ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, No.10, October 2010, pp.2445−2453
doi: 10.3724/SP.J.1001.2010.03629

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

基于流形距离的半监督判别分析*

魏 莱[1+], 王守觉[2]

[1](同济大学 计算机科学与技术系,上海 201804)

[2](中国科学院 半导体研究所,北京 100083)

Semi-Supervised Discriminant Analysis Based on Manifold Distance

WEI Lai[1+], WANG Shou-Jue[2]

[1](Department of Computer Science and Technology, Tongji University, Shanghai 201804, China)

[2](Institute of Semiconductor, The Chinese Academy of Sciences, Beijing 100083, China)

+ Corresponding author: E-mail: weily105@hotmail.com, http://www.tongji.edu.cn

Wei L, Wang SJ. Semi-Supervised discriminant analysis based on manifold distance. *Journal of Software*, 2010,21(10):2445−2453. http://www.jos.org.cn/1000-9825/3629.htm

Abstract: Rich unlabeled data contains valuable information, which is useful for classification. Using information efficiently to improve the accuracy of classification is the major purpose of semi-supervised learning. This paper proposes a kind of semi-supervised classification approach called Semi-Supervised Discriminant Analysis that is based on Manifold Distance, SSDA. The intra-class neighbors, the inter-class neighbors, and the total neighbors of a selected point can be determined by the proposed manifold distance. The similarity between these neighbors and the point can be defined based on the manifold distance. The object function is defined using the similarity. As the experiments operated on the database ORL and YALE show, compared with the existing algorithms, the proposed algorithm can improve the accuracy of classified algorithms based on distance. When dealing with nonlinear dimensionality reduction problem, the Kernel SSDA (namely, kernel semi-supervised discriminant analysis based on manifold distance) is proposed. Also, the experimental results show the efficiency of this algorithm.

Key words: principal component analysis; linear discrimininat analysis; manifold distance; semi-supervised classfication

摘 要: 大量无类别标签的数据具有对分类有用的信息,有效地利用这些信息来提高分类精确度,是半监督分类研究的主要内容.提出了一种基于流形距离的半监督判别分析(semi-supervised discriminant analysis based on manifold distance,简称 SSDA)算法,通过定义的流形距离,能够选择位于流形上的数据点的同类近邻点、异类近邻点以及全局近邻点,并依据流形距离定义数据点与其各近邻点之间的相似度,利用这种相似度度量构造算法的目标函数.通过在 ORL,YALE 人脸数据库上的实验表明,与现有算法相比,数据集通过该算法降维后,能够使基于距离的识别算法具有更高的分类精确度.同时,为了解决非线性降维问题,提出了 Kernel SSDA,同样通过实验验证了算法的有效性.

关键词: 主成分分析;线性判别分析;流形距离;半监督判别分析

中图法分类号: TP181 **文献标识码**: A

* Received 2008-07-04; Revised 2009-02-13; Accepted 2009-03-31

随着信息技术的发展,数据普遍呈现出高维数的特点,要对这些高维数的数据进行分析、处理,从中获得有用的知识,则需占用大量的资源,如计算机内存、CPU 运算时间等等.并且,由于数据的高维数,使得一些成熟、有效的算法(如 KNN,SVM)性能急剧下降,面临所谓的"维数灾"问题.解决上述问题的一个理想的方法是,在保持数据集某种内在信息的情况下,对数据进行合理的维数约简或称属性提取.

在众多的维数约简算法中,主成分分析(principal component analysis,简称 PCA)[1]和线性判别分析(linear discriminant analysis,简称 LDA)[1,2]是两种应用最为广泛的线性降维算法.PCA 期望找到一个低维子空间,在这个子空间中,约简后数据集的方差达到最大.PCA 不考虑数据的类别,因而是一种无监督的降维算法.LDA 则利用了数据点的类别信息,目的是最大化类间的离散度,同时最小化类内的离散度.由于利用了数据点的类别信息,因此在一些识别和分类任务上,使用 LDA 降维要比使用 PCA 更加有效[3].但需要指出的是,PCA 和 LDA 都假设数据集具有全局的线性结构,而随着数据分析技术的发展,人们发现高维数的数据集往往具有一些低维的非线性结构(这种低维结构数学上称为流形)[4],在这种情况下,PCA 和 LDA 就受到了运用上的局限.

近几年兴起的流形学习算法[5],如 ISOMAP[6],LLE[7],Laplacian Eigenmaps[8]等能够有效解决这个问题.它们假设数据分布于一个嵌入在高维空间中的低维流形上,通过保持数据集的某种全局度量——测地线距离(ISOMAP)或局部结构(Laplacian Eigenmaps,简称 LLE),来对数据集进行维数约简.但是,经典的流形学习算法都是批处理的,如果要计算新增加的数据点(测试点)的低维坐标,则需要将新数据点加入训练集中重新运算整个算法,这样的做法显然是比较费时的.一些学者也提出了改进方法[9,10],但效果并不突出.保局部映射(locality preserving projection,简称 LPP)[11]为解决这个问题开辟了一条新路.LPP 是 Laplacian Eigenmaps 的一个线性逼近,它不对数据集作有全局线性结构的假设,只是期望保持数据点原本在流形上的邻域关系来对数据集进行降维.与 PCA 及 LDA 一样,LPP 能够直接得到一个线性变换矩阵,通过这个矩阵,测试点可以方便地得到低维坐标.因此,相比于经典的流行学习算法,LPP 在识别应用上的优势是显而易见的.LPP 虽然是非监督的降维算法,但受到 LPP 算法的启发,很多利用数据点邻域信息的线性监督降维算法也被不断地提出,如 LDE(local discriminant embedding)[12],MFA(marginal Fisher analysis)[13],SNNDA(stepwise nearest neighbor discriminant analysis)[14],ANMM(average neighborhood margin maximization)[15]等.

虽然这些监督降维算法在应用上取得了很好的效果,但是对于某些识别问题来说,我们并不能得到足够多的具有类别标签的数据,因为对数据标注类别是一件非常耗时的工作.而相反地,无标签的数据相对来说更容易得到,半监督分类就是通过希望利用这些无标签的数据具有的信息来增加分类的精确度.本文提出了一种基于流形距离的半监督判别分析(semi-supervised discriminant analysis based on manifold distance,简称 SSDA)方法.通过定义的流形距离来选择流形上数据的同类近邻点、异类近邻点和全局近邻点(不考虑类别的近邻点),这样选择的近邻点更符合数据具有低维流形分布的假设,并根据流形距离得到相应的相似度度量.利用这些相似性度量,与 MFA 类似地构造一个 Fisher 判别函数,在投影空间最大化领域内不同类点的边界,同时保持了全局数据集的内在流形结构.我们通过在 ORL 人脸数据库以及 YALE 人脸数据库上的实验,验证了与 PCA,LDA,LPP,MFA 等算法相比,本文提出的算法具有更高的识别精度.此外,我们利用核函数,提出了 Kernel SSDA 用来进行非线性维数约简,同样,通过实验验证了算法的有效性.

本文第 1 节介绍 MFA 算法.第 2 节介绍基于密度敏感距离的领域选择算法.第 3 节提出基于密度敏感距离的半监督判别分析方法.第 4 节是实验分析.最后得出结论.

1 边界 Fisher 判别分析

MFA 是一种线性监督降维算法[13],它解决了在数据集不为 Gauss 分布时,对数据集进行维数约简的问题.MFA 利用数据点邻域内的信息,在邻域内选择同类近邻点和不同类近邻点,然后在投影空间中将这两类点的平均边界尽量扩大.假设数据集为$\{(x_1,l_1),(x_2,l_2),\ldots,(x_N,l_N)\}$,其中,$x_i\in R^m$,表示一个 m 维的向量,$l_i\in L=\{1,2,\ldots,c\}$,为 x_i 的类别标签,L 是类别标签集.通过投影矩阵 $A_{m\times n}$,可以得到数据点 x_i 的低维映射 $y_i\in R^n$,$n<m$,即 $y_i=A^Tx_i$.

MFA 定义矩阵 S_w 来表征同类数据点的离散程度:

$$S_w=\sum_{i=1}^{N}\sum_{i\in N_{k_1}(j)\ \text{or}\ j\in N_{k_1}(i)}\|A^Tx_i-A^Tx_j\|^2=2A^TX(D-W)X^TA,$$

其中,W 是同类数据相似度邻接矩阵,$W_{ij}=\begin{cases}1, & i\in N_{k_1}(j)\ \text{or}\ j\in N_{k_1}(i)\\ 0, & \text{else}\end{cases}$,$N_{k_1}(i)$ 表示与 x_i 同类的 k_1 个近邻点,D 是一个对角阵,$D_{ii}=\sum_{j\neq i}W_{ij},\ \forall i$.同时,定义 S_b 来表征异类数据点的离散程度:

$$S_b=\sum_{i=1}^{N}\sum_{(i,j)\in P_{k_2}(l_i)\ \text{or}\ (i,j)\in P_{k_2}(l_j)}\|A^Tx_i-A^Tx_j\|^2=2A^TX(D^p-W^p)X^TA,$$

其中,W^p 是不同类数据相似度邻接矩阵,$W_{ij}^p=\begin{cases}1, & (i,j)\in P_{k_2}(l_i)\ \text{or}\ (i,j)\in P_{k_2}(l_j)\\ 0, & \text{else}\end{cases}$,$P_{k_2}(l_i)$ 表示与 x_i 不同类的 k_2 个近邻点,D^p 是一个对角阵,$D_{ii}^p=\sum_{j\neq i}W_{ij}^p,\ \forall i$.MFA 通过最大化如下一个广义 Rayleigh 商来求解 A:

$$J(A)=\frac{A^TX(D^p-\tilde{W}^p)X^TA}{A^TX(D-\tilde{W})X^TA},\text{满足 }A^TA=I \tag{1}$$

文献[13]中指出,在 MFA 得到的子空间中,采用最近邻分类器,数据的分类精度也是比较高的.但是我们也发现,MFA 中存在着两个问题:第一,MFA 是根据欧式距离来选择近邻点的,而由于数据的低维流形分布,使得这样选择的近邻点可能并不能表征流形上数据的真正邻域结构;第二,MFA 是完全监督的.正如前文所述,在具有类别的标签其数据不足时,MFA 的降维效果可能并不理想.为了解决这样两个问题,我们提出了基于流形距离的半监督判别分析 SSDA 算法.下一节我们先来讨论近邻点选择.

2 基于流形距离的近邻选择

2.1 流形距离

欧氏距离是最常用的距离度量,但是在数据集不具有全局线性结构时,欧氏距离就不是一种合理的数据间的距离度量.为了解决位于流形上的数据的邻域点选择,我们这里首先定义一种流形距离.

定义 1. 流形上两点间 x_i,x_j 的线段长度定义为 $L(x_i,x_j)=\mathrm{e}^{\frac{d(x_i,x_j)}{\sigma}}-1$.

$d(x_i,x_j)$表示 x_i,x_j 之间的欧氏距离,σ是可调参数.而流形上任意两点间的距离则定义如下:

定义 2. 将数据点看作是图 $G=(V,E)$的顶点,V 是顶点集合,E 是边集.P_{ij} 表示图上连接数据点 x_i,x_j 的所有路径集合,则 x_i,x_j 之间的流形距离为

$$MD(x_i,x_j)=\min_{p\in P_{ij}}\sum_{k=1}^{|p|-1}L(p_k,p_{k+1}) \tag{2}$$

即图上所有连接这两点路径上线段总长的最小值.

需要指出的是,这里定义的流形距离本质上和文献[16]是一样的,但是文献[16]中并没有给出伸缩因子ρ的指导性确定方法,而这里定义中的σ,在实验中可设为不同类中心平均距离的两倍.

2.2 基于流形距离的近邻选择算法(NSMD)

有了上述定义,我们可以计算数据集任意两点之间的流形距离,然后选择合适的近邻点.但是由于图 $G=(V,E)$是稠密的,通过 Dijkstra 算法计算出任意两点间流形距离的整个算法时间复杂度为 $O(|V|^3)$,比较费时.而事实上,我们也只需通过流形距离来选择合适的邻域.为此,提出基于流形距离的近邻选择(neighbors selecting by manifold distance,简称 NSMD)算法:假设 $X=\{x_1,x_2,\ldots,x_N\}$,基于欧氏距离的邻域大小参数 K_d,基于流形距离的邻域大小参数 K_D,$K_D\leqslant K_d$.

Step 1. 通过欧氏距离计算任意一点 x_i 的欧氏邻域 $N_d(x_i)$,按照距离升序排列 $N_d(x_i)$,且令

$$MD(x_i,x_j)=L(x_i,x_j)=e^{\frac{d(x_i,x_j)}{\sigma}}-1$$

$d(x_i,x_j)$表示 x_i,x_j 之间的欧氏距离;

Step 2. for i=1 to N

令 k=1;

while $k<K_d$

取 $N_d(x_i)$中的第 k 点,记为 x_{ik}

for j=1 to K_D

令 $N_d(x_{ik})$中的第 j 点,记为 x_{ikj}

if $MD(x_i,x_{ik})+MD(x_{ik},x_{ikj})<MD(x_i,x_{iK_d})$ 且 x_{ikj} 不在 $N_d(x_i)$中

将 x_{ikj} 按升序插入 $N_d(x_i)$,且令 $MD(x_i,x_{ikj})=MD(x_i,x_{ik})+MD(x_{ik},x_{ikj})$, break;

endif $MD(x_i,x_{ik})+MD(x_{ik},x_{ikj})<MD(x_i,x_{iK_d})$

更新 $MD(x_i,x_{ikj})=MD(x_i,x_{ik})+MD(x_{ik},x_{ikj})$,将按照距离升序排列 $N_d(x_i)$; break;

endif

if $MD(x_i,x_{ik})+MD(x_{ik},x_{ikj})>MD(x_i,x_{iK_d})$ or

($MD(x_i,x_{ikj})=MD(x_i,x_{ik})+MD(x_{ik},x_{ikj})$且 $j==K_D$)

$k=k+1$; break;

endif

endfor

endwhile

endfor

Step 3. 对于任意取 x_i,取 $N_d(x_i)$中前 K_D 个点构成 x_i 流形距离的邻域 $N_D(x_i)$及记录任意 $x_{ij}\in N(x_i)$与 x_i 间的流形距离 $D(x_{ij})$;

算法中,Step 2 事实上完成了这样一个工作:对 x_i 的欧氏近邻点分别计算其与 x_i 的流形距离,然后按照流形距离来重新选择 x_i 的近邻点.由于两种距离的特点,所以在算法中首先选择较多的欧氏近邻点,通过 Step 2 计算后,从中选择一部分以流形距离计算的近邻点.

NSMD 算法的时间复杂度最坏情况下为 $O(|V|K_DK_d)(<O(|V|^3))$,因此能有效的用来选取数据点的近邻点.图 1、图 2 是以欧式距离和流形距离对点 x 近邻的选择结果.黑色实心点表示 x,与其有线段相连的点表示其近邻点.

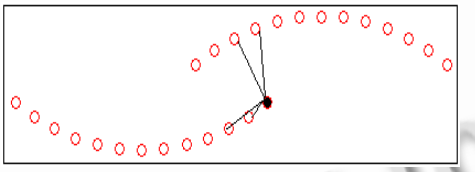


Fig.1 Neighbors selected by Euclidean distance

图 1 以欧式距离选择的近邻点

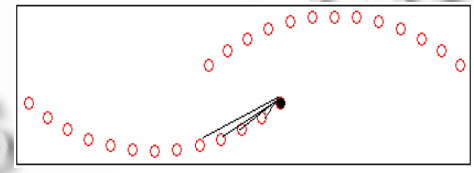


Fig.2 Neighbors selected by manifold distance

图 2 以流形距离选择的近邻点

3 基于流形距离的半监督判别分析(SSDA)

3.1 目标函数

首先给出一些符号说明.设 $X=\{X_N,X_M\}=\{x_1,x_2,\ldots,x_N,x_{N+1},\ldots,x_{N+M}\}$,前 N 个数据点带有类别标签,后 M 个点不带类别标签,其他符号假设同上文.我们知道,通过 NSMD 算法可以选择位于低维流形上数据点的近邻点,同时得到它们之间的流形距离.利用这些信息以及数据点的类别,可以构造同类数据点的相似度邻接矩阵 $\tilde{W}_w$:

$$\tilde{W}_{wij}=\begin{cases}\dfrac{1}{MD_w(x_i,x_j)+1}, & x_j\in N_{wMDk_1}(x_i)\\ 0, & \text{else}\end{cases}.$$

这里,$N_{wMDk_1}(x_i)$ 是由与 x_i 同类别的 k_1 个数据点以流形距离构成的同类点邻域,$MD_w(x_i,x_j)$为 x_i,x_j 之间的流形距离.但这样得到的 $\tilde{W}_w$ 可能不对称,因此,再令 $\tilde{W}_{wij}=\max\{\tilde{W}_{wij},\tilde{W}_{wji}\}$,然后我们也可以定义表征同类数据点离散程度的矩阵:

$$\tilde{S}_w=\sum_{i=1}^{N}\sum_{j\in N_{wDk_1}(x_i)}\tilde{W}_{wij}\|A^Tx_i-A^Tx_j\|^2=2A^TX_N(D_w-\tilde{W}_w)X_N^TA \tag{3}$$

D_w 是一个对角阵,$D_{wii}=\sum_{j\neq i}\tilde{W}_{wij},\forall i$, $X_N=\{x_1,x_2,\ldots,x_N\}$.

同样,还可以构造异类数据点的相似度邻接矩阵 $\tilde{W}_b$:

$$\tilde{W}_{bij}=\begin{cases}\dfrac{1}{MD_b(x_i,x_j)+1}, & x_j\in N_{bMDk_2}(x_i)\\ 0, & \text{else}\end{cases}\text{ 且 }\tilde{W}_{bij}=\max\{\tilde{W}_{bij},\tilde{W}_{bji}\}.$$

这里,$N_{bMDk_1}(x_i)$ 是由与 x_i 不同类别的 k_2 个数据点以流形距离构成的异类点邻域.同样,$MD_b(x_i,x_j)$为 x_i,x_j 之间的流形距离.于是,表征异类数据点离散程度的矩阵 $\tilde{S}_b$ 可以表示为

$$\tilde{S}_b=\sum_{i=1}^{N}\sum_{j\in N_{bDk_2}(x_i)}\tilde{W}_{bij}\|A^Tx_i-A^Tx_j\|^2=2A^TX_N(D_b-\tilde{W}_b)X_N^TA \tag{4}$$

D_b 是一个对角阵,$D_{bii}=\sum_{j\neq i}\tilde{W}_{bij},\forall i$.

我们期望的是要寻找一个投影矩阵 A,能够使 $\tilde{S}_b$ 尽量地大而使 $\tilde{S}_w$ 尽量地小.但是如上文所述,通常情况下,具有类别的数据点在数据集中只占少数,只满足这样一个目标的投影矩阵,可能并不是最佳的.而大量的不带类别标签的数据包含对分类有用的信息,我们相信,这些信息就是数据点与其全局近邻点(不再考虑类别)之间的邻接关系[17].为此,我们希望通过投影变换 A,能够保持数据点之间原有的邻接关系.这样的目的可以通过使如下定义的 $\tilde{S}_t$ 尽量小来满足.

定义数据点的全局相似度邻接矩阵 $\tilde{W}_t$:

$$\tilde{W}_{tij}=\begin{cases}\dfrac{1}{MD_t(x_i,x_j)+1}, & x_j\in N_{tDk_3}(x_i)\\ 0, & \text{else}\end{cases}\text{ 且 }\tilde{W}_{tij}=\max\{\tilde{W}_{tij},\tilde{W}_{tji}\}.$$

这里,$N_{tMDk_3}(x_i)$ 是由 x_i 的 k_3 个数据点以流形距离构成的邻域,这里都不考虑数据点类别.令

$$\tilde{S}_t=\sum_{i=1}^{N+M}\sum_{j\in N_{tDk_3}(x_i)}\tilde{W}_{tij}\|A^Tx_i-A^Tx_j\|^2=2A^TX(D_t-\tilde{W}_t)X^TA \tag{5}$$

D_t 是一个对角阵,$D_{tii}=\sum_{j\neq i}\tilde{W}_{tij},\forall i$. 于是,综上所述,给出基于流形距离的半监督判别分析的目标函数:

$$\max_{A^TA=I}J(A)=\max_{A^TA=I}\frac{\tilde{S}_b-\alpha\tilde{S}_t}{\tilde{S}_w}=\max_{A^TA=I}\frac{A^TX_N(D_b-\tilde{W}_b)X_N^TA-\alpha A^TX(D_t-\tilde{W}_t)X^TA}{A^TX_N(D_w-\tilde{W}_w)X_N^TA} \tag{6}$$

其中,α是一个正实数.令 $\tilde{I}=\begin{pmatrix}I_N & 0\\ 0 & 0\end{pmatrix}_{M+N}$,$I_N$ 为 N 阶单位阵,上式可以写成

$$\max_{A^TA=I}\frac{A^TX(\tilde{I}(D_w-\tilde{W}_w)-\alpha(D_t-\tilde{W}_t))X^TA}{A^TX\tilde{I}(D_w-\tilde{W}_w)X^TA} \tag{7}$$

由于存在数据的高维数以及样本个数的限制,$X\tilde{I}(D_w-\tilde{W}_w)X^T$ 不一定可逆.为了得到 A 的稳定解,我们求解如下一个广义特征问题[1]:

$$X(\tilde{I}(D_w-\tilde{W}_w)-\alpha(D_t-\tilde{W}_t))X^TA=(X(\tilde{I}(D_w-\tilde{W}_w))X^T+\beta I)A \quad (8)$$

其中,β为一个小的正实数.

3.2 算法流程

因此,基于流形距离的半监督判别分析(SSDA)算法的流程如下:

Step 1. 对任意 $x_i\in X_N$,通过 NSMD 算法选择其同类点邻域 $N_{wMDk_1}(x_i)$、异类点邻域 $N_{bMDk_2}(x_i)$ 以及全局邻域 $N_{tMDk_3}(x_i)$;对任意 $x_i\in X_M$ 通过 NSMD 算法选择全局邻域 $N_{tMDk_3}(x_i)$;

Step 2. 利用流形距离构造同类点相似度邻接矩阵、异类点相似度邻接矩阵 $\tilde{W}_b$ 以及全局相似度邻接矩阵 $\tilde{W}_t$,同时得到 D_w,D_b,D_t;

Step 3. 通过求解广义特征问题 $X(\tilde{I}(D_w-\tilde{W}_w)-\alpha(D_t-\tilde{W}_t))X^TA=(X(\tilde{I}(D_w-\tilde{W}_w))X^T+\beta I)A$,得到投影矩阵 A;

Step 4. 对任意 $x_i\in X$,求其投影点 $y_i=A^Tx_i$.

通过 SSDA 算法对数据进行降维后,能够使得基于距离的分类算法识别准确率得以提高.实验分析见下节.

但需要指出的是,SSDA 算法本质上仍是一种线性降维算法,当数据集分布呈高度分线性时,SSDA 算法并不能发现数据集构成的低维流形的内在结构.为此,我们利用核函数(kernel function)将 SSDA 加以扩展,提出 Kernel SSDA,使其能够进行非线性维数约简.具体分析见附录.

4 实验分析

在这一节中,我们进行一些实验,以对比本文提出的算法与现存算法之间的优劣.

4.1 实验数据集

我们采用两个人脸数据库,即 ORL 人脸数据库[17]和 YALE 人脸数据库[18].

ORL 人脸库包含了 40 个人共 400 张人脸照片,每人 10 张,分别是在不同的表情、光照以及一些面部细节情况下拍摄的结果.原始图像为 92×112 的 256 级灰度图,在实验中将其正规化到 32×32.图 3 是 ORL 数据库的一些图片样本.

Fig.3 Image samples of ORL face

图 3 ORL face 的图片样本

YALE 人脸库包含了 15 个人每人 11 张灰度照片,同样,这些照片也是在不同的光照条件以及不同的面部表情条件下拍摄的结果.在实验中,我们将图片都正规化到 32×32 大小.图 4 是 YALE 数据库的一些图片样本.

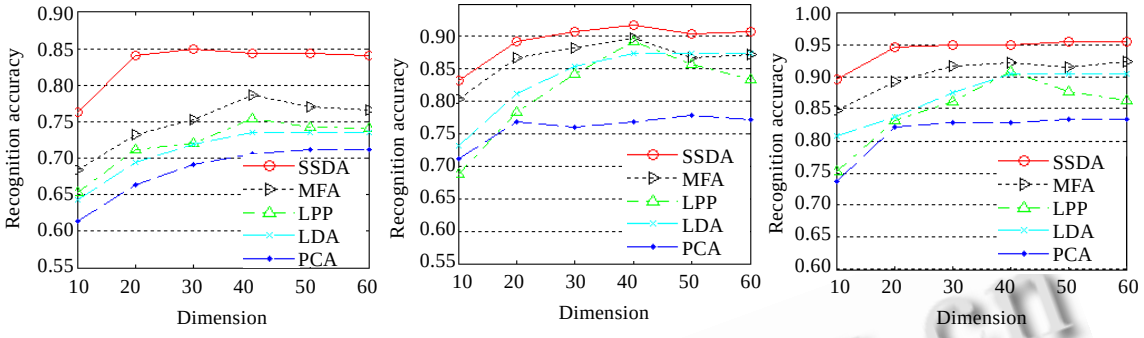
Fig.4 Image samples of YALE face

图 4 YALE face 的图片样本

4.2 实验测试

对于 ORL 数据库,我们从每个人的照片中随机抽取出 l(l=2,3,4)张照片作为训练样本,余下 10–l 张照片作为测试样本.同样,在 YALE 数据库中,我们从每个人的照片中随机抽出 l(l=2,3,4)张照片作为训练样本,而余下 11–l

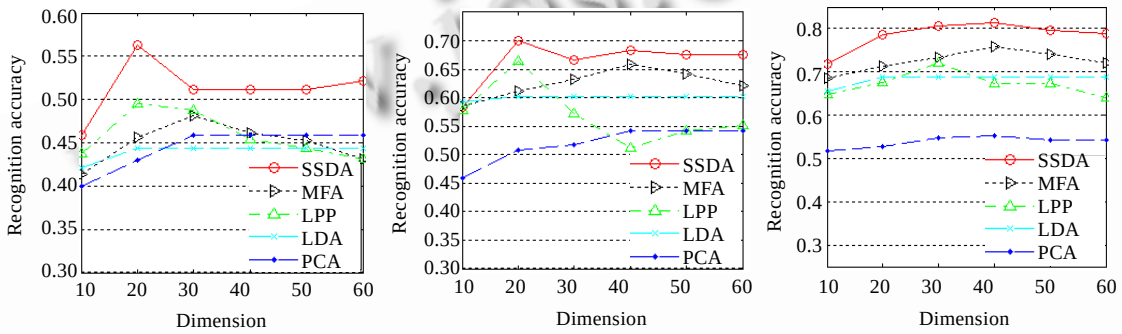
张照片作为测试样本.我们采用 PCA,LDA,LPP,MFA 等算法作比较.在实验中,将同类近邻点、异类近邻点、全局近邻点的最大值都设为 10,计算流形距离是将可调参数σ设为不同类中心平均距离的两倍,参数α通过交叉实验来确定.识别算法采用最近邻方法,整个实验重复 20 次,最后得出各种算法的识别平均结果,绘成图 5、图 6(随机选取 2 幅~4 幅照片作为带标签的训练数据集).



(a) 2 train (b) 3 train (c) 4 train

Fig.5 Recognition accuracy on the ORL database

图 5 ORL 数据库上的识别准确率



(a) 2 train (b) 3 train (c) 4 train

Fig.6 Recognition accuracy on the YALE database

图 6 YALE 数据库上的识别准确率

需要说明的是,作为半监督学习算法的 SSDA,在算法训练中需要使用到测试样本.但是,在使用测试样本时,我们不利用其带有的类别信息.而 LDA 和 MFA 作为监督降维算法,在算法训练时只利用带标签的训练样本.

对于 Kernel SSDA,我们采用 Gauss 核函数,核参数 t 也通过交叉验证的方法来确定,比较算法采用 KPCA 和 KDA(表 1 的括号中是相应的投影空间维数).

Table 1 Recognition accuracy of kernel approach on ORL & YALE database (%)

表 1 核方法在 ORL,YALE 数据库上的识别准确率(%)

Method	ORL			YALE		
	2 train	3 train	4 train	2 train	3 train	4 train
KPCA	63.21(48)	77.14(52)	80.34(58)	49.34(45)	54.97(48)	61.19(54)
KDA	80.29(38)	87.75(38)	94.24(39)	53.25(15)	65.17(14)	71.29(16)
Kernel SSDA	**82.45(44)**	**92.67(50)**	**94.37(59)**	**56.17(54)**	**70.32(69)**	**83.14(64)**

5 结 论

大量的实验结果表明,高维数的数据点分布在一个低维的流形上,本文首先利用所提出的流形距离来选择

位于流形上的数据点的同类近邻点、异类近邻点以及全局近邻点,然后根据这些近邻点和原数据点之间的流形距离得到相应的相似度度量,最后利用这些相似度度量构造基于流形距离的半监督判别分析算法(SSDA)的目标函数.通过在标准人脸数据库上的实验,验证了提出的算法相比于现有算法,更能提高基于距离算法的分类准确率.同时,为了解决非线性降维问题,我们将核函数引入 SSDA,提出了 Kernel SSDA,同样,通过实验也验证了其有效性.

References:

[1] Duda RO, Hart PE, Stork DG, Wrote; Li HD, Yao TX, *et al*., Trans. Pattern Classification. 2nd ed., Beijing: China Machine Press, 2003. 94−102 (in Chinese).

[2] Martinez AM, Kak AC. PCA versus LDA. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2001,23(2):228−233. [doi: 10.1109/34.908974]

[3] Seung HS, Lee DD. The manifold ways of perception. Science, 2000,290(5500):2268−2269. [doi: 10.1126/science.290.5500.2268]

[4] Luo SW, Zhao LW. Manifold learning algorithm based on spectral graph theory. Journal of Computer Research and Development, 2006,43(7):1173−1179 (in Chinese with English abstract). [doi: 10.1360/crad20060707]

[5] 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]

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

[7] Belkin M, Niyogi P. Laplacian Eigenmaps and spectral techniques for embedding and clustering. In: Advances in Neural Information Processing System, Vol.14. British Columbia, 2001. 585−591. http://books.nips.cc/nips14.html

[8] Kouropteva O, Okun O, Pietiknen M. Incremental locally linear embedding. Pattern Recognition, 2005,38(10):1764−1767. [doi: 10.1016/j.patcog.2005.04.006]

[9] Law MHC, Jain AK. Incremental nonlinear dimensionality reduction by manifold learning. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2006,28(3):337−391. [doi: 10.1109/TPAMI.2006.46]

[10] He X, Yan S, Hu Y, Niyogi P, Zhang H. Face recognition using Laplacianfaces. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005,27(3):328−340. [doi: 10.1109/TPAMI.2005.55]

[11] Chen HT, Chang HW, Liu TL. Local discriminant embedding and its variants. In: Proc. of the Int'l Conf. on Computer Vision and Pattern Recognition. San Diego, 2005. 846−853. http://www.informatik.uni-trier.de/~ley/db/conf/iccv/iccv2005-2.html

[12] Yan S, Xu D, Zhang B, Zhang H. Graph embedding: A general framework for dimensionality reduction. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. San Diego, 2005. 830−837. http://www.informatik.uni-trier.de/~ley/db/conf/iccv/iccv2005-2.html

[13] Qiu X, Wu L. Face recognition by stepwise nonparametric margin maximum criterion. In: Proc. of the 10th IEEE Int'l Conf. on Computer Vision. Beijing, 2005. 1567−1572. http://www.informatik.uni-trier.de/~ley/db/conf/iccv/iccv2005-2.html

[14] Wang F, Zhang CS. Feature extraction by maximizing the average neighborhood margin. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Minnesota, 2007. 1−8. http://iccv2007.rutgers.edu/

[15] Wang L, Bo LF, Jiao LC. Density-Sensitive spectral clustering. ACTA ELECTRONICA SINICA, 2007,35(8):1577−1581 (in Chinese with English abstract).

[16] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research, 2006,7(11):2399−2434.

[17] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html. 1994.

[18] http://cvc.yale.edu/projects/yalefaces/yalefaces.html. 1997.

附中文参考文献:

[1] Duda RO, Hart PE, Stork DG,著;李宏东,姚天翔,等,译.模式分类.第 2 版,北京:机械工业出版社,2003.94−102.

[4] 罗四维,赵连伟.基于谱图理论的流形学习.计算机研究与发展,2006,43(7):1173−1179. [doi: 10.1360/crad20060707]

[15] 王玲,薄列峰,焦李成.密度敏感的谱聚类.电子学报,2007,35(8):1577−1581.

附 录

Kernel SSDA.

假设存在一个非线性映射$\Phi:R^D\to H$,H为一个高维内积空间.$\Phi(x_i)$表示x_i在H中的值,对于在空间H中两点间$\Phi(x_i),\Phi(x_j)$的距离,定义成$\|\Phi(x_i)-\Phi(x_j)\|=\sqrt{K_{ii}+K_{jj}-2K_{ij}}$,其中,$K_{ij}=\langle\Phi(x_i),\Phi(x_j)\rangle$是核矩阵$K$的$(i,j)$个元素.$K(x,y)=\langle\Phi(x),\Phi(y)\rangle$表示$H$中的内积运算,称为核函数.常用的核函数有Guass核$K(x,y)=\exp\left(\frac{-\|x-y\|^2}{t}\right)$以及多项式核$K(x,y)=(1+x^Ty)^d$.

需要注意的是,由于我们将数据映射到了一个高维空间,数据点的距离通过内积运算得到,因此,数据点的邻域与用欧式距离选取的可能不一致.因此,通过NSMD算法得到的邻域也可能不一样.去除公式(3)~公式(5)中的常数,同时通过映射函数得到:

$$\tilde{S}_w=A^T\Phi(X_N)(D_w-\tilde{W}_w)\Phi(X_N)^TA,\ \tilde{S}_b=A^T\Phi(X_N)(D_b-\tilde{W}_b)\Phi(X_N)^TA,\ \tilde{S}_t=A^T\Phi(X)(D_t-\tilde{W}_t)\Phi(X)^TA,$$

其中,$\Phi(x_N)=\{\Phi(x_1),\Phi(x_2),\ldots,\Phi(x_N)\}$,$\Phi(X)=\{\Phi(x_1),\Phi(x_2),\ldots,\Phi(x_{N+M})\}$.根据核函数性质,变换矩阵$A=\{a_1,a_2,\ldots,a_n\}$的列向量$a_t(1\leqslant t\leqslant n)$位于由$\Phi(x_1),\Phi(x_2),\ldots,\Phi(x_{N+M})$张成的空间中,因此可以得到$a_t=\sum_{i=1}^{N+M}b_i^t\Phi(x_i)$,于是有,

$$\tilde{S}_w=A^T\Phi(X_N)(D_w-\tilde{W}_w)\Phi(X_N)^TA=B^TK_N(D_w-\tilde{W}_w)K_NB,$$
$$\tilde{S}_b=A^T\Phi(X_N)(D_b-\tilde{W}_b)\Phi(X_N)^TA=B^TK_N(D_b-\tilde{W}_b)K_NB,$$
$$\tilde{S}_t=A^T\Phi(X)(D_t-\tilde{W}_t)\Phi(X)^TA=B^TK(D_t-\tilde{W}_t)KB,$$

其中,$K_N=\Phi(X_N)^T\Phi(X)$,$K=\Phi(X)^T\Phi(X)$.于是,公式(8)就转化为

$$K(\tilde{I}(D_w-\tilde{W}_w)-\alpha(D_t-\tilde{W}_t))KB=K(\tilde{I}(D_w-\tilde{W}_w))KB,$$

其中,$B=\{b_1,b_2,\ldots,b_{N+M}\}$为所需求解的变换矩阵.于是,对于测试点$x$,我们可以得到其投影坐标为$y=B^TK_x$,$K_x=\{K(x,x_1),K(x,x_2),\ldots,K(x,x_{N+M})\}$.

魏莱(1980－),男,江苏苏州人,博士生,主要研究领域为流形学习,仿生模式识别.

王守觉(1925－),男,教授,博士生导师,中国科学院院士,主要研究领域为仿生模式识别.