基于测地 Gabriel图的非线性流形判别分析

陈华杰, 韦 巍
(浙江大学电气工程学院, 浙江杭州 310027)

摘 要: 针对位于非线性流形上类别数据的判别分析问题, 提出了一种基于测地 Gabriel图的局部判别器融合算法. 利用测地距离表征流形的内在几何结构, 由此构造测地 Gabriel图确定异类数据相互靠近的局部临界区域, 进而训练得到局部线性的判别器. 整体的非线性判别器由多个局部判别器融合得到: 基于柔性边界准则函数, 以迭代优化的方式, 为每个局部判别器分配最佳的权重系数, 整体上逐步提高异类样本间的区分度. 在人工合成数据集以及人脸图像库上的实验证明了本文算法的有效性.

关键词: 非线性流形; 测地距离; 测地 Gabriel图; 多判别器融合
中图分类号: TP181 文献标识码: A 文章编号: 0372-2112 (2006) 08-1405-05

Discriminant Analysis on Nonlinear Manifold Based on Geodesic Gabriel Graph

CHEN Hua-jie, WEI Wei
(College of Electrical Engineeming, Zhejiang University, Hangzhou, Zhejiang 310027, China)

Abstract As for the discriminant analysis on nonlinear manifold, a geodesic Gabriel graph based local discriminant merging approach was proposed. Using geodesic distance to discover the intrinsic geometry of the manifold, the geodesic Gabriel graph was constructed to locate the critical local regions where the local linear discriminants would be learned. The global nonlinear discriminant was achieved by merging the multiple local discriminants. The soft margin criterion based merging algorithm assigned the best weight to each local discriminant in an iterative way and upgraded the detection accuracy stepwise. The superiority of this algorithm was confirmed by experiments both on synthesized data and face image set.

Key words nonlinear manifold; geodesic distance; geodesic Gabriel graph; multi-discriminant tm erging

1 引言

在计算机视觉、模式识别等领域通常要对高维的数据如人脸图像等进行判别分析, 这些高维数据的很大一部分可以用高维空间的非线性嵌入子流形来建模[1~3]. 简略地说, 一个流形是一个集合且带有一个拓扑, 使得在局部上此集合可看作是欧式空间. 已提出的流形学习方法, 如 LLE[1], Isomap[2], Laplacian Eigenmaps[3]等具有表达流形上非线性结构的能力; 但所采用的都是无导师学习的方式, 所获取的特征, 往往并非判别意义上的最优. 现有的用于判别的扩展流形学习算法, 根据流形上数据的所属类别的异同, 对同类与异类数据施加不同的处理[4~6]. Wu等[4]分别计算类内、类间的测地距离. Chen等[5]在保持同类数据相互靠近的情况下, 尽量扩大映射后的异类数据的距离. Geng等[6]综合数据之间的测地距离以及类别来获取相似度, 进而降维. 上述扩展算法, 由于在降维处理时需要数据的类别信息, 对于未标定类别的测试样本数据, 无法直接获得映射后的低维数值.

流形学习的一个基本思想是: 先获取流形上局部区域的几何结构, 再对局部的几何结构调整并扩展至整体, 从而获得非线性流形的整体表达[7]. 此种先局部后整体的策略可以借鉴用来对流形进行判别分析. 对于整体非线性可分的流形, 在若干选定的局部区域, 则可能是线性可分的.

本文提出了一种非线性流形的判别分析方法. 采用测地距离[2]来表达数据间的非线性关系, 构造测地 Gabriel图在非线性流形上划分出异类样本相互靠近的局部区域, 进而在这些局部区域上进行线性判别分析. 而整体的非线性判别器通过多个局部线性判别器融合而得. 由此, 将整体非线性流形的判别分析问题分解为局部线性判别分析以及局部判别器的融合问题.

收稿日期: 2005-10-21; 修回日期: 2006-05-18
基金项目: 浙江省青年人才培养资助项目(No. R105341)

2 测地 Gabriel图

在m维空间上的一个d维流形M (d<m)由下面的映射函数构成:

$$f:\ C \rightarrow R^m, C \subset R^d$$

式中: C是R^d上的紧子集. 采样自M上的数据集合X = $\{x_1, x_2, \cdots, x_N\}$, $x_i \in R^m$, 每个数据有对应的标号为y_i, 对于二分类问题, 有$y_i \in \{-1, 1\}$. 流形上异类数据相互靠近的区域可以通过构造测地Gabriel图来确定.

2.1 Gabriel图

近邻法是常用的一种模式识别的方法. 采用k近邻法进行模式分类时, 对于一个新样本, 根据其k个最近邻的训练样本的优势类别进行分类. Gabriel图[8]可用来从训练样本集中找出异类样本相对靠近的临界区域(critical region).

当2个点A、B满足以AB为直径的超球中不包含任何其他点时, 亦即对其他任何一个点, 满足条件

$$d^2(D, A) + d^2(D, B) > d^2(A, B) \qquad (1)$$

时, 则称它们为Gabriel近邻; 其中d是两个点之间的距离测度.

图1给出了一个采用欧式距离的Gabriel近邻的例子, 其中A与B是Gabriel近邻, 而B与C不是. 所有的Gabriel近邻以边相连接则构成了Gabriel图. 不同类别的数据间的Gabriel图可以采用Gabriel剪辑算法得到[8].

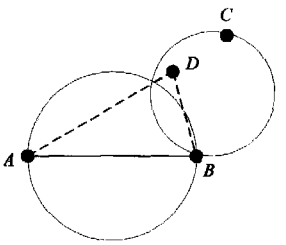


图1 Gabriel近邻

2.2 Minkowski距离与测地距离

Minkowski距离[9]是很常用的一种测度方式. 对x, y∈R^n这2个点, 其Minkowski距离$d_m(x, y)$定义为:

$$d_m(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^r \right)^{1/r}$$

式中: r是Minkowski因子. 当r设为2时, 就是欧式距离. 当对于复杂流形上的数据, 测地距离比Minkowski距离更适合用来表征它们之间的相互关系[2].

测地曲率处处为0的曲线称为测地线. 以曲面为例, 一般情况下经过曲面上2点的测地线是这2点间的短程线, 2点间的测地线长度为测地距离. 测地距离反映了流形的内在几何结构(intrinsic geometry), 当一个弯折的曲面被展开成一个平面时, 点与点之间的测地距离保持不变. 图2显示了测地距离(实线)与欧式距离(虚线)的比较.

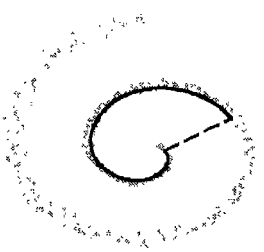


图2 测地距离与欧式距离

2.3 训练样本与测试样本的测地距离估计

流形上测地距离估计的基本思路是: 流形上相邻点之间的测地距离可以直接由欧式距离来估计, 非相邻点之间的测地距离则是通过一系列相邻点之间的最短路径来估计.

首先根据某一测度方式, 如欧式距离$d_X(x_i, x_j)$, 确定训练样本x_i的近邻点集$Z(x_i)$. 一种方法是取距离x_i最近的k个点构成$Z(x_i)$; 另外一种方法是取位于x_i的ε领域内的点构成$Z(x_i)$. 本文采用了第一种方法. 样本点间的近邻关系由矩阵G表示, 其中

$$d_G(x_i, x_j) = \begin{cases} d_X(x_i, x_j) & x_j \in Z(x_i) \\ \infty & x_j \notin Z(x_i) \end{cases}$$

继而估计流形上任一数据点对之间的测地距离$g(x_i, x_j)$. 由于嵌入流形未知, 用G上x_i与x_j之间的最短路径来近似$g(x_i, x_j)$. 最短路径的估计方法如Floyd-Warshall算法[2]:

$$g(x_i, x_j) = \min\{d_G(x_i, x_j), d_G(x_i, x_k) + d_G(x_k, x_j)\},$$

由此可见: 流形上的2个数据之间的欧式距离仅取决与它们本身, 而它们的测地距离则还取决于其他数据点所构成的整体结构.

上述估计方法针对的是训练样本. 而针对给定一个测试样本x_t, 首先采用相同的规则确定其近邻点集$Z(x_t)$, 则x_t到非近邻的训练样本的测地距离采用下式估计:

$$g(x_t, x_i) = \begin{cases} \min\limits_{x_j \in Z(x_t)} \{g(x_j, x_i) + d_X(x_t, x_j)\} & x_i \notin Z(x_t) \\ d_X(x_t, x_i) & x_i \in Z(x_t) \end{cases}$$

2.4 测地 Gabriel图

当采用测地距离来描述流形上数据的相互关系的时候, 式(1)被改写为:

$$g^2(D, A) + g^2(D, B) > g^2(A, B) \qquad (2)$$

式中: g是测地距离. 则满足式(2)的点对, 称它们为测地Gabriel近邻(Geodesic Gabriel Neighbor GGN), 所有GGN之间以边相连接所构成的图, 称之为测地Gabriel图(Geodesic Gabriel Graphic GGG). 为了便于比较, 将基于欧式距离的Gabriel近邻以及Gabriel图分别称为欧式Gabriel近邻(Euclidean Gabriel Neighbor EGN)、欧式Gabriel图(Euclidean Gabriel Graphic: EGG). 相比于其他区域, 临界区域在判别分析时更值得关注; 而临界区域的划分与构造Gar

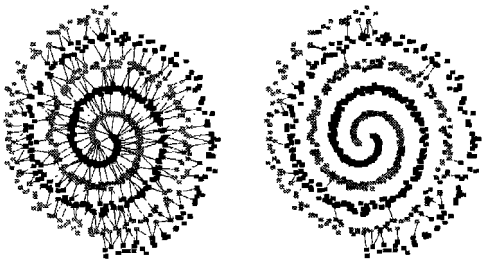


图3 EGG与GGG的比较

briel图时所采用的测度方式相关. 图 3给出了 EGG(左)与 GGG(右)的一个实例上的比较, 其中每对 Gabriel近邻用实线相连. 此例的数据点位于复杂的非线性流形上; 数据点间的测地距离一定程度上包含了此流形的整体结构信息, 而欧式距离则无此能力. 正是由于此原因, GGG 比 EGG 对临界区域作了更简洁有效的划分.

3　基于 GGG 的局部判别分析

在异类数据 Gabriel近邻所确定的临界区域进行局部判别分析. 首先根据流形上样本数据到 GGN 的测地距离, 估计属于对应的局部判别器的概率, 然后采用加权正则化线性判别分析的方法构造局部判别器.

3.1　确定流形上数据属于局部分类器的概率

假设 $\{(x_1^k, x_2^k) \mid k = 1, 2, \cdots, K\}$ 为 GGN 集, 近邻对 (x_1^k, x_2^k) 对应的局部判别器为 f_k. 给定流形上的数据 x, 其属于局部判别器的概率 $p(f_k \mid x)$ 可采用下式获得[10]:

$$p(f_k \mid x) = p^k(x) \Big\backslash \sum_{j=1}^{K} p^j(x)$$

式中: $p^k(x) = \exp(-\beta^k(x))$, $\beta^k(x)$ 是样本 x 相对于局部分类器 f_k 的活跃信号量(activity signal). 根据样本与 GGN 之间的测地距离来估计 $\beta^k(x)$:

$$\beta^k(x) = \frac{(g(x, x_1^k) + g(x, x_2^k))^2}{t}$$

式中: t 是宽度常数. 样本数据与 GGN 在流形上相距越远, 即 $g(x, x_1^k) + g(x, x_2^k)$ 越大, $\beta^k(x)$ 也越大, 则样本数据属于对应局部判别器的概率 $p(f_k \mid x)$ 就越低.

3.2　基于测地距离的加权正则化线性判别分析

构造局部判别器 f_k 时, 样本数据到 GGN 的测地距离构成了新的特征空间:

$$x_i \rightarrow \boldsymbol{z}_i = [g(x_i, x_1^k), g(x_i, x_2^k)]^{\mathrm{T}}$$

给定当前的样本的概率 $\{p(f_k \mid x_i)\}$, 则类内离散度为:

$$\boldsymbol{S}_{\mathrm{w}} = \boldsymbol{w}^{\mathrm{T}} \left(\sum_{j=1,2} (\boldsymbol{Z}^j - \bar{\boldsymbol{Z}}^j)(\boldsymbol{Z}^j - \bar{\boldsymbol{Z}}^j)^{\mathrm{T}} \right) \boldsymbol{w} = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{M}_{\mathrm{w}} \boldsymbol{w}$$

式中: $\boldsymbol{w}$ 是投影映射矩阵; $\boldsymbol{Z}^j \in \mathrm{R}^{2 \times N_j}$ (N_j 是 j类样本的个数), $\boldsymbol{Z}^j$ 的第 i个列向量为 $\boldsymbol{Z}^j(i) = \boldsymbol{z}_i \cdot \sqrt{p(f_k \mid x_i)}$; $\bar{\boldsymbol{Z}}^j$ 的每一列向量是 j类样本的加权均值.

类间离散度为:

$$\boldsymbol{S}_{\mathrm{B}} = \boldsymbol{w}^{\mathrm{T}} \left(\sum_{j=1,2} N_j (\bar{\boldsymbol{Z}}^j - \bar{\boldsymbol{Z}})(\bar{\boldsymbol{Z}}^j - \bar{\boldsymbol{Z}})^{\mathrm{T}} \right) \boldsymbol{w} = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{M}_{\mathrm{B}} \boldsymbol{w}$$

式中: $\bar{\boldsymbol{Z}}$ 是全部样本的加权均值.

传统的 Fisher线性分析通过求解 $\boldsymbol{M}_{\mathrm{B}}$ 与 $\boldsymbol{M}_{\mathrm{W}}$ 的广义特征值问题获取投影映射矩阵, 使得 $\boldsymbol{S}_{\mathrm{B}} / \boldsymbol{S}_{\mathrm{W}}$ 最大. 该方法的一个缺点是难以解决小样本(Small Sample Size: SSS)问题[11]. 考虑到实际应用中可能会遇到数据分布稀疏的情况, 采用了正则化线性分析的方法[11]:

$$\arg\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{M}_{\mathrm{B}} \boldsymbol{w}}{\lambda \boldsymbol{w}^{\mathrm{T}} \boldsymbol{M}_{\mathrm{B}} \mathrm{w} + \boldsymbol{w}^{\mathrm{T}} \boldsymbol{M}_{\mathrm{W}} \boldsymbol{w}} \qquad (3)$$

式中: $0 \leqslant \lambda \leqslant 1$ 称为正则化参数. 通过对 $\boldsymbol{M}_{\mathrm{B}}$ 以及 $\boldsymbol{M}_{\mathrm{W}}$ 的转化矩阵的特征值分解来求解式(3)的优化问题, 从而克服了小样本问题. 详细的算法可参考文献[11]. 则局部分类器为:

$$f_k(x_i) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{z}_i$$

4　局部判别器融合算法

在局部判别分析的基础上, 通过融合多个局部线性判别器, 可获得整体上非线性的判别器. 整体判别器的形式为:

$$F(x) = \sum_{k=1}^{K} \alpha_k p(f_k \mid x) f_k(x) \qquad (4)$$

式中: f_k 是第 k 个局部判别器, α_k 是 f_k 在整体判别器中所占的权重系数. 局部判别器的融合学习中, 根据一定的全局优化指标, 确定每个局部分类器的权重系数.

Fisher准则函数可以用来作局部判别器融合的全局优化指标, 但是数据得以最好区分的前提是: 每个类的样本数据的投影值为高斯分布[12]. 对于所关注的非线性流形上的数据, 如人脸图像等, 不一定就能满足这个前提条件. 而基于边界(margin)的判别优化指标[13], 关注样本的边界参数(如与分类面间的距离), 判别分析的结果与投影数值的分布情况无关. 本文的融合算法的基本思想是: 调整各个局部判别器的权重系数, 优先扩大距分类面最近的那部分样本的边界; 对应的优化指标是:

$$\arg\min_{\alpha} \sum_{i} \exp(-y_i F(x_i)) \qquad (5)$$

式中: F 是整体判别器; 该优化指标是一种柔性边界(soft margin).

优化算法流程由表 1给出. 在每次迭代过程中, 采用加权正则化线性判别分析的方法获取当前样本权重分配情况下的每个局部判别器的权重系数; 离分类面比较近的样本在下一次迭代过程中增加其权重. 经过多次迭代, 整体上逐渐地拉大样本与分类面之间的距离.

表 1　局部判别器融合算法流程

给定训练样本 $\{x_i \mid i = 1, 2, \cdots N\}$:

(1) 样本权重的初始设定 $w_i = 1/N$, $i = 1, 2, \cdots N$
迭代过程 $m = 1, 2, \cdots M$

(2) (a)在当前样本的权重分配情况下, 采用加权正则化线性判别方法获取各个局部判别器的权重系数 $\{\alpha_k^m\}$, 根据式(4)构造判别器 $F_m(x)$.
(b)更新样本权重 $w_i \leftarrow w_i \exp(-y_i F_m(x_i))$, $i = 1, 2, \cdots N$, 然后归一处理使得 $\sum_i w_i = 1$.

(3) 最终的权重系数为 $\alpha_k = \sum_m \alpha_k^m$.

5　实验

分别在人工合成的数据集以及真实的人脸图像数据库上进行相应的实验来验证本文算法.

5.1　合成数据集上的实验

该数据集上正例样本与反例样本数据相互缠绕, 整体上呈现较强烈的非线性特征, 如图 3所示; 图上亦给出了相应的 Gabriel图. 根据 Gabriel图划分局部区域, 进行局部线性判别分析. 实验中对 EGG 与 GGG、基于 Fisher准则函数与基于边界准则函数的融合算法进行了比较.

采用 EGG 划分临界区域时, 对应的 EGN 的个数是 264; 而采用 GGG 时, 对应的 GGN 个数为 37, 仅为 EGN 个数的 14%. 这项数据的比较, 也证明了 GGG 比 EGG 更适合用来提取此例中的非线性流形的结构信息.

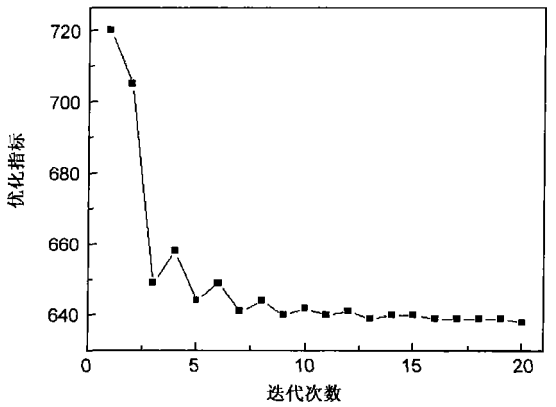


图 4 迭代过程的优化指标值

本质上基于 Fisher准则函数的融合算法可视为基于边界准则函数融合算法的一次迭代. 图 4给出了基于边界准则函数融合算法迭代过程中公式(5)所示的优化指标的数值变化情况. 可以看到, 随着迭代次数的增大, 优化指标总体上呈下降趋势.

图 5给出了训练样本数据的判别结果, 可以看到, GGG 与边界准则函数相结合方法的结果优于其他两种算法. GGG 对流形数据整体结构的把握, 以及边界准则函数对异类数据的有效区分, 均有助于对复杂流形的判别分析.

5.2 人脸识别实验

人脸识别需要处理诸如光照、姿态与表情等各种变化因素; 这些变化因素对实际人脸图像的影响往往是非线性的, 由此得到的人脸图像集可视为位于高维空间的非线性子流形上.

在 Yale[14], Umist[15] 人脸图像库上, 进行了比较实验. 其中 Yale数据库的图像在光照与表情上有比较大的变化; 而 Umist数据库的图像则覆盖了从正面到侧面的各个姿态角度. 图 5 给出了部分人脸样本图像, 其中第一行来自 Yale数据库, 第二行来自 Umist数据库. Yale数据库上包含了 11个对象的 165张图像. 在原始图像上剪切出人脸图像部分, 然后降采样至 28(高) × 24(宽)大小. 随机选择每个对象的 7 张图像为训练样本, 其他 8张为检测样本. UMIST人脸库包含了 20个对象的 564张图像. 首先我们对图像进行挑选, 每个对象选出 10张, 每张都对应一个特定的角度. 随机选择每个对象的 5张图像作为训练样本,

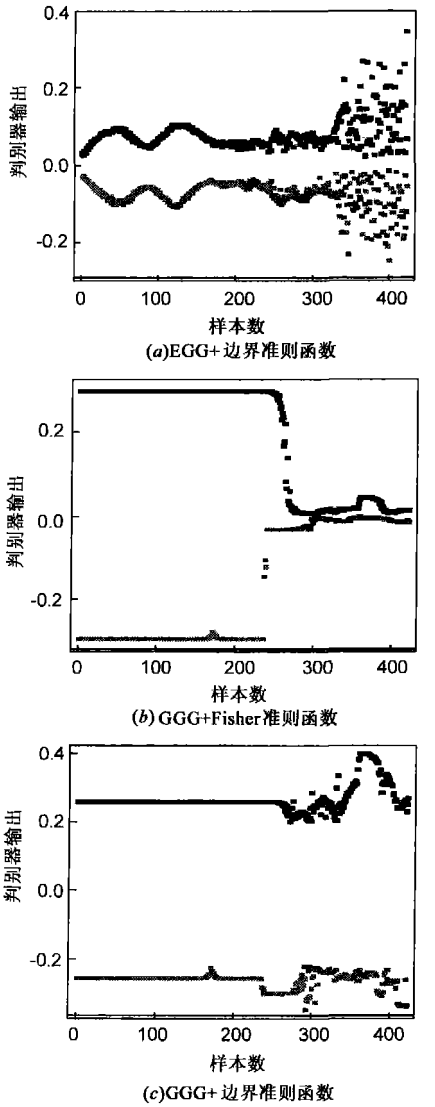


图 5 合成数据集上算法比较

另外 5张图像作为测试样本. 缩放后的图像大小为 28(高) × 23(宽).

图 6 部分人脸图像样本

与本文算法进行比较的算法包括了 Eigenface, Fisherface, Isomap与 LLE. Eigenface与 Fisherface是传统的子空间方法, 将图像数据投射到低维子空间, 是一种线性的处理方法. 而包括 Isomap, LLE 与本文算法在内的流形学习算法则直接采用了非线性的处理方法. Eigenface, Isomap与 LLE采用了无导师学习的方式, 而 Fisherface与本文算法采用了有导师学习的方式. 其中, Fisherface, Isomap, LLE 采用近邻法对降维后的数据进行分类. 经过多次实验比

较,选取最优的实验参数与检测结果.

实验结果在表 2,表 3中给出. 可以看到, 采用有导师学习的 Fisherface与本文算法的检测精度较高; 采用非线性处理方式的本文算法的检测精度高于采用线性处理方式的 Fisherface方法.

表 2　Yale数据库上的实验结果

算法	错误率(%)
Eigenface	20.45
Fisherface	13.64
Isomap, k= 20	18.18
LLE, k= 25	19.32
本文算法	9.09

表 3　UMIST数据库上的实验结果

算法	错误率(%)
Eigenface	23.0
Fisherface	14.0
Isomap, k= 20	21.0
LLE, k= 20	22.0
本文算法	9.00

6　结语

本文提出了一种基于测地 Gabriel图的局部线性判别器融合的方法对非线性流形上的数据进行判别分析. 对于复杂流形上的数据,较之 Minkowsk距离, 测地距离能更好地刻画流形之整体结构, 基于此构造的测地 Gabriel图对流形作了有效地划分, 将整体上的非线性判别分析, 转化为局部的线性判别分析. 基于柔性边界准则函数的迭代优化融合算法, 不依赖于样本数据的分布, 能够有效地将局部线性判别器整合为整体的非线性判别器.

参考文献:

[1] Roweis S T, et al. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500): 2323-2327.

[2] Joshua B, et al. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(5500): 2319-2323.

[3] Belkin M, et al. Laplacian eigenmaps for dimensionality reduction and data representation[J]. Neural Computation, 2003, 15(6): 1373-1396.

[4] Wu Y M, et al. An extended isomap algorithm for learning multi-class manifold[A]. IEEE International Conference on Machine Learning and Cybernetics[C]. Shanghai: IEEE Computer Society, 2004, 6: 3429-3433.

[5] Chen H T, et al. Local discriminant embedding and its variants[A]. Computer Vision and Pattern Recognition[C]. San Diego: IEEE Computer Society, 2005, 2: 846-853.

[6] Geng X, et al. Supervised nonlinear dimensionality reduction for visualization and classification[J]. IEEE Trans on Systems, Man, and Cybernetics-Part B: Cybernetics, 2005, 35(6): 1098-1107.

[7] Zhang Z Y, et al. Principal manifolds and nonlinear dimension reduction via local tangent space alignment[A]. 4th International Conference on Intelligent Data Engineering and Automated Learning[C]. Hong Kong: Lecture Notes in Computer Science, 2003. 477-481.

[8] Zhang W, et al. A study of the relationship between support vector machine and Gabriel graph[A]. International Joint Conference on Neural Networks[C]. Hawaii, 2002, 1: 239-244.

[9] Batchelor Pattern recognition: ideas in practice[M]. New York: Plenum Press, 1978: 71-72.

[10] Roweis S, et al. Global coordination of local linear models[J]. Advances in Neural Information Processing System, 2001, 14(2): 889-896.

[11] Lu J W, et al. Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition[J]. Pattern Recognition Letter, 2005, 26(2): 181-191.

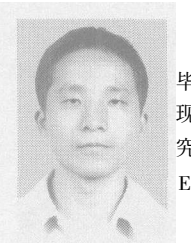
[12] Kim T K, et al. Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2005, 27(3): 318-327.

[13] Schapire R, et al. Boosting the margin: a new explanation for the effectiveness of voting methods[J]. Annals of Statistics, 1998, 26(5): 1651-1686.

[14] Yale Univ. Face Database[DB/OL]. http://cvc.yale.edu/projects/yalefaces/yalefaces.html, 2005.

[15] Graham D. B, et al. Characterizing virtual eigensignatures for general purpose face recognition[A]. Face Recognition: From Theory to Applications[C]. Berlin: NATO ASI Series F, Computer and Systems Sciences, 1998, 163: 446-456.

作者简介:



陈华杰　男, 1978年生于福建闽侯. 2001年毕业于浙江大学工业自动化专业, 获学士学位. 现为浙江大学控制理论与控制工程专业博士研究生. 主要研究方向为: 模式识别、机器学习.

E-mail: binwhit@163.com