Fisher 鉴别特征的最近邻凸包分类*)

姜文瀚　杨静宇　周晓飞

(南京理工大学计算机科学与技术学院　南京 210094)

摘　要　基于 Fisher 准则的特征提取方法是模式识别技术的重要分支，其中，Foley-Sammon 变换和具有统计不相关性的最佳鉴别变换是这一技术典型代表，本文将它们与一种新型分类器—最近邻凸包分类器相结合，从而实现 Fisher 鉴别特征的有效分类。最近邻凸包分类器是一类以测试样本点到各类训练集生成类别凸包的距离为分类判别依据的模式分类新方法，具有非线性性，无参性，多类别适用性等特点。实验证实了本文方法的有效性。

关键词　特征提取，最近邻凸包分类器，凸包，模式分类

Nearest Neighbor Convex Hull Classification of Fisher Discriminant Features

JIANG Wen-Han　YANG Jing-Yu　ZHOU Xiao-Fei

(Department of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094)

Abstract　Feature extraction based on Fisher criteria is a branch of pattern recognition. Foley-Sammon algorithm and the Uncorrelated Fisher Linear Discriminant Analysis (ULDA) are the classic two of those correlative methods. As preprocessors, they usually cooperate with other classification algorithm such as the minimum distance classifier, the nearest neighbor classifier or support vector machines (SVMs). In this paper, the results of them are used as inputs to a new classification method named the nearest neighbor convex hull (NNCH) classifier, which takes the convex hull of one class training data as a new unit class. The test sample will belong to the class of the nearest convex hull in the feature space. Nonlinearity, no parameters and multi-class applicability are the characters of NNCH. The experiments compared with the other cooperators mentioned above, indicate the good performance of the proposed methods.

Keywords　Feature extraction, Nearest neighbor convex hull classifier, Convex hull, Pattern classification

1　引言

Fisher 鉴别基本思想始见于 Fisher 在 1936 年发表的经典论文[1]。基于该思想的特征提取方法，如 Fisher 线性分类分析(Fisher Linear Discriminant Analysis, FLDA)[2,3]，Foley-Sammon[4,5]变换，具有统计不相关性的最佳鉴别变换(Uncorrelated FisherLinear Discriminant Analysis, ULDA)[6]等，在模式识别领域有着广泛的应用。FLDA 是由 Wilks 和 Duda 提出的，旨在依据 Fisher 准则，寻找一组鉴别矢量。样本在由该矢量集构成的子空间内的投影即为鉴别特征。Foley-Sammon 变换和 ULDA 可以说是在此基础上发展起来的一对姊妹技术。前者选择的是一组满足正交条件的最佳鉴别矢量，后者的最优鉴别矢量集则要求满足共轭正交条件，从而提取样本的统计不相关特征。综合而言，这类方法一方面实现了数据的降维压缩，另一方面有效地提取了数据的鉴别特征。在保证系统性能的同时，大大加快了计算速度，降低了运算复杂度。

模式识别系统的另一重要环节是对提取特征的分类判别。通常这一任务交由最小距离分类器，最近邻分类器或支持向量机(Support Vector Machines- SVMs)[7]等技术手段来完成。最近邻凸包分类器(Nearest Neighbor Convex Hull-NNCH)是一类以测试样本点到各类训练集生成类别凸包的距离为分类判别依据的模式分类新方法。该方法具有非线性、无参性、多类别适用性等诸多特点，本文将其与上述特征提取方法相结合，取得了较好的实验效果。

本文在第 2 部分分别对 FLDA，Foley-Sammon 变换和具有统计不相关性的最佳鉴别变换等三种特征提取方法进行了简单概述。第 3 部分介绍了最近邻凸包分类方法。第 4 部分在 ORL(Olivetti Research Lab)人脸数据库实验平台上，针对 Foley-Sammon 变换和 ULDA 两种特征提取方法抽取的鉴别特征，我们分别采用最小距离分类器，最近邻分类器，线性核函数支持向量机和最近邻凸包分类器四种不同的方法进行了分类比较，并加以分析总结。

2　FLDA、Foley-Sammon 变换、ULDA

FLDA、Foley-Sammon 变换、具有统计不相关性的最佳鉴别变换，三者是基于 Fisher 鉴别准则，一脉相承的特征提取方法。其中后两者是前者思想基础上的改进。FLDA 提取的是满足正交条件的一组最佳鉴别矢量。ULDA 选择共轭正交的最佳鉴别矢量集构成样本特征子空间，以抽取统计不相关的样本特征。

2.1　Fisher 线性分类分析(FLDA)

FLDA 是传统的特征提取手段之一，其主要思想是依据 Fisher 准则，寻求一组鉴别投影矢量，使得投影变换后的样本特征具有最佳可分性，即最大化类间距离与类内距离比。简单表述如下：

*)本课题得到国家自然科学基金(60472060)资助。**姜文瀚**　博士研究生，主要研究方向为人工智能、模式识别。**杨静宇**　教授，博士生导师，主要研究方向为计算机视觉，机器学习，智能机器人与模式识别等。**周晓飞**　博士研究生，主要研究方向为人工智能与模式识别。

设 $\omega=\{\omega_1,\omega_2,\cdots,\omega_c\}$ 是包含 c 个类别的类集，其中类别 $\omega_i(i+1,\cdots,c)$ 含有 n_i 个 l 维训练样本 $x_j^{(i)}(j=1,\cdots,n_i)$，$\sum_{i=1}^{c}n_i=n$，$p(\omega_i)(i=1,\cdots,c)$ 是 ω_i 类的先验概率，这里定义 $p(\omega_i)=n_i/n$。

Fisher 准则函数定义为：

$$J(\varphi)=\varphi^T S_b\varphi/\varphi^T S_w\varphi \tag{1}$$

其中，φ 为待求最佳鉴别矢量，

样本类间散布矩阵：

$$S_b=\sum_{i=1}^{c}p(\omega_i)(m_i-m)(m_i-m)^T \tag{2}$$

样本类内散布矩阵：

$$S_w=\sum_{i=1}^{c}(p(\omega_i)/n_i)\sum_{j=1}^{n_i}(x_j^{(i)}-m_i)(x_j^{(i)}-m_i)^T \tag{3}$$

W_i 类样本均值：

$$m_i=\frac{1}{n_i}\sum_{j=1}^{n_i}x_j^{(i)} \tag{4}$$

总体样本均值：

$$m=\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n_i}x_j^{(i)} \tag{5}$$

最大化上述目标函数(1)的解，即最佳鉴别矢量，为以下广义特征方程最大特征值对应的特征向量：

$$S_b\varphi=\lambda S_w\varphi \tag{6}$$

样本 x 在最佳鉴别矢量上的投影，$y=\varphi^T x$，即为 x 的鉴别特征。

2.2 Foley-Sammon 变换

FLDA 经典算法主要是针对两类问题的。Foley-Sammon 变换在正交条件下把 FLDA 推广到多类。该方法首先利用 FLDA 经典算法求得最佳鉴别矢量 φ_1。在条件 $\varphi_{i+1}^T\varphi_i=0(i=1,\cdots,r)$ 约束下，第 $r+1$ 个最佳鉴别矢量是以下广义特征方程最大特征值对应的特征向量：

$$PS_b\varphi_{r+1}=\lambda\Sigma_w\varphi_{r+1} \tag{7}$$

$$P=I-D^T(DS_w^{-1}D^T)^{-1}DS_w^{-1},D=[\varphi_1,\varphi_2,\cdots,\varphi_r]^T \tag{8}$$

通过迭代，求解以上方程，从而得到一组正交最佳鉴别矢量。

2.3 具有统计不相关性的最佳鉴别变换(ULDA)

具有统计不相关性的最佳鉴别变换将 Foley-Sammon 变换的正交条件替换为共轭正交条件 $\varphi_{r+1}^T S_i\varphi_i=0$，$Sbst=S_b+S_w$，$i=1,\cdots,r$。相应的待求解广义特征方程为：

$$PS_b\varphi_{r+1}=\lambda S_w\varphi_{r+1} \tag{9}$$

$$P=I-S_tD^T(DS_tS_w^{-1}S_tD^T)^{-1}DS_tS_w^{-1},D=[\varphi_1,\varphi_2,\cdots,\varphi_r]^T \tag{10}$$

同样迭代求解方程得到一组共轭正交最佳鉴别矢量。

2.2 和 2.3 两节所述方法，样本 x 在所得到的最佳鉴别矢量集上的投影，$y=[\varphi_1,\varphi_2,\cdots,\varphi_k]^T x$，即为 x 的鉴别特征矢量。ULDA 投影所得各个特征分量具有统计不相关的性质。

上述三种方法既提取了数据的有效特征，也实现了样本的降维压缩，客观上降低了后续处理工作的计算复杂度。这一点对于处理高维数据集具有着重要意义。

3 最近邻凸包分类

我们基于同类相聚的模式识别假定，认为同类样本会分布在同一凸包里或离凸包较近的区域。测试时，依据最近邻凸包的类别确定测试样本归属。按照这个原则设计的分类器称为最近邻凸包分类器。从这个角度出发分析，支持向量机的直观几何意义也是最大间隔地实现训练样本凸包的相互分离。

首先应当明确以下的概念：

定义 1(凸包)[8] 设集合 $S\subset R^n$ 是由 R^n 中的 k 个点组成的集合，即 $S=\{x_1,\cdots,x_k\}$。定义 S 的凸包 $co(S)$ 为 $co(S)=\{x=\sum_{j=1}^{k}\lambda_j x_j\mid\sum_{j=1}^{k}\lambda_j=1,\lambda_j\geqslant 0,j=1,\cdots,k\}$。

定义 2(凸集)[8] 设集合 $S\subset R^n$，称 S 是凸集，如果对任意 $x_1,x_2\in S$ 和任意的 $\lambda\in[0,1]$，都有 $\lambda x_1+(1-\lambda)x_2\in S$。

定理 1(闭凸集投影定理)[8] 设 S 是 R^n 中一闭凸集，并且 $x'\notin S$。考虑 S 中的各点 x 到 x' 的最小距离问题 $\begin{cases}\text{minimise}\parallel x-x'\parallel\\ \text{s. t. } x\in s\end{cases}$，则 S 中存在唯一的点 $\bar{x}$，使 $\bar{x}$ 与 x' 的距离为最小，即上述最优化问题有唯一极小点，而且，$\bar{x}$ 是极小点的充要条件是对任意的 $x\in S$，有 $(x-\bar{x})^T(\bar{x}-x')\geqslant 0$。

依据定理 1，对于含有 k 个训练样本的一类样本集，$S\subset R^n$，$S=\{x_i\mid x_i\in R^n,i=1,\cdots,k\}$，我们定义样本点 $x\in R^n$ 到 S 的凸包的距离为：

$$dist(x,co(S))=\inf_{y\in co(S)}(\parallel x-y\parallel) \tag{11}$$

令 $\alpha=(\alpha_1,\cdots,\alpha_k)^T$，按照定义 1，

$$dist(x,co(S))=\text{minimise}_\alpha\parallel x-\sum_{i=1}^{k}\alpha_i x_i\parallel$$

$$\text{s. t. }\sum_{i=1}^{k}\alpha_i=1,\alpha_i\geqslant 0 \tag{12}$$

向量 2-范数意义下有：

$$dist^2(x,co(S))=\text{minimise}_\alpha\parallel x-\sum_{i=1}^{k}\alpha_i x_i\parallel_2^2$$

$$=\text{minimise}_\alpha(x^Tx-2x^T(x_1,\cdots,x_l)\alpha+\alpha^T(x_1,\cdots,x_k)^T(x_1,\cdots,x_k)\alpha) \tag{13}$$

$$\text{s. t. }\sum_{i=1}^{k}\alpha_i=1,\alpha_i>0$$

方程(13)两边除 2，这里 x^Tx 是优化无关项，在优化过程中可省略，进而有：

$$\text{minimise } W(\alpha)=\frac{1}{2}\alpha^T(x_1,\cdots,x_k)^T(x_1,\cdots,x_k)\alpha-x^T(x_1,\cdots,x_k)\alpha \tag{14}$$

$$\text{s. t. }\sum_{i=1}^{k}\alpha_i=1,\alpha_i\geqslant 0$$

若 α^* 是以上方程的优化解，则

$$dist^2(x,co(S))=2W(\alpha^*)+x^Tx \tag{15}$$

对于多类问题，我们把每一类训练样本生成的凸包作为一个新的集类，依据最近邻分类原则决定测试样本的类别归属。c 类 $\omega=\{\omega_1,\omega_2,\cdots,\omega_c\}$ 问题，若 $dist(x,co(S_j))=\min_{i=1,\cdots,c}(dist(x,co(S_i)))$，则 $x\in\omega_j$。

最近邻凸包分类器是一类以测试样本点到各类别凸包的距离为分类判别依据的分类器，这一点有别于最近邻分类器。而最近邻分类原则的吸收，又有效借鉴了后者的非线性分类特征及多类适用性的特点。另外，该算法的无参性，即不存在任何需要预先设定的经验参数，也使得这一分类器更为方便适普。

4 实验及结论

实验采用 AT&T 剑桥实验室(AT&T Laboratories Cambridge)的 ORL(Olivetti Research Lab)人脸数据库(http://www.cl.cam.ac.uk/Research/DTG/attarchive/face database.html)作为实验平台。该数据库包括 40 人，每人 10 幅，共计 400 幅 PGM 格式人脸灰度图像。256 灰度级，92×112 像素。类别标识为：s×(×是 1～40 自然数)；样本标识为×：(×是 1～10 自然数)。图像采集条件：不同时间，变化光照，不同表情(如睁眼，闭眼，微笑，不笑)，取舍饰物(如戴眼

镜,不戴眼镜)等。所有图像均采用单一黑色背景,以前正向脸为主,兼有些许缩放,旋转和侧移,如图1。

实验中,全部图像转换为JPG格式,双三次插值缩至16×16大小,并列向拉伸。对于特征提取算法中遇到的样本类内散布矩阵 S_w 不可逆问题,本文采用增加扰动的方法加以处理。即 $S_w=S_w+\mu I$,μ 是常数,I 是单位矩阵。至于特征抽取的维数问题,我们参见论文[9],统一选择 c−1 维鉴别特征矢量(c是类别数,本实验即 c=40)。

根据Fisher鉴别特征提取方法的不同,我们设计了两项实验。实验一采用Foley-Sammon(FS)变换进行特征提取,分别由最小距离分类器(MD),最近邻分类器(NN),线性核函数支持向量机(SVM(linear))和最近邻凸包分类器(NNCH)进行分类。这里的最小距离分类器是把测试样本点到各类别中心的距离作为判别依据的分类方法。实验二采用具有统计不相关性的最佳鉴别变换(ULDA)提取特征,分类判别选用同样的一组分类器方法。每项实验均分为10组进行测试。每组顺序选择各类别第 i 个样本构成测试集,其余样本作为训练集。例如,在第1组实验中,我们提取各类别的第1个样本,共计40个,组成测试样本集,余下的360个样本共同构成该组实验的训练集。实验结束,我们分别将两项实验的各10组数据取平均值,作为最终评价依据。详细实验结果见表1和表2。

图1 ORL人脸库图像示例

表1 Foley-Sammon(FS)变换鉴别特征的四种分类器分类结果比较

实验分组	1	2	3	4	5	6	7	8	9	10	平均
FS+MD	92.5%	97.5%	100%	97.5%	97.5%	100%	95%	97.5%	97.5%	92.5%	96.75%
FS+NN	97.5%	100%	100%	100%	97.5%	100%	95%	95%	97.5%	92.5%	97.5%
FS+SVM(linear)	97.5%	100%	100%	95%	97.5%	97.5%	97.5%	97.5%	97.5%	95%	97.5%
FS+NNCH	97.5%	100%	100%	97.5%	97.5%	100%	97.5%	97.5%	97.5%	92.5%	97.75%

表2 具有统计不相关性的最佳鉴别变换(ULDA)鉴别特征的四种分类器分类结果比较

实验分组	1	2	3	4	5	6	7	8	9	10	平均
ULDA+MD	95%	92.5%	87.5%	85%	85%	90%	85%	85%	85%	87.5%	87.75%
ULDA+NN	95%	92.5%	87.5%	87.5%	90%	92.5%	77.5%	85%	87.5%	85%	88%
ULDA+SVM(linear)	92.5%	90%	90%	85%	82.5%	87.5%	85%	82.5%	87.5%	85%	86.75%
ULDA+NNCH	95%	92.5%	85%	90%	92.5%	90%	80%	85%	87.5%	87.5%	88.5%

以上两表中的实验数据均为正确识别百分率(正确识别样本数/测试样本数×100%)。各组合方法的平均实验结果统计在表的最后一列。我们用粗体标注各组实验的最佳结果。

通过平均值的比较,我们看到:最近邻凸包分类器组合表现最优且较为稳定。实验一的10组实验中,FS+NNCH有8个最佳结果,且平均值最高。其他三种分类器MD,NN,SVM(linear)依次是5个,7个,8个。SVM(linear)组合相比略次之。在实验二的10组实验中,ULDA+NNCH有7个实验达到最好水平,同样是最佳的平均结果。其他的分别是MD 5个,NN 5个,SVM(linear)3个。比照实验一,线性核函数支持向量机在该项试验中则略显不稳定。同时也应注意:实验样本对识别结果有着重要影响。如实验一的第4,第10两组实验中,FS+NNCH的结果分别低于FS+NN和FS+SVM(linear)。实验二的第3,第6,第7三组实验的ULDA+NNCH实验结果也同样略低于其他分类器组合。究其原因可能在于样本分布(包括测试样本和训练样本)及分类器的推广泛化能力的差异。

综合而言,本文提出的把Fisher鉴别特征提取技术与最近邻凸包分类器相结合的方法具有着相对较好的分类性能。相信会有更为广泛的应用前景。诚然,任何特征提取方法和分类器都有其优点和不足,两方面技术的改进及有机融合将是我们今后进一步研究的发展方向。

参考文献

1 Fisher R A. The use of multiple measurements in taxonomic problems [J]. Annals of Eugenics, 1936, 7: 178~188

2 Wilks S S. Mathematical Statistics [M]. New York: Wiley, 1962. 577~578

3 Duda R, Hart P. Pattern Classification and Scene Analysis [M]. New York: Wiley, 1973

4 Foley D H, Sammon J W Jr. An optimal set of discriminant vectors [J]. IEEE Trans. Computer, 1975, 24(3): 281~289

5 Duchene J, Leclercq S. An optimal transformation for discriminant and principal component analysis [J]. IEEE Trans. Pattern Anal. Machine Intell, 1988, 10(6): 978~983

6 Jin Z, Yang J Y, Tang Z M, Hu Z S. A theorem on uncorrelated optimal discriminant vectors [J]. Pattern Recognition, 2001, 34(10): 2041~2047

7 Vapnik V N. The Nature of Statistical Learning Theory [M]. Springer, 1995

8 邓乃扬,田英杰著. 数据挖掘中的新方法—支持向量机[M]. 北京:科学出版社, 2004

9 杨静宇,金忠,胡钟山. 具有统计不相关性的最佳鉴别特征空间的维数定理[J]. 计算机学报, 2003, 26(1):110~115