

一种基于完全 2DPCA 的二次特征选择方法

武妍, 夏莹

(同济大学计算机科学与技术系, 上海 200092)

摘要: 提出一种基于完全二维主元分析(2DPCA)的二次特征选择方法用于人脸识别。该方法针对 PCA 及其改进方法的不足, 结合完全 2DPCA 方法, 用免疫算法和基于最近邻分类器的 5 阶交叉验证方法完成对人脸特征二次选择。基于 ORL 人脸数据库和 Yale 人脸数据库的实验结果表明, 该方法识别效果较好。

关键词: 人脸识别; 完全二维主元分析; 免疫算法; 最近邻分类器

Second-order Feature Selection Method Based on Complete 2DPCA

WU Yan, XIA Ying

(Department of Computer Science and Technology, Tongji University, Shanghai 200092)

【Abstract】 A second-order feature selection method based on complete 2DPCA is proposed for face recognition. The method based on complete 2DPCA uses immune algorithm and five-fold cross-validation method which is based on the nearest neighbor classifier to complete second-order selection of facial features in order to overcome the demerits of PCA and some improved PCA methods. Experiments based on ORL face library and Yale face library show that the method can achieve good recognition effect.

【Key words】 face recognition; complete Two-dimensional Principle Component Analysis(2DPCA); immune algorithm; nearest neighbor classifier

人脸识别是模式识别研究领域的重要课题,也是目前一个研究热点^[1]。由于人脸非常相似,如何有效抽取人脸图像的识别特征是解决人脸识别问题的关键。主元分析法(Principle Component Analysis, PCA)是人脸识别领域的一种经典的特征提取技术。近年来,文献[2]提出一种二维主元分析法(Two-dimensional Principle Component Analysis, 2DPCA)用于人脸识别,识别效果优于PCA。文献[3]在 2DPCA 的基础上提出了一种完全 2DPCA,识别率超过 2DPCA。但是,上述方法都是基于最小均方误差原则的特征压缩方法,因此,用提取的人脸特征来表达和重建人脸图像是最优的,但对于分类来说并非最优。

1 基于完全 2DPCA 的二次特征选择方法

1.1 本文的方法

文献[3]采用的完全 2DPCA方法分别选择 2 个协方差矩阵最大的若干个特征值所对应的特征向量构造投影矩阵,提取的特征对于分类来说不是最优的。本文依次用免疫算法和基于最近邻分类器的 5 阶交叉验证方法寻找特征向量的最优组合,完成对人脸特征的二次选择。假设 A_1, A_2, \dots, A_M 为人脸图像库中的训练样本,本方法步骤如下:

Step1 根据式(1)、式(2)计算图像的 2 个协方差矩阵 G_1 和 G_2 :

$$G_1 = \frac{1}{M} \sum_{j=1}^M (A_j - \bar{A})(A_j - \bar{A})^T \quad (1)$$

$$G_2 = \frac{1}{M} \sum_{j=1}^M (A_j - \bar{A})^T (A_j - \bar{A}) \quad (2)$$

其中, \bar{A} 为所有训练样本的平均。

Step2 分别计算 G_1 和 G_2 的特征值及其对应的特征向量。

Step3 依次用免疫算法和基于最近邻分类器的 5 阶交叉验证方法获得最优的特征向量组合,构造最优的投影矩阵 U

和 V , 从而完成对人脸特征的二次选择。

Step4 由公式 $Y_j = UA_jV$ ($j=1, 2, \dots, M$)将每个训练样本 A_j ($j=1, 2, \dots, M$)通过投影转换成特征矩阵 Y_j ($j=1, 2, \dots, M$)。

Step5 对于人脸库中的任一测试样本 X , 由公式 $Y=UXV$ 得到 X 对应的特征矩阵 Y 。

1.2 人脸特征的第 1 次选择

本文用免疫算法选择特征向量,利用免疫算法强大的全局搜索能力寻找优化的特征向量组合,完成人脸特征的第 1 次选择,具体过程如下:

Step1 编码抗体。协方差矩阵 G_1 和 G_2 分别有 m 和 n 个特征向量。要用免疫算法解决的问题是从这 m 和 n 个特征向量中分别找到 p 和 q 个特征向量,得到特征向量的最优组合。因此,对抗体采用二进制编码,就是用 $m+n$ 位的由 0 或 1 构成的字符串表示一种特征向量的组合。其中, 1 表示对应的特征向量被选中, 0 表示未被选中。

Step2 产生初始抗体群体。先随机生成 N 个抗体,再随机生成 N_m 个抗体放入记忆库中,则初始抗体群的规模为 $N + N_m$ 。在可容忍的计算复杂度情况下,一般选取 N 为 50~100, N_m 为 5~10。

Step3 对各抗体进行评价。以期望繁殖率 e_V 作为抗体评价的标准,计算步骤如下:

(1)计算抗体 V 对抗原的亲合力 ax_V ,它反映抗体对抗原的匹配程度。为了得到对分类最有效的特征向量组合,本文用多分类问题的类内类间距离来计算 ax_V ,计算公式如下:

基金项目: 国家自然科学基金资助项目(60475019)

作者简介: 武妍(1967-),女,教授、博士后、博士生导师,研究方向:神经网络,计算智能,模式识别;夏莹,硕士研究生

收稿日期: 2007-05-10 **E-mail:** yanwu@mail.tongji.edu.cn

$$ax_v = \frac{|\tilde{S}_b(\mathbf{V})|}{|\tilde{S}_w(\mathbf{V})|} \quad (3)$$

其中, $\tilde{S}_b(\mathbf{V})$ 和 $\tilde{S}_w(\mathbf{V})$ 分别表示将训练样本 A_1, A_2, \dots, A_M 向抗体 \mathbf{V} 所对应的特征向量组合所构成的投影矩阵投影后得到的特征矩阵 B_1, B_2, \dots, B_M 的类间离散度矩阵和类内离散度矩阵。根据式(3), 投影后, 在特征空间中各类样本应尽量分开些, 即 $\tilde{S}_b(\mathbf{V})$ 越大越好, 同时各类样本内部应尽量密集, 即 $\tilde{S}_w(\mathbf{V})$ 越小越好, 因此, ax_v 的值越大, 分类效果就越好。 $\tilde{S}_b(\mathbf{V})$ 和 $\tilde{S}_w(\mathbf{V})$ 的计算过程如下:

假设训练样本集中的样本分成 C 个类别, 令 B_j^i 表示属于第 i 类的第 j 个训练样本的特征矩阵, 定义 $\bar{B}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} B_j^i$ 为属于第 i 类的样本的特征矩阵的平均, $\bar{B} = \sum_{i=1}^C p(\omega_i) \bar{B}_i$ 为全体特征矩阵的平均, 其中, N_i 表示属于第 i 个类别的样本数, $p(\omega_i)$ 为属于第 i 个类别的先验概率, 则

$$\tilde{S}_b(\mathbf{V}) = \sum_{i=1}^C p(\omega_i) (\bar{B}_i - \bar{B})(\bar{B}_i - \bar{B})^T \quad (4)$$

$$\tilde{S}_w(\mathbf{V}) = \sum_{i=1}^C p(\omega_i) \frac{1}{N_i} \sum_{j=1}^{N_i} (B_j^i - \bar{B}_i)(B_j^i - \bar{B}_i)^T \quad (5)$$

(2) 计算抗体 \mathbf{V} 的浓度 C_v , 它表示与抗体 \mathbf{V} 相似的抗体的规模, 计算公式如下:

$$C_v = \frac{1}{N + N_m} \sum_{w=1}^{N+N_m} ac_{vw} \quad (6)$$

其中, $ac_{vw} = \begin{cases} 1 & ay_{vw} \geq T \\ 0 & \text{otherwise} \end{cases}$; ay_{vw} 是用信息熵的方法得出的抗体 \mathbf{V} 和 \mathbf{W} 的相似度, 具体方法见参考文献[4]; T 为一个预先确定的阈值, 一般为 0.75~0.85。

(3) 计算抗体 \mathbf{V} 的期望繁殖率 e_v :

$$e_v = \frac{ax_v}{C_v} \quad (7)$$

Step4 生成父代抗体群, 并更新记忆库。将抗体群按 e_v 从大到小的顺序排列, 选择前 N 个抗体组成父代抗体群, 并将前 N_m 个抗体存入记忆库中。

Step5 判断是否满足终止条件, 若满足, 则结束; 否则继续下一步操作。终止条件为记忆库中的抗体不再变化。随着算法过程的进行, 记忆库中的元素最终收敛于各个峰值, 并且随着搜索过程的继续, 仍能保持收敛结果。

Step6 生成新一代抗体群。将 Step4 生成的父代抗体群中的抗体进行交叉和变异的遗传操作(交叉采用两点交叉, 交叉概率 P_c 一般取 0.75~0.85; 变异采用单点的位点变异, 变异概率 P_m 一般为 0.01~0.03), 得到新抗体, 同时取出记忆库中的抗体, 共同组成新一代抗体群。

Step7 转 Step3。

1.3 人脸特征的第 2 次选择

以上免疫算法可以得到 N_m 种优化的人脸特征选择方式。为了找到这 N_m 种方式中最优的一个, 本文用基于最近邻分类器的 5 阶交叉验证方法来完成对人脸特征的第 2 次选择。

假设 $C_1^i, C_2^i, \dots, C_M^i$ ($i=1, 2, \dots, N_m$) 是用第 i 种特征选择方式选取的训练样本 A_1, A_1, \dots, A_M 的特征矩阵的集合。当采用最近邻分类器进行交叉验证时。任意 2 个特征矩阵 C_p^i 和 C_q^i ($1 \leq i \leq N_m, 1 \leq p, q \leq M$) 之间的距离用欧几里德距离定义为

$$d(C_p^i, C_q^i) = \sqrt{\sum_{x=1}^k \sum_{y=1}^k (C_p^i(x, y) - C_q^i(x, y))^2} \quad (8)$$

假设 C_x^i 为一个用于测试的特征矩阵, 如果 $d(C_x^i, C_y^i) = \min_j d(C_x^i, C_j^i)$, 则 C_x^i 属于 C_j^i 所属的那个类别。基于最近邻分类器的 5 阶交叉验证方法如下:

Step1 $i \leftarrow 1$;

Step2 将 $C_1^i, C_2^i, \dots, C_M^i$ 平均分成 5 个子集 $Subset_j$ ($j=1, 2, \dots, 5$);

Step3 $k \leftarrow 0$;

Step4 $k \leftarrow k+1$;

取子集 $Subset_k$ 作为测试集, 其他 4 个子集合在一起作为训练集, 用最近邻分类器进行测试, 并计算识别率 $RecogRate_k^i$;
until $k=5$

Step5 计算平均识别率 $RecogRate^i = \frac{1}{5} \sum_{j=1}^5 RecogRate_j^i$;

Step6 $i \leftarrow i+1$, 若 $i = N_m+1$, 则继续下一步操作; 否则转 Step2;

Step7 比较几个 $RecogRate^i$ ($1 \leq i \leq N_m$), 取其中的最大值, 则它所对应的特征选择方式是最优的, 用这种方式提取的人脸特征也是最有利于分类的。

2 实验结果及分析

本文采用标准的 ORL 人脸数据库和 Yale 人脸数据库进行实验。ORL 人脸库有 400 幅分辨率为 92×112 的人脸图像, 具体为 40 个人, 每人 10 幅图像。Yale 人脸库中共有 165 幅分辨率为 320×243 的不同表情、遮掩和光照条件下的人脸图像, 具体为 15 个人, 每人 11 幅图像。对于 2 个人脸库, 均从每个人的几幅人脸图像中随机选择 k ($k=4, 5, 6$) 幅图像作为训练样本, 其余的图像作为测试样本, 且对于每一个 k 值, 分别进行 30 次不同的实验, 并算出平均识别率。实验中先对所有的人脸图像进行灰度分布标准化^[5], 将图像的灰度均值和灰度方差调整到给定值, 以消除一些光照的影响, 再分别用 PCA、2DPCA、完全 2DPCA 和本文的方法提取人脸图像的特征, 最后用最近邻分类器进行识别。实验还记录了 $k=5$ 的情况下每次实验所需 CPU 时间, 并计算出了 4 种方法所需的平均 CPU 时间。实验结果如表 1、表 2 所示。本方法中的一些参数设置为: 抗体群规模为 $55(N=50, N_m=5)$; 浓度阈值 $T=0.8$; 交叉概率 $P_c=0.8$; 变异概率 $P_m=0.03$ 。

表 1 ORL 人脸数据库上 4 种方法识别效果的比较

识别方法	识别率(%)			$k=5$ 时平均 CPU 时间/s
	$k=4$ 时	$k=5$ 时	$k=6$ 时	
PCA	81.89	87.12	88.64	18.32
2DPCA	89.45	90.94	92.12	7.86
完全 2DPCA	90.83	92.11	93.25	2.95
本文的方法	92.56	93.75	95.19	4.28

表 2 Yale 人脸数据库上 4 种方法识别效果的比较

识别方法	识别率(%)			$k=5$ 时平均 CPU 时间/s
	$k=4$ 时	$k=5$ 时	$k=6$ 时	
PCA	83.39	88.24	89.53	24.65
2DPCA	90.71	91.85	93.08	9.58
完全 2DPCA	92.36	93.48	95.16	3.72
本文的方法	94.68	95.91	97.22	5.39

用本方法选取的特征向量最优组合中, 一些特征向量往往不是那些最大特征值所对应的特征向量, 对于分类却更为 (下转第 227 页)