\boldsymbol{L}^f)\phi^T(\boldsymbol{X})\phi(\boldsymbol{X})\boldsymbol{\alpha}=\lambda\phi(\boldsymbol{X})(\boldsymbol{L}^m+\beta\boldsymbol{L}^n)\phi^T(\boldsymbol{X})\phi(\boldsymbol{X})\boldsymbol{\alpha}\\&\Rightarrow\phi^T(\boldsymbol{X})\phi(\boldsymbol{X})(\boldsymbol{L}^c+\alpha\boldsymbol{L}^f)\phi^T(\boldsymbol{X})\phi(\boldsymbol{X})\boldsymbol{\alpha}=\lambda\phi^T(\boldsymbol{X})\phi(\boldsymbol{X})(\boldsymbol{L}^m+\beta\boldsymbol{L}^n)\phi^T(\boldsymbol{X})\phi(\boldsymbol{X})\boldsymbol{\alpha}\\&\Rightarrow K(\boldsymbol{L}^c+\alpha\boldsymbol{L}^f)\boldsymbol{K}\boldsymbol{\alpha}=\lambda\boldsymbol{K}(\boldsymbol{L}^m+\beta\boldsymbol{L}^n)\boldsymbol{K}\boldsymbol{\alpha}\end{aligned} \tag{13}$$

其中,$\boldsymbol{K}$ 为核矩阵(Gram 矩阵), $K_{ij}=\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$.

令列向量 $\boldsymbol{\alpha}^1,\boldsymbol{\alpha}^2,...,\boldsymbol{\alpha}^d$ 为式(13)的最大 d 个非零特征值所对应的特征向量,则对于一个新来样本 $\boldsymbol{x}$,它在第 k 个特征向量 $\boldsymbol{w}^k$ 上的投影是

$$(\boldsymbol{w}^k\cdot\phi(\boldsymbol{x}))=(\phi(\boldsymbol{X})\boldsymbol{\alpha}^k\cdot\phi(\boldsymbol{x}))=\sum_{i=1}^{n}\boldsymbol{\alpha}_i^k(\phi(\boldsymbol{x}_i)\cdot\phi(\boldsymbol{x}))=\sum_{i=1}^{n}\boldsymbol{\alpha}_i^k\kappa(\boldsymbol{x}_i,\boldsymbol{x}) \tag{14}$$

其中, $\boldsymbol{\alpha}_i^k$ 是 $\boldsymbol{\alpha}^k$ 的第 i 个分量.因此,如果令 $\boldsymbol{\theta}=[\boldsymbol{\alpha}^1,\boldsymbol{\alpha}^2,...,\boldsymbol{\alpha}^d]\in\boldsymbol{R}^{n\times d}$, 则新样本 $\boldsymbol{x}$ 的 d 维嵌入可以表示为 $\boldsymbol{x}\mapsto\boldsymbol{y}=\boldsymbol{\theta}^T\boldsymbol{K}(\cdot,x)$, 其中, $\boldsymbol{K}(\cdot,\boldsymbol{x})=[\kappa(\boldsymbol{x}_1,\boldsymbol{x}),...,\kappa(\boldsymbol{x}_n,\boldsymbol{x})]^T$. 对于训练样本,它们的嵌入是 $\boldsymbol{Y}=\boldsymbol{\theta}^T\boldsymbol{K}$,矩阵 $\boldsymbol{Y}$ 的第 i 列就是 $\boldsymbol{x}_i$ 的 d 维嵌入.

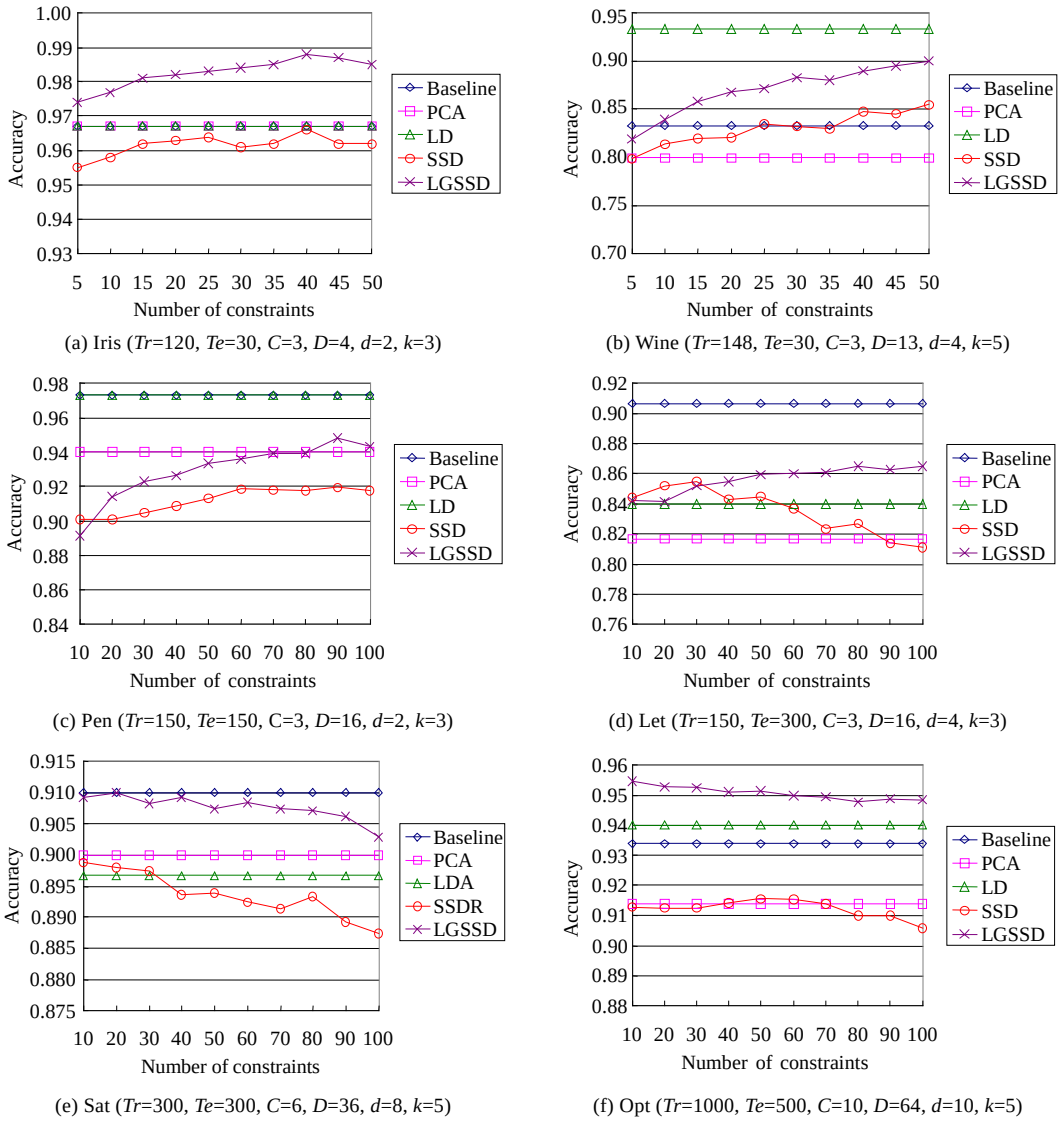
5 实验与分析

为了验证 LGSSDR 算法的有效性,本节通过几个实验展示该算法与其他算法相比较的结果,比较指标是降维后的低维投影在分类问题上的效果好坏(假设在分类时已知所有训练样本的标签,所使用的分类方法为最近邻分类法).与之相比较的算法包括基线方法(直接在原始输入空间上使用最近邻分类法)、PCA+最近邻分类法、LDA+最近邻分类法、SSDR+最近邻分类法(参考文献[13]中的 SSDR-CMU 方法).实验所用数据集是一组 UCI(University of California,Irvine)数据集[24]以及 CMU(Carnegie Mellon University)的 PIE(pose, illumination, and expression)人脸数据集[25].在下面的实验中,边信息都是通过从训练样本中随机选取样本点对来获取,如果某一样本点对的两个样本属于同一类,则把该点对放入正约束中;反之,则放入负约束中.在实验过程中,假设只知道边信息而不知道训练样本的标签信息(LDA 算法除外).另外,如无特别说明,LGSSDR 算法中的参数 α 和 β 都设置为 0.05.

5.1 UCI数据集

在本实验中,为了看清不同成对约束的数量对分类精度的影响,我们把 LGSSDR 算法与基线方法、PCA 算法、LDA 算法、SSDR 算法在 Iris,Wine,Pen-Based Handwritten Digits,Letter,Landsat Satellite 以及 Optical Handwritten Digits 数据集上的分类性能作一比较.对于 Iris 数据集,随机选择其中的 120 个样本作为训练集,剩下的 30 个样本作为测试集,这些样本分为 3 类,每个样本 4 维.对于 Wine 数据集,随机选择其中的 148 个样本作为训练集,剩余的 30 个样本作为测试集。该数据集共分为 3 类,每个样本 13 维.Pen-Based Handwritten Digits 数据集共有 10 992 个样本,分为 10 类(数字 0~9),每个样本 16 维,随机选择数字 3,8,9 的各 50 个样本作为训练集,50 个样本作为测试集.Letter 数据集共有 20 000 个样本,分为 26 类(字母 A~Z),每个样本 16 维,随机选择字母 I,J,L 的各 50 个样本作为训练集,100 个样本作为测试集.Landsat Satellite 数据集共有 6 435 个样本,随机选择其中的 300 个样本作为训练集,300 个样本作为测试集,这些样本共分为 6 类,每个样本 36 维.Optical Handwritten Digits 数据集共有 5 620 个样本,随机选择其中的 1 000 个样本作为训练集,500 个样本作为测试集,这些样本共分为 10 类(数字 0~9),每个样本 64 维.边信息按照本节开头所述的方法获取.所有 SSDR 和 LGSSDR 的实验结果都是 200

次不同边信息情况下的平均值(基线方法、PCA、LDA 不受边信息影响).实验结果如图 2 所示,其中 *Tr* 代表训练样本数,*Te* 代表测试样本数,*C* 代表样本类数,*D* 代表输入数据维数,*d* 代表约减数据维数,*k* 是 LGSSDR 算法的邻域参数(在本实验中,基线方法使用原始 *D* 维特征,LDA 算法使用 *C*–1 维判别特征,其他算法使用 *d* 维约减特征).



(a) Iris (*Tr*=120, *Te*=30, *C*=3, *D*=4, *d*=2, *k*=3)

(b) Wine (*Tr*=148, *Te*=30, *C*=3, *D*=13, *d*=4, *k*=5)

(c) Pen (*Tr*=150, *Te*=150, C=3, *D*=16, *d*=2, *k*=3)

(d) Let (*Tr*=150, *Te*=300, *C*=3, *D*=16, *d*=4, *k*=3)

(e) Sat (*Tr*=300, *Te*=300, *C*=6, *D*=36, *d*=8, *k*=5)

(f) Opt (*Tr*=1000, *Te*=500, *C*=10, *D*=64, *d*=10, *k*=5)

Fig.2 Classification accuracy on 6 UCI datasets with different number of constraints (NOC)

图 2 不同算法在不同约束数量(NOC)情况下,在 6 种 UCI 数据集上的分类精度

由图 2 中可以看出如下几点:

(1) LGSSDR 算法在大多数数据集上,无论是在少量约束数量还是在大量约束数量条件下,都能得到比 PCA 算法和 SSDR 算法更好的分类精度.其中,LGSSDR 算法在 Iris 和 Optical Handwritten Digits 数据集上的分类精度是所有算法中最好的,甚至好于 LDA 算法.LGSSDR 算法比 PCA 算法更好是因为 LGSSDR 得到了约束信息和先验知识,即流形假设,而 PCA 则纯粹是一种无监督的算法;LGSSDR 算法比 SSDR 算法更好是因为 LGSSDR 在利用约束信息的同时保持了数据集的局部以及全局结构,而 SSDR 在利用约束信息的同时仅保持了数据集的

全局结构,这说明局部结构的保持对算法性能的提升有着重要的作用.另外,我们发现 SSDR 算法在 Iris, Pen-Based Handwritten Digits 和 Landsat Satellite 数据集上的性能是最差的,甚至不如 PCA 算法,这说明仅仅利用边信息和保持全局信息是不够的,可能会导致降维效果不如无监督算法,这也从另一个侧面说明了保持数据集局部结构的重要性.

(2) 随着成对约束数量的增加,LGSSDR 算法的分类效果也在逐渐增加,但也有例外情况.比如,在 Landsat Satellite 和 Optical Handwritten Digits 数据集上,随着成对约束数量的增加,LGSSDR 算法的分类精度反而有小幅度下降,但其受影响的程度比 SSDR 要小.这是由于目前对约束的选取是随机的,而不是有目的的选择,这就使得在一些特殊的数据分布情况下,如多峰分布,过多的约束信息反而可能对降维效果起反作用.这说明,并不是对所有的数据集,约束信息都是越多越好,需要具体情况具体分析.

(3) 在 Letter 数据集上,SSDR 算法的分类精度随着成对约束数量的增加先有小幅度增加而后却有大幅度下降,这说明仅靠成对约束信息来指导降维并不一定能够取得很好的效果.而结合了邻域假设和非邻域假设的 LGSSDR 算法的分类精度随着成对约束数量的增加在一直增加,这说明该算法保持了数据集的内在结构,证明了所提出假设的合理性.

(4) 在 Pen-Based Handwritten Digits 数据集上,LGSSDR 算法与 SSDR 算法的性能是最差的,这是由于所提出的流形假设与数据集的实际情况有出入,说明当数据分布特性不满足流形假设时,使用流形假设并不能得到较好的结果.因此,在选用算法时应该具体情况具体分析,这与“没有免费的午餐”定理[1]也是相吻合的.

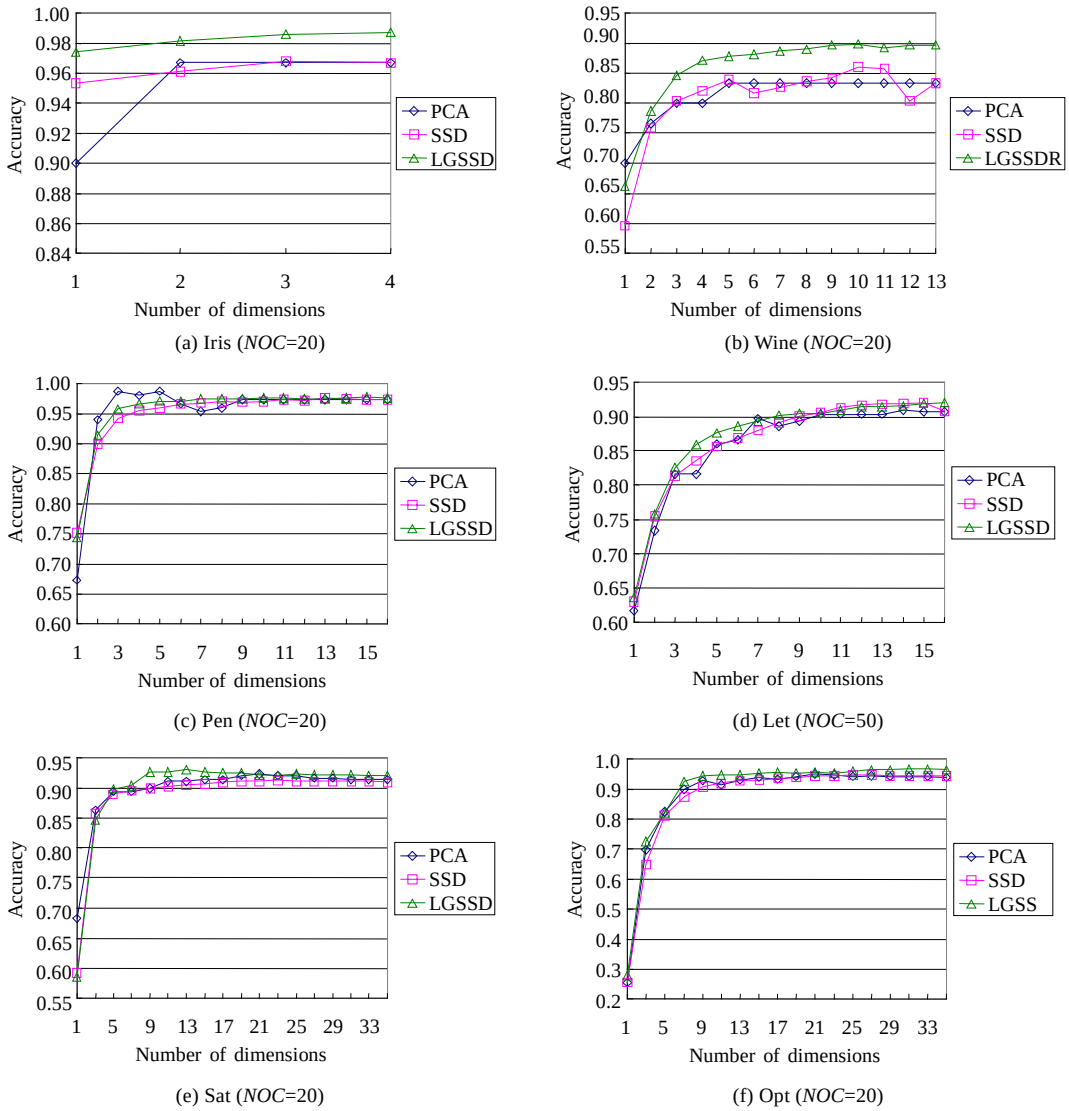
为了看清不同的投影维数对分类精度的影响,我们把 LGSSDR 算法与 PCA 算法和 SSDR 算法在上述 6 个数据集上的分类性能作一比较,结果如图 3 所示.从图中可以看出,在大多数情况下,LGSSDR 算法都能得到最好的分类效果,其分类效果随着维数的增加基本上呈递增趋势,并且其判别信息基本上集中在前面几个维数.这说明,LGSSDR 在很大范围的维数之内都能得到令人满意的结果.

5.2 PIE人脸数据集

CMU 的 PIE 人脸数据集共包括 68 个人的 41 368 张脸部图片.这些脸部图片是在不同的姿态、光照和表情的条件下采集到的.在本实验中,我们随机选择其中 24 个人,每人 170 张照片作为实验用例,这些照片都被裁剪为 32×32 像素大小,每张照片都是 256 级的灰度图片(如图 4 所示).这样,每张图片都可以看成是 1 024 维向量空间中的一个点.随机选取每人的 50 张图片作为训练集,剩余的 120 张图片作为测试集.边信息的获取方法与上一节相同.为了减少计算时间,避免奇异问题,提高计算精度,首先利用 PCA 算法对训练样本进行预处理(保留其中 98%的主成分).所有 SSDR 和 LGSSDR 的实验结果都是 100 次不同边信息情况下的平均值.

图 5(a)给出了不同算法在不同约束数量情况下在 PIE 数据集上的分类精度.其中类别数 C=24,输入数据维数 D=1024,约减数据维数 d=30,LGSSDR 算法的邻域参数 k=1.从中可以看出,在不同成对约束数量的情况下,LGSSDR 算法总是能够得到比基线方法、PCA 算法、SSDR 算法更好的结果,并且在少量约束的条件下也能得到比较理想的结果,而在成对约束数量较大的情况下所得结果甚至与 LDA 算法相当.

为了揭示邻域参数 k 对算法的影响,我们需要计算不同邻域大小情况下 LGSSDR 算法的分类精度.当训练样本较少时(每人 50 张图片作为训练集,剩余的 120 张图片作为测试集),实验结果如图 5(b)中最下面的曲线所示,当取 k=1 时,可以得到最好的结果;k=2 及 k=3 的结果要次于 k=1 而好于 k=0;而 k=4 和 k=5 的结果甚至比 k=0 还要差,可见,当训练样本较少时,选择一个合适的邻域参数能够极大地提高算法的性能.增加训练样本(每人 100 张图片作为训练集,剩余的 70 张图片作为测试集),实验结果如图 5(b)中间的曲线所示,可以看出,k 在 1~5 的范围内所得结果都比 k=0 要好,k=1 时的结果仍然是最好的.继续增加训练样本(每人 130 张图片作为训练集,剩余的 40 张图片作为测试集),实验结果如图 5(b)中最上面的曲线所示,可以看出,k 在 1~5 的范围内所得结果都比 k=0 要好,当 k=2 时取得最好的结果,并且不同 k 取值所得分类精度差别不大.可见,训练样本较多时,算法对 k 的取值就不那么敏感了.



(a) Iris (*NOC*=20) (b) Wine (*NOC*=20)

(c) Pen (*NOC*=20) (d) Let (*NOC*=50)

(e) Sat (*NOC*=20) (f) Opt (*NOC*=20)

Fig.3 Influence of LGSSDR algorithm under different NOC and different neighborhood parameters

图 3 不同约束数量及不同邻域参数对 LGSSDR 算法的影响

Fig.4 Some sample face images from the CMU PIE face database

图 4 CMU PIE 人脸数据集中的一些人脸图像样本

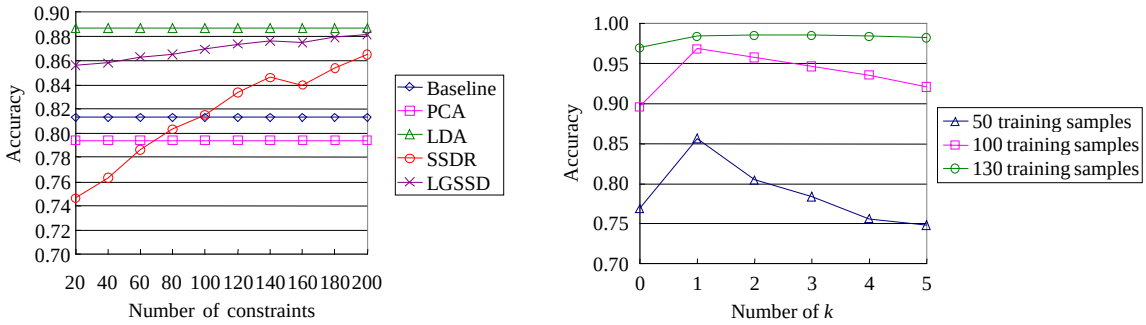


Fig.5

图 5

6 结束语

在本文中,我们提出了一种基于局部与全局保持的半监督维数约减方法(LGSSDR),该算法能够充分利用正负约束信息以及隐藏在未标记数据中的潜在信息,同时又能很好的保持数据集的全局以及局部结构.另外,我们还扩展了图嵌入理论,使之不仅能够适用于监督式以及无监督式学习,也能够适用于半监督学习,并且用该理论解释了 LGSSDR 算法的合理性.实验结果充分验证了该算法的有效性.

然而,LGSSDR 算法也存在着一些未解决的问题,比如我们的算法是建立在认为流形假设成立的基础上的,是否在所有数据集下该假设都成立?如果不成立应该如何处理?另外,目前参数 k,α和β都是靠经验指定,找到更好的方法来确定这些参数也需要进一步研究.再有,在有些数据集上过多的约束信息可能会对降维的效果起到反作用,找到一种方法(如主动学习的方法)消除这种负面影响也是一个有趣的工作.这些都是需要继续深入探讨的问题.

References:

[1] Duda RO, Hart PE, Stork DG. Pattern Classification. 2nd ed., New York: John Wiley & Sons, 2001.

[2] Turk MA, Pentland AP. Face recognition using eigenfaces. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Madison: IEEE Computer Society, 1991. 586−591.

[3] Martinez AM, Kak AC. PCA Versus LDA. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2001,23(2):228−233.

[4] Zhu XJ. Semi-Supervised learning literature survey. Technical Report, 1530, Department of Computer Sciences, University of Wisconsin at Madison, 2006. http://www.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf

[5] Wagstaff K, Cardie C. Clustering with instance-level constraints. In: Proc. of the 17th Int'l Conf. on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2000. 1103−1110.

[6] Klein D, Kamvar SD, Manning CD. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In: Sammut C, Hoffmann AG, eds. Proc. of the 19th Int'l Conf. on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2002. 307−314.

[7] Shental N, Hertz T, Weinshall D, Pavel M. Adjustment learning and relevant component analysis. In: Shental N, Hertz T, Weinshall D, Pavel M, eds. Proc. of the 7th European Conf. on Computer Vision. London: Springer-Verlag, 2002. 776−792.

[8] Bar-Hillel A, Hertz T, Shental N, Weinshall D. Learning a Mahalanobis metric from equivalence constraints. Journal of Machine Learning Research, 2005,6(6):937−965.

[9] Xing EP, Ng AY, Jordan MI, Russell S. Distance metric learning, with application to clustering with Side-information. In: Becker S, Thrun S, Obermayer K, eds. Advances in Neural Information Processing Systems 15. Cambridge: MIT Press, 2003. 505−512.

[10] Tang W, Zhong S. Pairwise constraints-guided dimensionality reduction. In: Proc. of the 2006 SIAM Int'l Conf. on Data Mining Workshop on Feature Selection for Data Mining. 2006. 59−66.

[11] Yeung DY, Chang H. Extending the relevant component analysis algorithm for metric learning using both positive and negative equivalence constraints. Pattern Recognition, 2006,39(5):1007−1010.

[12] Wu F, Zhou YL, Zhang CS. Relevant linear feature extraction using side-information and unlabeled data. In: Proc. of the 17th Int'l Conf. on Pattern Recognition. Washington: IEEE Computer Society, 2004. 582−585.

[13] Zhang DQ, Zhou ZH, Chen SC. Semi-Supervised dimensionality reduction. In: Proc. of the 7th SIAM Int'l Conf. on Data Mining. 2007. 629−634.

[14] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation, 2003,15(6): 1373−1396.

[15] Cai D, He XF, Han JW. Semi-Supervised discriminant analysis. In: Proc. of the 11th IEEE Int'l Conf. on Computer Vision. 2007. 1−7.

[16] Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000,290(22):2323−2327.

[17] Belhumeur PN, Hespanha JP, Kriegman DJ. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1997,19(7):711−720.

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

[19] Yang J, Zhang D, Yang JY, Niu B. Globally maximizing, locally minimizing: Unsupervised discriminant projection with application to face and palm biometrics. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2007,29(4):650−664.

[20] Schölokopf B, Smola AJ, Muller KR. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 1998,10(5):1299−1319.

[21] Mika S, Rätsch G, Weston J, Schölkopf B, Müller KR. Fisher discriminant analysis with kernels. In: Hu YH, Larsen J, Wilson E, Douglas S, eds. Proc. of the 1999 IEEE Signal Processing Society Workshop. Madison: IEEE Computer Society, 1999. 41−48.

[22] He XF, Niyogi P. Locality preserving projections. In: Thrun S, Saul L, Scholkopf B, eds. Advances in Neural Information Processing Systems 16. Cambridge: MIT Press, 2004. 153−160.

[23] Schölkopf B, Smola AJ. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Cambridge: MIT Press, 2002.

[24] Asuncion A, Newman DJ. UCI machine learning repository. School of Information and Computer Science, University of California, Irvine. 2007. http://mlearn.ics.uci.edu/MLRepository.html

[25] Sim T, Barker S, Bsat M. The CMU pose, illumination, and expression database. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2003,25(12):1615−1618.

韦佳(1982－),男,江西南昌人,博士生,主要研究领域为人工智能,流形学习,半监督学习.

彭宏(1956－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为机器学习,数据挖掘.