

融合过滤和封装方式的特征选择算法

代 旺, 方昱春, 李 杨

(上海大学计算机工程与科学学院, 上海 200072)

摘 要: 已有特征选择算法不能有效降低特征维数, 且稳定性较低。为此, 提出一种融合过滤和封装方式的特征选择算法。在封装式算法中, 设计能保持图像之间拓扑结构的特征选择判据, 在过滤式算法中, 以 Fisher Score 为判据, 采用单独最优的特征搜索策略。实验结果表明, 将算法应用于人脸识别中, 能提高识别率, 降低特征维数, 且具有较好的稳定性。

关键词: 特征选择; 过滤式方法; 封装式方法; Fisher Score 判据; 人脸识别; 降维

Feature Selection Algorithm Fused with Filtering and Packaging Mode

DAI Wang, FANG Yu-chun, LI Yang

(School of Computer Engineering and Science, Shanghai University, Shanghai 200072, China)

【Abstract】 The feature selection algorithm can not effectively reduce the feature dimension, and the stability is lower. In order to solve this problem, this paper proposes a feature selection algorithm of fusing filtering and packaging mode. In packaging type algorithm, it designs the feature selection criterion which can maintain topological structure between image, uses Fisher Score as criterion in the filtering type algorithm, and the individual optimum search strategy is used in this paper. Experimental results show that this algorithm can improve recognition rate, reduces the feature dimension, and has good stability in face recognition application.

【Key words】 feature selection; filtering type method; packaging type method; Fisher Score criterion; face recognition; dimension reduction

DOI: 10.3969/j.issn.1000-3428.2012.24.039

1 概述

随着人脸识别在身份认证、人流监控、安检、人员查找、支付手段等方面的广泛应用, 人脸识别技术已经成为模式识别和图像处理方面的热点。文献[1]提出基于 Gabor 变换和局部二值模式(Local Binary Pattern, LBP)的特征提取方法——局部 Gabor 二值模式直方图序列(Local Gabor Binary Pattern Histogram Sequence, LGBPHS), 近年来被证明是人脸识别领域最有效的特征提取算法之一, 然而, LGBPHS 特征过高的维数不仅在计算的时间和存储空间上会降低系统的性能, 而且对大多数分类器来说, 在训练样本数量不变的情况下, 特征维数的增加会使分类器的参数估计的可靠性下降, 进而直接影响学习算法的性能和效率, 这就是所谓的“维数灾难”问题^[2]。

特征选择是模式识别中降低特征维数的方法之一。它依据某个准则从众多原始特征中选择部分最能反映模式类别的统计特性的相关特征^[3]。与引入空间变换的特征抽取算法相比较, 通过特征选择获得的特征更能反映原始输入图像的物理特性。

从结构上讲, 特征选择算法包括判据的定义和搜索策略的选取。判据是对所选特征子集好坏的判断准则的定义, 搜索策略就是采取何种策略算法从原始特征集上生成特征子集。从特征选择时如何判断与何时判断特征子集的好坏的角度, 特征选择方法大致可分为 3 类: 过滤式, 封装式和嵌入式^[4]。过滤式方法独立于具体的学习模型, 直接从分析数据的特性来给出判据准则; 封装式方法需要预设学习算法, 并在训练过程中以此学习模型作为判据进行特征选择; 嵌入式方法将特征选择包含在模型适应/训练的过程之中, 并以最优化学习模型的目标函数作为特征选择的判据。

与封装式和嵌入式方法相比较, 过滤式方法具有结构简单、训练速度快、独立于具体训练模型、易于设计和理解等优点, 这也是大多数特征选择算法是过滤式方法的原因。大部分过滤式方法都是单独最优的训练方法, 常被采用的判据有: Fisher Score, Laplacian Score^[5], 信息增益, 互信息^[6]等。文献[7]在特征之间相关性与冗余性的研究中指出: 单独最优的特征选择算法在去除不相关特征中比较

基金项目: 国家自然科学基金资助项目(61170155); 上海市重点学科建设基金资助项目(J50103); 上海大学研究生创新基金资助项目(SHUCX1 12152)

作者简介: 代 旺(1988—), 男, 硕士, 主研方向: 图像处理, 模式识别; 方昱春, 副研究员、博士; 李 杨, 硕士

收稿日期: 2012-02-28 **修回日期:** 2012-03-27 **E-mail:** ycfang@shu.edu.cn

有效,但不能有效减小特征之间的冗余。而基于特征子集搜索的封装式方法考虑到了特征维之间的相互作用,能有效去除特征之间的冗余。

针对已有算法存在的问题,本文提出一种将过滤式和封装式相结合的特征选择算法。综合过滤式方法训练速度快、结构简单以及封装式方法有效去除特征之间冗余的优点,在封装式方法部分,提出一种基于保持图像之间拓扑结构(Preserving Images Structure, PIS)的特征选择判据。

2 LGBPHS 特征提取方法

研究发现,对二维人脸图像进行 Gabor 变换能够抓住图像局部区域内多个方向的空间频率和局部结构特征,这相当于增强了人脸面部的关键部位如眼睛、鼻子、嘴巴等部位的信息,从而使得在总体上保留人脸信息的同时增强局部特性成为可能^[8]。LGBPHS 特征提取过程包括 4 个步骤:

(1)对输入图像做 Gabor 变换

首先对人脸图像做标准化处理,然后用多尺度、多方向的 Gabor 滤波器对输入图像进行 Gabor 变换,得到多个 Gabor 振幅图像(Gabor Magnitude Pictures, GMPs)。具体就是用 Gabor 滤波器 $\psi_{\mu,\nu}(z)$ 与输入人脸图像 $f(x,y)$ 做卷积:

$$G_{\psi f}(x,y,\mu,\nu) = f(x,y) * \psi_{\mu,\nu}(z) \quad (1)$$

其中, * 代表卷积。

Gabor 滤波器 $\psi_{\mu,\nu}(z)$ 定义如下:

$$\psi_{\mu,\nu}(z) = \frac{\|k_{\mu,\nu}\|^2}{\sigma^2} e^{-(\|k_{\mu,\nu}\| - 1)^2 / 2\sigma^2} [e^{ik_{\mu,\nu}z} - e^{-\sigma^2/2}] \quad (2)$$

其中, μ 和 ν 分别代表 Gabor 滤波中的方向和尺度; $z = (x,y)$, $\|\cdot\|$ 是范数运算。波矢量 $k_{\mu,\nu} = k_{\nu} e^{i\phi_{\mu}}$, $k_{\nu} = k_{\max} / \lambda^{\nu}$, $\phi_{\mu} = \pi\mu/8$; λ 是核函数距离间隔因子。

(2)用 LBP 算子^[9]对每个 GMPs 处理,得到局部二值模式 Gabor 图。

(3)将每个 Gabor 图分成大小相同的矩形块,统计每个块内的二值模式分布直方图^[9]。

(4)将每个子图的每块直方图序列拼接在一起作为描述人脸图像的特征。

针对 LGBPHS 特征维数过高的问题,比较经典的方法是文献[10]提出的 GFC 方法:先对特征进行均匀下采样,然后进行主成分分析和线性判别分析。尽管一定程度上解决了维数问题,但简单的下采样会导致大量判别特征的丢失,从而使得分类精度下降。本文提出一种融合的特征选择算法解决 LGBPHS 特征维数过高的问题。

3 融合过滤式和封装式特征选择算法

本文模式识别系统流程包括 4 个部分,即图像输入、特征提取、特征选择和分类识别,其中,特征选择的训练部分是离线进行的,融合特征选择系统流程如图 1 所示。

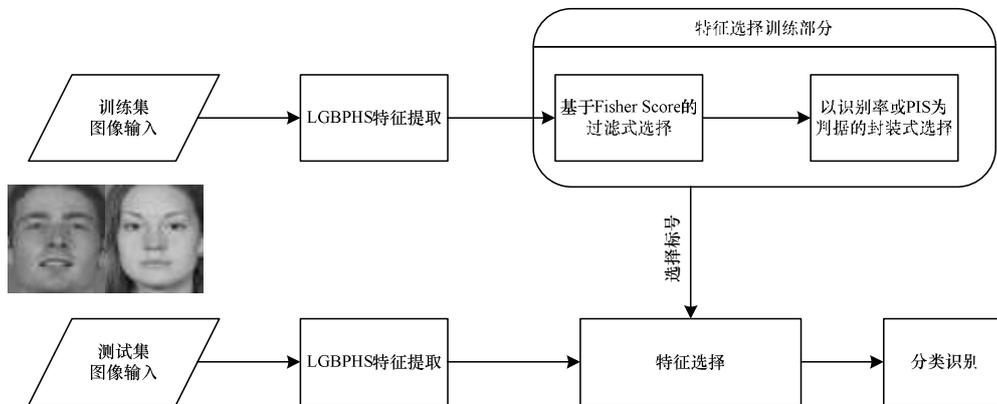


图 1 融合特征选择系统流程

在图 1 中,特征选择训练部分以识别率为判据是封装式特征选择算法中比较常用的做法,本文用的是基于最近邻分类器的识别率。给定类别标号已知的数据集: $\{x_i, y_i\}_{i=1}^n, y_i \in \{1, 2, \dots, c\}$, 其中, x_i 为 N 维特征向量; 第 i 类的样本个数为 n_i 。特征空间上的第 r 列(即每个样本的第 r 维特征组成的特征向量)记为 $f_r (r=1, 2, \dots, N)$ 。

3.1 Fisher Score 判据

以 Fisher 准则为判据的线性鉴别方法充分利用训练样本的类别信息,以最大化类间离散度和最小化类内离散度为目标,因为能获取更具有分类鉴别性质的数据而被广泛应用在模式识别各个领域。本文将 Fisher 准则作为过滤式特征选择算法的判据,具体定义如下:

(1)对于第 r 维特征: α_i 和 β_i^2 分别是第 i 类的均值和方

差, $i=1, 2, \dots, c$; α 和 β^2 分别是整个数据集的均值和方差。

基于 Fisher 准则的过滤式特征选择算法的判据为:

$$J_r = \frac{\sum_{i=1}^c n_i (\alpha_i - \alpha)^2}{\sum_{i=1}^c n_i \beta_i^2} \quad (3)$$

(2)以 J_r 为判据值的单独最优的过滤式特征选择过程为:计算各个单维特征的判据值 $J_r (r=1, 2, \dots, N)$, 对判据值加以排序(降序),取前 d 个判据值对应的特征维数标号作为选择结果。

该算法的优点是计算效率高和分类效果较好,但是基于单独最优的特征选择算法没有考虑到特征维之间的相关性,文献[7]在特征之间相关性与冗余性的研究中发现:单独最优的特征选择算法在去除不相关特征中比较有效但不能有效去除特征之间的冗余。因此,在过滤式特征选

择基础上,本文选取考虑到特征之间相关性的封装式特征选择算法,从而起到了互补的效果。

3.2 保持图像之间拓扑结构的特征选择判据

近年来,保局投影(Locality Preserving Projection, LPP)方法^[11]在特征降维上取得了较大成功。保局投影方法的基本思想是在保持数据集的样本间局部邻域结构信息的同时,降低数据集空间的维数。本文从这一思想出发,尝试从特征选择的角度达到这一目标。首先,定义图像之间的拓扑结构矩阵 Q :

$$Q_{ij} = \begin{cases} 1 & \text{若 } x_i, x_j \text{ 属于同一类} \\ 0 & \text{若 } x_i, x_j \text{ 不属于同一类} \end{cases}$$

矩阵 Q 的意义是:

在特征选择过程中,若一个被选中的特征子集形成的图像之间的拓扑结构越接近 Q , 认为该特征子集越好。对于 m 维已选特征子集为 $S_{\text{selected}} = \{f_{i_1}, f_{i_2}, \dots, f_{i_m}\}$, 当选择第 $m+1$ 维特征时($r = t_{m+1}$), 将在选中特征集下形成的特征之间的拓扑结构和原始特征集下拓扑结构矩阵 Q 之间的相似度(即特征选择的判据)定义为(取 J_{PIS} 的最小值):

$$J_{\text{PIS}}(r) = \frac{\sum_{i,j} \text{Dis}_{(S_{\text{selected}} \cup f_r)}(i,j) * Q_{ij}}{J_r} \quad (4)$$

其中, * 代表卷积; J_r 是第 r 列特征类间和类内方差的比值(参见式(3)), 加入 J_r 的目的在于让同类样本尽量聚集而不同类样本尽量远离, 另外, 可以避免选中某一列全为 0 或全相等的特征(如图像的公共背景区域); $\text{Dis}_{(S_{\text{selected}} \cup f_r)}(i,j)$ 代表在已选特征子集和第 r 维结合的特征子集上样本 i 和 j 之间的距离, 这里取绝对值距离, 即: $\text{Dis}_{(S_{\text{selected}} \cup f_r)}(i,j) = \sum_{k=i_1, i_2, \dots, i_m} |f_{ki} - f_{kj}| + |f_{ri} - f_{rj}|$, 由 $\text{Dis}(i,j)$ 的对称性, 判据可约简为:

$$J_{\text{PIS}}(r) = \frac{\sum_{i=1}^n \sum_{j < i} \text{Dis}_{(S_{\text{selected}} \cup f_r)}(i,j) * Q_{ij}}{J_r} \quad (5)$$

3.3 封装式特征选择算法

特征选择的目的是从 N 维的特征中选择出 $d(d < N)$ 维的特征子集, 在满足降低维数的同时, 使判据值最大。

设原始特征集为 $\bigcup_{n=1,2,\dots,N} S_n$, 已选入 m 维特征的已选特征子集为 S_{selected} , S_{selected} 初始为空, 具体算法步骤如下:

步骤 1 依次用顺序前进法^[7]从剩余特征子集 $\bigcup_{n=1,2,\dots,N} S_n - S_{\text{selected}}$ 中选 L 维特征(每一步都是以 3.2 节提出的 PIS 判据值最大为约束条件):

$$S_{i_1}, S_{i_2}, \dots, S_{i_L}, S_{\text{selected}} = S_{\text{selected}} \cup \{S_{i_1}, S_{i_2}, \dots, S_{i_L}\}, m = m + L$$

步骤 2 依次用顺序后退法(步骤和顺序前进法相反)从已选特征子集 S_{selected} 中剔除 R 维:

$$S_{i_1}, S_{i_2}, \dots, S_{i_R}, S_{\text{selected}} = S_{\text{selected}} - \{S_{i_1}, S_{i_2}, \dots, S_{i_R}\}, k = k - R$$

如果 $k = d$, 停止迭代, 否则, 转至步骤 1。

增 L 减 R 法是顺序前进法和顺序后退法的推广和补充, 它克服了顺序前进法对已选特征不能剔除的缺陷和顺序后退法对已剔除特征不能再选中的缺陷。

4 实验结果与分析

4.1 数据集介绍

本文实验的原始 LGBPHS 特征维数是 132 160 维。随着人脸识别技术的不断进步, 人脸识别系统的错误率在 FRVT(2002)上的测试结果显著减小, 为了给人脸识别的研究人员提供更具挑战性的问题, FRGC 的设计者们给出了 FRGC 图像库和相应的富有挑战性的问题。本文图像库就是取自 FRGC 的正面 2D 人脸图像库的一个子集, 选取了 459 个人, 每人随机选取了 6 幅图像(包括光照控制和非控制的), 共 2 754 幅图像。在特征选择中, 选取其中 200 个人的 1 200 幅图像作为训练集, 259 个人的 1 554 幅图像用于测试。实验特征是用 3.1 节介绍的 LGBPHS 特征提取方法提取的特征, 本文特征提取参数取 5 个尺度、8 个方向, 即: $\nu \in \{0, 1, \dots, 4\}, \mu \in \{0, 1, \dots, 7\}, \lambda = \sqrt{2}$, 经 Gabor 滤波后, 每幅图像得到 40 幅 Gabor 子图 GMPs, 原始特征维数为 132 160 维。实验用图举例如图 2 所示。



图 2 实验用图举例

4.2 非融合与融合算法实验结果比较与分析

本文比较实验的 4 个方法定义如下:

方法 1 以 Fisher Score 为判据, 以单独最优为搜索策略的过滤式算法。

方法 2 以 PIS 为判据, 以增 L 减 R 为判据的封装式算法。

方法 3 先用方法 1 做初步选择, 再用以识别率为判据, 以增 L 减 R 法为搜索策略的封装式算法。

方法 4 先用方法 1 做初步选择, 再用方法 2 做进一步选择(即本文算法)。

在原始特征集上, 直接用方法 1 进行特征选择, 测试结果如图 3 所示。

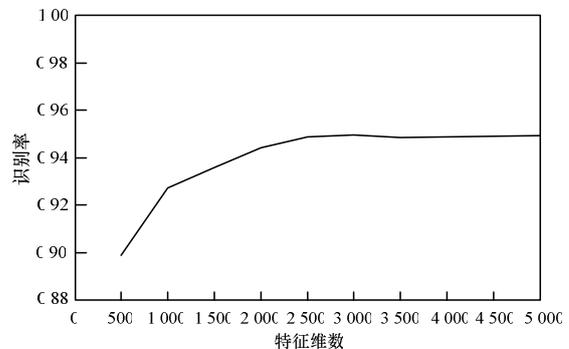


图 3 方法 1 的特征维数和识别率关系

由图 3 可知, 当用方法 1 选择出 3 000 维时, 以后增加的特征对识别率的增加没有显著贡献。因此, 本文在方法 3 和方法 4 的第 1 个步骤中, 先用方法 1 在原始特征集

上初步选出 3 000 维, 然后再用方法 3 和方法 4 的其余步骤做进一步特征选择。

在特征选择过程中, 增 L 减 R 法的 $L=4, R=3$ 。L 和

R (均大于 1)的变化和减小对实验结果影响不显著, 关于 L 和 R 参数值选择参见文献[12]。4 种方法选出不同维数特征集上的识别率比较如表 1 所示。

表 1 4 种方法选出不同维数特征集上的识别率比较

方法	500 维	1 000 维	1 500 维	2 000 维	2 500 维
方法 1	0.898 327	0.927 928	0.935 650	0.945 302	0.947 876
方法 2	0.902 831	0.914 414	0.924 710	0.929 215	0.929 215
方法 3	0.906 692	0.934 363	0.943 372	0.944 659	0.945 946
方法 4	0.921 493	0.937 580	0.945 946	0.949 807	0.949 807

由表 1 可知, 单独使用方法 1 和方法 2 都不能取得很好的效果, 而方法 4 都取得了较好的效果, 特别是对特征过滤后, 方法 4 将原始特征集从 132 160 维降到 1 000 维时, 识别率大于原始特征集上的识别率, 即: 方法 4 在保持原始识别率的情况下, 可将维数降低 132 倍。

4.3 融合算法和经典方法比较

对 Gabor 特征经典的降维方法是文献[10]提出的 GFC(Gabor-fisher Classifier)方法, 该方法具体如下: 先对特征进行均匀下采样, 然后进行主成分分析(Principal Component Analysis, PCA)和线性判别分析(Linear Discriminant Analysis, LDA)。本文直接在 LGBPHS 特征上, 先用 PCA 降维, 再用 LDA 方法进一步降维。

图 4 是方法 4 和 PCA+LDA 方法(文献[10]方法)在特征降到 500 维和 1 000 维时的识别率比较。在 PCA+LDA 方法中, 先用 PCA 方法将特征降到 1 200 维(在训练中, 有 1 200 个样本, 由于 PCA 方法在小样本上的限制, 最多能降到 1 200 维), 再用 LDA 方法降到 500 维。由图 4 可以看出, 在特征降到相同维数时, 方法 4 在提高系统识别精度上优于 PCA+LDA 方法。

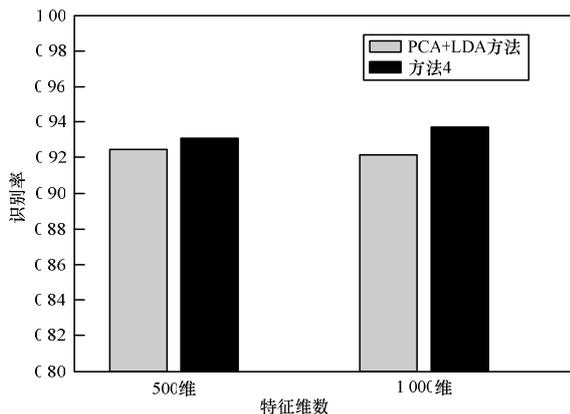


图 4 方法 4 和 PCA+LDA 方法的识别率比较

文献[13]在提出的 LGBPHS 特征基础上, 根据 LGBPHS 特征计算中各个 Gabor 子图的可分离性大小(根据 Fisher 准则计算), 进一步探讨了如何在分类器设计阶段与统计方法进行结合的问题, 提出了统计 Fisher 加权的 LGBPHS 特征匹配方法。实验结果表明, 基于 Fisher 准则的 LGBPHS 特征加权匹配方法十分有效, 但是采用加权方法进行匹配时并没有降低特征的维数, 因此, 在存储和计算上都没有减轻系统的负担。文献[13]方法与方法 4 的

识别率比较如图 5 所示。

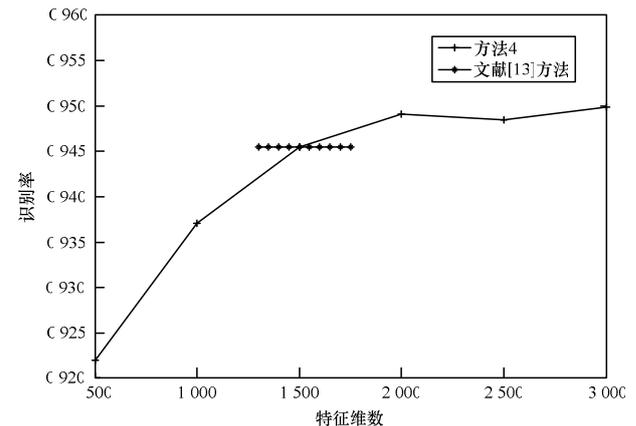


图 5 方法 4 与文献[13]方法的识别率比较

由图 5 可以看出, 其中, 文献[13]方法所用特征维数是不变的(132 160 维), 方法 4 当特征降到 1 500 维(或大于 1 500 维)时, 比文献[13]方法能获得更高的识别率。

4.4 交叉验证实验

交叉验证方法^[14]是一种常用的验证和模型选择技术, 广泛应用于统计学和机器学习中, 常用的方法是将样本数据分成若干子集, 然后在特征子集的不同组合上进行统计分析。对于特征选择问题, 训练集和测试集的不同划分会得到不同的特征选择结果和测试结果。

针对训练样本较少的问题, 本文实验结合了经典的交叉验证方法 Hold-Out 方法^[14]和 K-CV 方法^[14]的优点, 给出了适合人脸识别中特征选择问题的交叉验证方法: 对于总人数为 459 个, 每人 6 幅的图像库, 随机地将数据分为 2 个部分: 230 个人的训练集和 229 个人的测试集。这样重复训练测试 10 次。最后, 取 10 次测试集上的平均识别率为测试结果。用 4.2 节介绍的方法 1、方法 2 和方法 4 分别选择出 500 维时的交叉验证结果进行比较, 3 种方法在交叉验证实验上的平均识别率如表 2 所示, 其中, K 是分类器的参数, 当 $K=1$ 时, 是最近邻分类器。通过表 2 可以看出, 方法 4 较方法 1 和方法 2 在识别率上表现更好。

表 2 3 种方法在交叉验证实验上的平均识别率

K	方法 1	方法 2	方法 4
1	0.896	0.884	0.912
2	0.922	0.900	0.929
3	0.932	0.909	0.937
4	0.937	0.915	0.941
5	0.940	0.917	0.942

在交叉验证实验中,各个测试子集上结果的方差常被用于衡量一个模型的稳定性或适应性,即测试结果的方差是否会随着训练集的变化有较大的性能波动。3种特征选择方法的稳定性(方差)比较如图6所示。

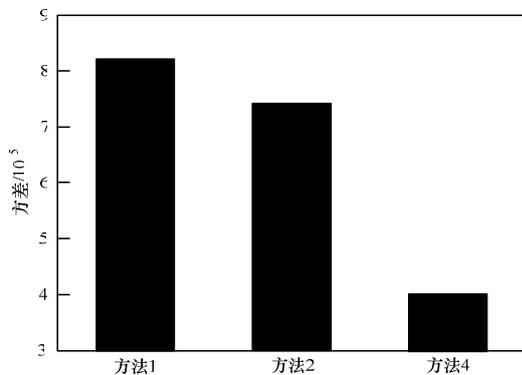


图6 3种特征选择方法的稳定性(方差)比较

由图6可知,方法4具有较好的稳定性:在不同的训练数据集上训练的结果波动不大。方法1波动最大,方法2稳定性介于方法1和方法4之间,可见PIS判据在特征选择的稳定性上表现较佳。

对于特征选择问题,除了可以用方差作为交叉验证实验中反映模型性能的指标,在每次交叉验证中,特征选择算法选中的特征维标号的交集大小也是反映特征选择算法的稳定性指标。图7给出了3种方法分别在10次交叉验证实验中每次训练选择出的特征维标号的交集比较。

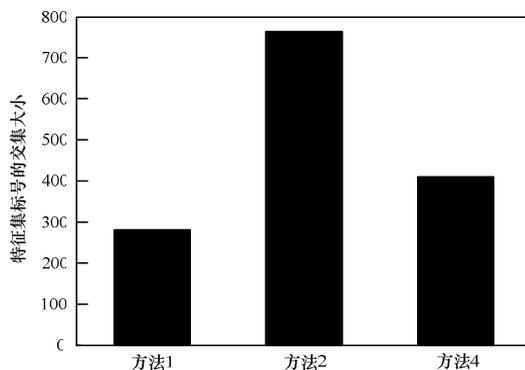


图7 3种方法在交叉验证中选特征维标号的交集比较

方法1、方法2、方法4在10次交叉验证中,特征降到1000维时,交集大小分别为279、764、434。其中,方法2得到的交集最大,且明显好于方法1,这说明PIS判据在特征选择中具有十分稳定的特点,这也从另一个方面有效证明了PIS判据的有效性。

5 结束语

本文提出一种过滤式和封装式相结合的特征选择算法。在封装式特征选择部分,给出保持图像之间拓扑结构的特征选择判据。实验结果表明,该算法在提高识别率的同时,可以大幅降低特征的维数,且PIS判据在特征选择

算法中的应用稳定性较好。

参考文献

- [1] Shan Shiguang, Gao Wen, Chen Xilin, et al. Local Gabor Binary Pattern Histogram Sequence(LGBPHS): A Novel Nonstatistical Model for Face Representation and Recognition[C]//Proc. of the 10th IEEE International Conference on Computer Vision. [S. l.]: IEEE Press, 2005.
- [2] Anil K J, Robert P W D, Mao Jianchang. Statistical Pattern Recognition: A Review[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(1): 4-37.
- [3] 宋枫溪,高秀梅,刘树海,等.统计模式识别中的维数削减与低损降维[J].计算机学报,2005,28(11):1915-1922.
- [4] Liu Huan, Motoda H, Setiono R, et al. Feature Selection: An Ever Evolving Frontier in Data Mining[EB/OL]. (2011-10-08). <http://jmlr.csail.mit.edu/proceedings/papers/v10/liu10b/liu10b.pdf>.
- [5] He Xiaofei, Cai Deng, Partha N. Laplacian Score for Feature Selection[C]//Proc. of NIPS'05. Cambridge, USA: MIT Press, 2005.
- [6] 徐燕,李锦涛,王斌,等.基于区分类别能力的高性能特征选择方法[J].软件学报,2008,19(1):82-89.
- [7] Yu Lei, Liu Huan. Efficient Feature Selection via Analysis of Relevance and Redundancy[J]. Journal of Machine Learning Research, 2004, 5: 1205-1224.
- [8] Wiskott F, Kruger J M, Malsburg V D. Face Recognition by Elastic Bunch Graph Matching[C]//Proc. of International Conference on Image Processing. [S. l.]: IEEE Press, 1997.
- [9] Ojala T, Pietikäinen M, Mäenpää T. Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 971-987.
- [10] Liu Chengjun, Wechsler H. Gabor Feature Based Classification Using the Enhanced Fisher Linear Discriminant Model for Face Recognition[J]. IEEE Transactions on Image Processing, 2002, 11(4): 467-476.
- [11] He Xiaofei, Partha N. Locality Preserving Projections[EB/OL]. (2010-11-21). <http://people.cs.uchicago.edu/~xiaofei/LPP.html>.
- [12] Dai Wang, Fang Yuchun, Hu Binbin. Feature Selection in Interactive Face Retrieval[C]//Proc. of the 4th International Congress on Image and Signal Processing. [S. l.]: IEEE Press, 2011.
- [13] 张文超,山世光,张洪明,等.基于局部Gabor变化直方图序列的人脸描述与识别[J].软件学报,2006,17(12):2508-2517.
- [14] Wikipedia. Cross-validation(Statistics)[EB/OL]. (2010-08-16). [http://en.wikipedia.org/wiki/Cross-validation_\(statistics\)](http://en.wikipedia.org/wiki/Cross-validation_(statistics)).

编辑 刘冰