关于最佳鉴别特征维数问题的讨论

李昭阳[1)] 王元全[2)] 夏德深[2)]

[1)](中国科学院软件研究所人机交互与智能信息处理实验室 北京 100080)

[2)](南京理工大学计算机科学系 南京 210094)

摘 要 该文对最佳鉴别特征的最佳维数问题进行了详细的讨论.文章首先对最佳维数问题进行了界定,然后指出了两种最佳特征维数为 c－1 维的情况即以某些基于矩的可分性判据(准则函数)为优化目标的最优特征和以某些特殊的分类器错误率为优化目标的最优特征.最后该文运用方差分析法对最佳鉴别特征进行特征选择使之代入最小距离分类器后识别率最大.

关键词 最佳鉴别特征;最优特征;维数问题;特征选择;方差分析

中图法分类号: TP301

The Dispute on the Dimension Problem of Optimal Discriminant Features

LI Zhao-Yang[1)] WANG Yuan-Quan[2)] XIA De-Shen[2)]

[1)](*Human Computer Interface and Intelligent Information Process Lab*, *Institute of Software*, *Chinese Academy of Science*, *Beijing* 100080)

[2)](*Department of Computer Science*, *Nanjing University of Science and Technocogy*, *Nanjing* 210094)

Abstract This paper makes a detail discussion on the optimal dimension problem of optimal discriminant feature set, gives the condition under which the dimension of optimal features sets is $c-1$. That is the optimal features set which optimizes the discriminant criterion with the form $f(M_1,\cdots,M_c,S_1,\cdots,S_c)$, where M_i and S_i are the first-and second-order moments, and the optimal features set which optimizes the error rate of some special classifiers, such as Bayesian linear classifier, minimum-distance classifier. Both of the optimal features are the posterior probabilities or their function. At last, variance analysis is introduced against estimation error to select the optimal discriminant features that minimize the error rate of minimum-distance classifier. This method can also be used in other feature selection problems.

Keywords optimal discriminant features; optimal features; optimal dimension problem; feature selection; variance analysis

1 引 言

文献[1,2]提出了有关最佳鉴别特征的算法及其维数问题.对算法[1]我们并无异议,但在维数问题上,我们认为,作者根据“Fukunaga 维数定理”作出“具有统计不相关的最佳鉴别变换,可抽取最有效的 c(模式类别数)－1 个模式鉴别特征”的结论[2]是错误的.本文首先针对文献[2]的含糊表述,重新界定了最佳维数问题,并指出了文献[2]错误的原因,在接下来的部分,本文将方差分析方法用于最佳鉴别特征的特征选择,以使最小距离分类器达到最大的识别率.在附录部分,详细讨论了最佳特征维数为 c－1的情况.

收稿日期:2001-10-24;修改稿收到日期:2002-06-06.**李昭阳**,男,1973 年生,博士,主要研究方向包括模式识别、人机界面. E-mail: lzy@iel.iscas.ac.cn. **王元全**,男,1973 年生,博士研究生,主要研究方向包括模式识别、医学图像分析. **夏德深**,男,1941 年生,教授,博士生导师,博士,研究领域包括模式识别、遥感信息系统、医学图像分析、计算机视觉.

2 最佳维数问题的提出

首先,我们必须明确维数问题的真正含义.文献[2]的引言提到了,但表述不清.为此,我们假定某模式识别问题的原始特征集 X 给定,模式类别为 c,优化目标为 $F(\Psi)$,所谓最佳变换是指使优化目标 F 达到最大的变换,即

$$F(\Psi_{\text{opt}}) = \max F(\Psi) \tag{1}$$
$$\Psi_{\text{opt}}: X \mapsto \boldsymbol{Y}_{\text{opt}}$$

最佳特征为 $\boldsymbol{Y}_{\text{opt}}$,$\boldsymbol{Y}_{\text{opt}}$ 的维数称为最佳特征维数(如 $\boldsymbol{Y}_{\text{opt}}$ 有多个解则以维数最少者称之),求最佳变换实际是个求泛函极值的问题.在原始特征集 $\boldsymbol{X}$ 给定的条件下,优化目标 $F(\Psi)$ 不同,最佳特征及其维数亦不相同.常见的目标函数有:分类器的识别率(或误判率);基于矩的准则函数如 $tr\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$,$\dfrac{tr(\boldsymbol{S}_b)}{tr(\boldsymbol{S}_w)}$ 等准则函数.离开优化目标 $F(\Psi)$ 讨论最佳特征及其维数问题是没有意义的.

3 关于 c−1 维的最佳维数的讨论

3.1 准则函数下的最佳特征

文献[2]的定理 2(Fukunaga 维数定理)是说以 Bayes 分类器的误判率作为目标函数(一个复杂的带(多重)积分的函数),最佳特征是 $\boldsymbol{Y}_{\text{opt}}=[q_1(\boldsymbol{X}),q_2(\boldsymbol{X}),\cdots,q_{c-1}(\boldsymbol{X})]^{\mathrm{T}}$,最佳维数为 $c-1$,其中 $q_i(\boldsymbol{X})$ 为后验概率.

而文献[2]提出的三种最佳鉴别变换可归纳成:

求取线性变换 $\boldsymbol{\phi}_{\text{opt}}=(\boldsymbol{\varphi}_1,\boldsymbol{\varphi}_2,\cdots,\boldsymbol{\varphi}_r)$ 使得 fisher 广义鉴别准则函数

$$F(\boldsymbol{\phi}) = \frac{tr(\boldsymbol{\phi}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{\phi})}{tr(\boldsymbol{\phi}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{\phi})} = \frac{tr(\boldsymbol{S}_b\boldsymbol{\phi})}{tr(\boldsymbol{S}_w\boldsymbol{\phi})} \tag{2}$$

达到最大,即

$$F(\boldsymbol{\phi}_{\text{opt}}) = \max_{\boldsymbol{\phi}} F(\boldsymbol{\phi})$$

并且满足约束条件 $\boldsymbol{\varphi}_{r+1}^{\mathrm{T}}\boldsymbol{\varphi}_i=0$ 或 $\boldsymbol{\varphi}_{r+1}^{\mathrm{T}}\boldsymbol{S}_t\boldsymbol{\varphi}_i=0$,$i=1,2,\cdots,r$.

文献[2]称 $\boldsymbol{\varphi}$ 为最佳鉴别方向,$\boldsymbol{Y}_{\text{opt}}=\boldsymbol{\phi}'\boldsymbol{X}$ 为最佳鉴别特征.除了广义最佳鉴别变换外,其他两种变换(Foley-Sammon 变换和具有统计不相关性的最佳鉴别变换)求解的过程是分步进行的,因此这两种鉴别变换得到的解不能从整体上保证达到 $F(\boldsymbol{\phi})$ 最大[3].

不论求解过程怎样,此最佳鉴别变换的目标函数与 Fukunaga 维数定理的目标函数相去甚远,因此其最佳维数不能由 Fukunaga 最佳维数 $c-1$ 来给定,以 Fukunaga 维数定理的 $c-1$ 维来界定此(线性)最佳鉴别变换的最佳维数是不合理的.

实际上 Fukunaga 并未笼统地提出维数定理.他的有关维数问题的一些结论都是在特定的目标函数下得到的,且结果多为非线性变换.例如,若我们不局限在线性变换中来最大化准则函数,易知该准则函数对位移和非奇异线性变换具有不变性.由文献[4]知,其最佳特征为 $\boldsymbol{Y}_{\text{opt}}=[q_1(\boldsymbol{X}),q_2(\boldsymbol{X}),\cdots,q_{c-1}(\boldsymbol{X}),q_c(\boldsymbol{X})]^{\mathrm{T}}$,这里因为 $\sum_{i=1}^{c}q_i(\boldsymbol{X})=1$,实际维数为 $c-1$(详细讨论参看附录).

3.2 最小距离分类器错误率下的最佳特征

文献[2]还以最小距离分类器的识别错误率来验证统计不相关的最佳鉴别变换的最佳维数是 $c-1$ 的结论,也就是说当统计不相关的最佳鉴别方向取前 $c-1$ 个时,最小距离分类器的错误率最小.这样的验证是缺乏效力的.

如前所述,由于目标函数的不同,最佳鉴别变换所得到的特征与这些特征代入最小距离分类器所得到的错误率无直接关系.我们只能大致地说,对于设计"合理"的分类器(分类器的设计可以是任意的,对模式样本空间的任意划分都可构成一个分类器,这样最佳鉴别变换所得到的特征对某分类器来说可能是很差的),前面的最佳鉴别方向要比后面的最佳鉴别方向鉴别效果好,代入分类器后错误率一般会低些.至于哪些最佳鉴别方向使分类器错误率达到最低,由于数学上无法证明,因此不能就此得出最佳鉴别变换的前 $c-1$ 个最佳鉴别方向使最小距离分类器的识别错误率最低的结论.

若以文献[2]的最小距离分类器的识别率作为目标函数,由附录可知最佳特征 $y_i(\boldsymbol{X})$,$i=1,2,\cdots,c$ 为后验概率 $q_1(\boldsymbol{X})$,$q_2(\boldsymbol{X})$,$\cdots$,$q_c(\boldsymbol{X})$ 的线性函数,最佳维数为 $c-1$.这时我们才可以说取这些个特征时,最小距离分类器错误率最小.当分类器改变时,一般地最佳特征和最佳维数都要改变,如当采用 Bayes 线性分类器时,虽然最佳维数仍为 $c-1$ 但最佳特征为 $\boldsymbol{Y}_{\text{opt}}=[q_1(\boldsymbol{X}),q_2(\boldsymbol{X}),\cdots,q_c(\boldsymbol{X})]^{\mathrm{T}}$.在此,最佳变换都为非线性变换.

即使在最佳鉴别特征中前 $c-1$ 个特征使最小距离分类器错误率达到最小,还很难想象这 $c-1$ 个特征也使其它的分类器(如 Bayes 线性分类器)错误

率最小.由于文献[2]对最佳特征问题表述上的含糊,给人的印象是想在最佳鉴别特征中得到对所有分类器都有最小误判率的最佳特征,这是不可能的,并且认定前 $c-1$ 个最佳鉴别特征为最佳特征也是不可靠的.因此可以解决的问题应是在最佳鉴别变换中选取使某分类器错误率达到最小的那些最佳鉴别方向.该问题实际是个特征选择问题,它的解与所用分类器的类型有关,文献[2]提出最佳鉴别变换有效度的概念与具体分类器无关,因此不能用来解决该问题.

此外,文献[1,2]的实验都用错误率的样本估计值$\hat{\varepsilon}(i)$代替 $\varepsilon(i)$(最佳鉴别方向个数为 i 时,最小距离分类器错误率的真值)来直接比较也是不妥的.随机变量$\hat{\varepsilon}(i)$间的比较要在一定的置信度 $\beta(\alpha=1-\beta$ 称为显著水平)基础上进行,例如 $P\{\hat{\varepsilon}(j)<\hat{\varepsilon}(i)\}=\beta$,表示$\hat{\varepsilon}(j)$小于$\hat{\varepsilon}(i)$的可能性为 β.当 β 接近 1(β 常取 0.9)时,称$\hat{\varepsilon}(j)$在统计上显著地小于$\hat{\varepsilon}(i)$,这时我们才能较有把握地作出$\hat{\varepsilon}(j)$小于$\hat{\varepsilon}(i)$的结论.下面运用方差分析法[5]来确定最佳鉴别变换的最佳维数.

4　最佳鉴别变换最佳维数(针对最小距离分类器)的确定

方差分析是考察各因素对响应指标是否有显著影响的数理统计方法.在此,我们把最小距离分类器的识别率(响应指标)看成是受各鉴别特征(因素)影响的结果.各鉴别特征的参选与否对识别率的影响有正向(增加识别率)和负向的;影响程度从统计上看有显著和不显著的.我们试图通过方差分析选出那些对识别率影响是正向的且统计上是显著的鉴别特征,从而得到最佳特征.

我们的实验采用文献[1]的 ORL 人脸图像数据库,由于作者没有说明将原图像压缩至 6×7 的算法,我们先对 92×112 图像进行 *K-L* 变换降维,再进行具有统计不相关的最佳鉴别变换(具体参文献[6]第三章).在得到一组最佳鉴别特征后,再应用方差分析法在其中选取使最小距离分类器的识别率达到最大的特征.从最佳鉴别变换的算法过程可以看出,对最小距离分类器而言,越靠后的特征其鉴别力越差,本实验模式类别较大($c=40$),我们猜测第 40 个以后的特征对识别率的影响几乎可以不予考虑(在后面的方差分析中将会验证这一点).因此我们就在这前 39 个鉴别特征中选取使最小距离分类器的识别率达到最大的特征集.

表 1　统计不相关最佳鉴别特征的方差分析表

$F_{11}^{1}(0.1)=3.23,\alpha=0.1$

特征	F	特征	F
1	17.80*	22	3.08
2	14.18*	23	2.95
3	13.11*	…	…
4	12.73*	33	1.32
…	…	…	…
20	3.71*	39	1.01
21	3.57*		

$F(i)=\dfrac{V_i}{V_e}$,V_i表示第 i 个特征的平均波动(方差),它衡量第 i 个特征对识别率的影响大小,V_e表示误差的平均波动.从表 1 看出第 33 个特征之后它们对识别率的影响与误差的影响相差不大,远小于阈值 3.23,可将它们的影响与误差合并.故忽略第 39 个以后的鉴别特征是合理的.分析中发现所有特征的影响都是正向的.前 21 个特征的 F 值都大于阈值 3.23 称为显著特征,表示它们对识别率的正向影响在统计上是显著的(至少有 90%的把握).所以我们预测最佳维数应在 21 维左右出现.

表 2 验证了方差分析的结果.表中的识别率是平均识别率,每个模式类别随机地抽取 5 个训练样本,用剩余的样本为测试样本计算各个特征维数下的最小距离分类器的识别率$\hat{\varepsilon}_1(i)$,i 为特征维数.更换测试样本类似地测试 10 次,以平均值$\bar{\hat{\varepsilon}}(i)=\dfrac{1}{10}\sum_{j=1}^{10}\hat{\varepsilon}_j(i)$作为识别率真值的估计,它具有较小的估计偏差和方差[7],更真实地反映识别率真值.

表 2　统计不相关最佳鉴别方向不同维数下的识别率

维数	识别率	维数	识别率
1	0.20	22	0.86
2	0.40	…	…
3	0.55	33	0.84
4	0.70	…	…
…	…	38	0.835
20	0.865	39	0.83
21	0.865		

从表 2 中看出最佳维数为 20,与方差分析预报的结果基本一致.最小错误率并未在 $c-1(=39)$维上出现,这说明文献[2]的结论是错误的.

5　结　论

综上所述,最佳鉴别变换的最佳特征取为 $c-1$ 维理论上是没有依据的,应用中也会带来不便.比

如,人脸识别中较现实问题是模式类别数 c 很大,提取 $c-1$ 维最佳鉴别特征难以达到降维的目的.而且上面的例子也显示了这 $c-1$ 维特征并非是最有效的(分类器错误率最小).本文提出的方差分析法找到了最有效特征,因此该方法可应用在特征选择的问题上.

参 考 文 献

1 JIN Zhong *et al*. An optimal set of uncorrelated discriminant features. Chinese Journal of Computers, 1999,22(10): 1105-1108(in Chinese)
(金 忠等.一种具有统计不相关的最佳鉴别矢量集.计算机学报,1999,22(10):1105-1108)

2 JIN Zhong *et al*. Effective extraction of optimal discriminant features and the dimension problem. Chinese Journal of Computers, 2000,23(1): 108-112(in Chinese)
(金 忠等. 有效最佳鉴别特征的抽取与维数问题.计算机学报,2000,23(1):108-112)

3 Guo Yue-Fei *et al*. An iterative algorithm for the generalized optimal set of discriminant vectors and its application to face recognition. Chinese Journal of Computers, 2000,23(11): 1189-1195(in Chinese)
(郭跃飞等. 求解广义最佳鉴别矢量集的一种迭代算法及人脸识别. 计算机学报,2000,23(11):1189-1195)

4 Fukunaga K S. No linear feature extraction with a general criterion function. IEEE Trans Information Theory, 1978, 24(5): 600-607

5 Xiang Ke-Feng *et al*. Experiment Design and Data Analysis. 2nd Edition. Shanghai: Shanghai Science and Technology Press, 1989(in Chinese)
(项可风等. 试验设计与数据分析. 第 2 版. 上海: 上海科技出版社,1989)

6 JIN Zhong. The feature extraction of face image and research on dimension[Ph. D. dissertation]. Nanjing: Nanjing university of Science&Technology, 1999(in Chinese)
(金 忠.人脸图像特征抽取与维数研究[博士学位论文]. 南京理工大学,南京,1999)

7 Fukunaga K S. Introduction to Statistical Pattern Recognition. New York:Academic Press, 1990

8 Xia Dao-Xing *et al*. Real function and functional analysis(1). 2nd edition. Beijing: Higher Education Press, 1984(in Chinese)
(夏道行等. 实变函数与泛函分析(上). 第 2 版. 北京:高等教育出版社,1984)

附录 A

设模式 $\boldsymbol{X}$ 为 n 维向量,来自 c 类.类条件概率密度和先验概率分别为 $p(x/\omega_i)$, $p(\omega_i)$,$i=1,2,\cdots,c$. $\boldsymbol{Y}(\boldsymbol{X})$ 为 m 维向量函数,其分量 $y_i(\boldsymbol{X})$ 为 $\boldsymbol{X}$ 的标量函数,即

$$\boldsymbol{Y}(\boldsymbol{X}) = [y_1(\boldsymbol{X}), y_2(\boldsymbol{X}), \cdots, y_m(\boldsymbol{X})]^{\mathrm{T}} \quad (附\ 1)$$

(1) 基于矩的准则函数下的最优特征

我们把目标函数 $J(.)$限定为具有如下形式:

$$J(\boldsymbol{Y}(\boldsymbol{X})) = f(\boldsymbol{M}_1, \boldsymbol{M}_2, \cdots, \boldsymbol{M}_c, \boldsymbol{S}_1, \boldsymbol{S}_2, \cdots, \boldsymbol{S}_c) \quad (附\ 2)$$

式中 $\boldsymbol{M}_i$和 $\boldsymbol{S}_i$为 $\boldsymbol{Y}$ 在类 ω_i下的均值向量和自相关矩阵,即

$$\begin{aligned} \boldsymbol{M}_i &= \int \boldsymbol{Y}(\boldsymbol{X}) p(\boldsymbol{X}/\omega_i) \mathrm{d}\boldsymbol{X} \\ \boldsymbol{S}_i &= \int \boldsymbol{Y}(\boldsymbol{X}) \boldsymbol{Y}(\boldsymbol{X})^{\mathrm{T}} p(\boldsymbol{X}/\omega_i) \mathrm{d}\boldsymbol{X} \end{aligned} \quad (附\ 3)$$

关于目标函数(附 2)下的最优特征有如下结论.

定理 1[4]. 若目标函数 J 满足条件

$$\frac{\partial J}{\partial \boldsymbol{S}_i} = w_i \frac{\partial J}{\partial \boldsymbol{S}} \quad (附\ 4)$$

式中 $\partial J/\partial \boldsymbol{S}$ 为与 i 无关的 $m \times m$ 矩阵,w_i为正数且$\sum_{i=1}^{c} w_i = 1$,则满足(附 5)的 $\boldsymbol{Y}(\boldsymbol{X})$都为最优特征.

$$\frac{\partial J}{\partial \boldsymbol{S}} \boldsymbol{Y}(\boldsymbol{X}) = -\sum_{i=1}^{c} \hat{p}(\omega_i/\boldsymbol{X}) \frac{1}{w_i} \frac{\partial J}{\partial \boldsymbol{M}_i} \quad (附\ 5)$$

式中$\hat{p}(\omega_i/\boldsymbol{X})$为如下定义的后验概率:

$$\hat{p}(\omega_i/\boldsymbol{X}) = \frac{w_i p(\boldsymbol{X}/\omega_i)}{\hat{p}(\boldsymbol{X})} \quad (附\ 6)$$

其中

$$\hat{p}(\boldsymbol{X}) = \sum_{i=1}^{c} w_i p(\boldsymbol{X}/\omega_i) \quad (附\ 7)$$

具有下面形式(附 8)的准则函数都满足(附 4):

$$J = f(\boldsymbol{M}_1, \boldsymbol{M}_2, \cdots, \boldsymbol{M}_c, \sum_{i=1}^{c} w_i \boldsymbol{S}_i) \quad (附\ 8)$$

具体例子如

$$tr\boldsymbol{S}_w^{-1}\boldsymbol{S}_b, \frac{tr(\boldsymbol{S}_b)}{tr(\boldsymbol{S}_w)}, \ln \frac{|\boldsymbol{S}_b|}{|\boldsymbol{S}_t|} \quad (附\ 9)$$

式中

$$\boldsymbol{S}_w = \sum_{i=1}^{c} P(\omega_i)(\boldsymbol{S}_i - \boldsymbol{M}_i \boldsymbol{M}_i^{\mathrm{T}}) \quad (类内散布矩阵)(附\ 10)$$

$$\boldsymbol{S}_b = \sum_{i=1}^{c} P(\omega_i)(\boldsymbol{M}_i - \boldsymbol{M}_0)(\boldsymbol{M}_i - \boldsymbol{M}_0)^{\mathrm{T}} \quad (类间散布矩阵)(附\ 11)$$

$$\boldsymbol{S}_t = \boldsymbol{S}_w + \boldsymbol{S}_b \quad (总散布矩阵)(附\ 12)$$

$$\boldsymbol{M}_0 = \sum_{i=1}^{c} P(\omega_i) \boldsymbol{M}_i \quad (附\ 13)$$

(附 5)表示若 $\partial J/\partial \boldsymbol{S}$ 为非奇异矩阵,则最优特征 $y_i(\boldsymbol{X})$为 $\hat{p}(\omega_1/\boldsymbol{X}), \hat{p}(\omega_2/\boldsymbol{X}), \cdots, \hat{p}(\omega_c/\boldsymbol{X})$的线性函数.但是,$\partial J/\partial S$ 可能是奇异的,这时的讨论很复杂,但若目标函数 J 进一步具有一些性质则可得到较简洁的结果.

定理 2[4]. 若 $J(.)$满足(附 4),且保持位移变换不变性(附 14)和非奇异线性变换不变性(附 15)

$$J(\boldsymbol{Y}(\boldsymbol{X})+\boldsymbol{Y}_0)=J(\boldsymbol{Y}(\boldsymbol{X})) \quad (附\ 14)$$

$$J(\boldsymbol{A}\boldsymbol{Y}(\boldsymbol{X}))=J(\boldsymbol{Y}(\boldsymbol{X})) \quad (附\ 15)$$

式中 $\boldsymbol{Y}_0$ 为任意常数向量，$\boldsymbol{A}$ 为任意非奇异的 $m\times m$ 矩阵. 此时 $\boldsymbol{Y}(\boldsymbol{X})$ 具有性质

$$\boldsymbol{Y}(\boldsymbol{X})=\boldsymbol{S}_0\boldsymbol{S}_0^{+}\boldsymbol{Y}(\boldsymbol{X}) \quad (附\ 16)$$

并且满足条件(附 17)的 $\boldsymbol{Y}(\boldsymbol{X})$ 为最优特征.

$$\sum_{i=1}^{c-1}\frac{\partial J}{\partial \boldsymbol{M}_i}(\boldsymbol{M}_i-\boldsymbol{M}_c)^{\mathrm{T}}\boldsymbol{S}_0^{+}\boldsymbol{Y}(\boldsymbol{X})$$

$$=\sum_{i=1}^{c-1}\frac{\partial J}{\partial \boldsymbol{M}_i}\left[\frac{1}{w_i}\hat{p}(\omega_i/\boldsymbol{X})-\frac{1}{w_c}\hat{p}(\omega_c/\boldsymbol{X})\right] \quad (附\ 17)$$

式中 $\boldsymbol{S}_0=\sum_{i=1}^{c}w_i\boldsymbol{S}_i$ 是权为 $\{w_i, i=1,2,\cdots,c\}$ 的总散布矩阵，$\boldsymbol{S}_0^{+}$ 为 $\boldsymbol{S}_0$ 的伪逆.

$\boldsymbol{Y}(\boldsymbol{X})$ 不能直接从(附 17)得到，因为 $\partial f/\partial \boldsymbol{M}_i$，$\boldsymbol{M}_i$ 和 $\boldsymbol{S}_0$ 取决于 $\boldsymbol{Y}(\boldsymbol{X})$. (附 17)的一个解是让等式两边 $\partial J/\partial \boldsymbol{M}_i$ 项的系数相等，即

$$(\boldsymbol{M}_i-\boldsymbol{M}_c)^{\mathrm{T}}\boldsymbol{S}_0^{+}\boldsymbol{Y}(\boldsymbol{X})=\frac{1}{w_i}\hat{p}(\omega_i/\boldsymbol{X})-\frac{1}{w_c}\hat{p}(\omega_c/\boldsymbol{X}),\quad 1\le i\le c-1 \quad (附\ 18)$$

只要(附 17)中的 $\partial J/\partial \boldsymbol{M}_i$ 是线性无关的.

下面讨论

$$J(\boldsymbol{Y}(\boldsymbol{X}))=f(\boldsymbol{M}_1,\boldsymbol{M}_2,\cdots,\boldsymbol{M}_c,\sum_{i=1}^{c}P(\omega_i)\boldsymbol{S}_i) \quad (附\ 19)$$

的最佳特征. 令

$$\sigma_{ij}=\int p(\omega_i/\boldsymbol{X})p(\omega_j/\boldsymbol{X})p(\boldsymbol{X})\mathrm{d}\boldsymbol{X} \quad (附\ 20)$$

$$(\boldsymbol{M}_i)j=\int p(\omega_j/\boldsymbol{X})p(\boldsymbol{X}/\omega_i)\mathrm{d}\boldsymbol{X}=\frac{\sigma_{ij}}{p(\omega_i)} \quad (附\ 21)$$

$$(\boldsymbol{S}_0)_{l,k}=\sigma_{lk}=\sum_{i=1}^{c}p(\omega_i)\int p(\omega_l/\boldsymbol{X})p(\omega_k/\boldsymbol{X})p(\boldsymbol{X}/\omega_i)\mathrm{d}\boldsymbol{X} \quad (附\ 22)$$

推论 1.　对于具有(附 19)形式且满足性质(附 14)和(附 15)的 $J(.)$，$\boldsymbol{Y}(\boldsymbol{X})=[p(\omega_1/\boldsymbol{X}),p(\omega_2/\boldsymbol{X}),\cdots,p(\omega_c/\boldsymbol{X})]^{\mathrm{T}}$ 是最优特征.

证明.　只要证明 $\boldsymbol{Y}(\boldsymbol{X})=[p(\omega_1/\boldsymbol{X}),p(\omega_2/\boldsymbol{X}),\cdots,p(\omega_c/\boldsymbol{X})]^{\mathrm{T}}$ 满足(附 18)即可.

因为 $\boldsymbol{Y}(\boldsymbol{X})$ 具有性质(附 16)，故(附 23)式成立

$$\boldsymbol{Y}=\boldsymbol{S}_0\boldsymbol{S}_0^{+}\boldsymbol{Y}\Leftrightarrow\boldsymbol{\sigma}\boldsymbol{\sigma}^{+}\boldsymbol{Y}=\begin{pmatrix}\boldsymbol{\sigma}_1^{\mathrm{T}}\\ \vdots\\ \boldsymbol{\sigma}_c^{\mathrm{T}}\end{pmatrix}\boldsymbol{\sigma}^{+}\boldsymbol{Y}$$

$$=\begin{pmatrix}\boldsymbol{\sigma}_1^{\mathrm{T}}\boldsymbol{\sigma}^{+}\boldsymbol{Y}\\ \vdots\\ \boldsymbol{\sigma}_c^{\mathrm{T}}\boldsymbol{\sigma}^{+}\boldsymbol{Y}\end{pmatrix}=\begin{pmatrix}y_1\\ \vdots\\ y_c\end{pmatrix} \quad (附\ 23)$$

式中

$$\boldsymbol{\sigma}\triangleq\begin{pmatrix}\sigma_{11}&\cdots&\sigma_{1c}\\ \vdots&\ddots&\vdots\\ \sigma_{c1}&\cdots&\sigma_{cc}\end{pmatrix}=\boldsymbol{S}_0,\ \boldsymbol{\sigma}_i=\begin{pmatrix}\sigma_{i1}\\ \vdots\\ \sigma_{ic}\end{pmatrix},\ i=1,2,\cdots,c$$

由(附 20)知

$$\boldsymbol{M}_i=\boldsymbol{\sigma}_i/p(\omega_i),\ i=1,2,\cdots,c \quad (附\ 24)$$

将 $\boldsymbol{Y}$，$\boldsymbol{M}_i$ 和 $\boldsymbol{S}_0$ 代入(附 18)式左边

$$(附\ 18)\ 左边=\left(\frac{\boldsymbol{\sigma}_i^{\mathrm{T}}}{p(\omega_i)}-\frac{\boldsymbol{\sigma}_c^{\mathrm{T}}}{p(\omega_c)}\right)\boldsymbol{S}_0^{+}\boldsymbol{Y}$$

$$=\frac{\boldsymbol{\sigma}_i^{\mathrm{T}}\boldsymbol{S}_0^{+}\boldsymbol{Y}}{p(\omega_i)}-\frac{\boldsymbol{\sigma}_c^{\mathrm{T}}\boldsymbol{S}_0^{+}\boldsymbol{Y}}{p(\omega_c)}=\frac{y_i}{p(\omega_i)}-\frac{y_c}{p(\omega_c)}$$

$$=\frac{p(\omega_i/\boldsymbol{X})}{p(\omega_i)}-\frac{p(\omega_c/\boldsymbol{X})}{p(\omega_c)}=(附\ 8)\ 右边,$$

$$i=1,2,\cdots,c-1 \quad (附\ 25)$$

这里 $w_i=P(\omega_i)$. 因此 $\boldsymbol{Y}(\boldsymbol{X})$ 满足(附 18)式，所以 $\boldsymbol{Y}(\boldsymbol{X})$ 为最优特征.　证毕.

(2) 以误判率为准则函数的最优特征

贝叶斯线性分类器可表示为[7]：

$$h_{i,1}(\boldsymbol{X})<0,h_{i,2}(\boldsymbol{X})<0,\cdots,h_{i,c}(\boldsymbol{X})<0$$

$$\rightarrow \boldsymbol{X}\in\omega_i,h_{i,i}\ 除外 \quad (附\ 26)$$

式中

$$h_{i,l}(\boldsymbol{X})=h_{i,l}(\boldsymbol{X},\boldsymbol{M}_i,\boldsymbol{M}_l,\boldsymbol{\Sigma})$$

$$=\frac{1}{2}(\boldsymbol{Y}(\boldsymbol{X})-\boldsymbol{M}_i)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}(\boldsymbol{X})-\boldsymbol{M}_i)-$$

$$\frac{1}{2}(\boldsymbol{Y}(\boldsymbol{X})-\boldsymbol{M}_l)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}(\boldsymbol{X})-\boldsymbol{M}_l)-$$

$$\ln\frac{p(\omega_i)}{p(\omega_l)}\quad i,l=1,2,\cdots,c,l\ne i \quad (附\ 27)$$

该分类器在各类模式服从正态等协方差阵分布时才是贝叶斯分类器，又因具有线性判别形式，故而称为贝叶斯线性分类器.

由文献[7]知 ω_i 被误判为 ω_l 的概率

$$\varepsilon_{i,l}(\boldsymbol{M}_i,\boldsymbol{M}_l,\boldsymbol{\Sigma})=\begin{cases}\displaystyle\int_{h_{i,l}(x)>0}p_i(\boldsymbol{X})\mathrm{d}\boldsymbol{X}=\frac{1}{2}+\\ \displaystyle\frac{1}{2\pi}\iint\frac{\mathrm{e}^{\mathrm{j}\omega h_{i,l}(\boldsymbol{X})}}{\mathrm{j}\omega}p_i(\boldsymbol{X})\mathrm{d}\omega\,\mathrm{d}\boldsymbol{X},\ l\ne i\\ 0,\quad l=i\end{cases}$$

(附 28)

式中 j 为虚数单位. 所以贝叶斯线性分类器的总误判率为

$$\varepsilon(\boldsymbol{M}_1,\boldsymbol{M}_2,\cdots,\boldsymbol{M}_c,\boldsymbol{\Sigma})=\sum_{i=1}^{c}\sum_{l=1}^{c}p(\omega_i)p(\omega_l)\varepsilon_{i,l} \quad (附\ 29)$$

容易证明 $h_{i,l}(i,l=1,2,\cdots,c,i\ne l)$ 对 $\boldsymbol{X}$ 的位移变换和线性变换都具有不变性，所以 ε 满足(附 14)，(附 15). 另外(附 29)具有准则函数(附 19)的形式，所以 ε 满足(附 4). 只要证明 $\partial\varepsilon/\partial \boldsymbol{M}_i$，$\partial\varepsilon/\partial\boldsymbol{\Sigma}$，$i=1,2,\cdots,c$ 存在，就可利用推论 1 的结果. 实际上只要证明 $\partial\varepsilon_{il}/\partial \boldsymbol{M}_i$，$\partial\varepsilon_{il}/\partial \boldsymbol{M}_l$，$\partial\varepsilon_{il}/\partial\boldsymbol{\Sigma}$ 存在就可以了. 先介绍文献[8]中的一个定理.

引理 1[8]. 设 $f(x,t)$ 是定义在矩形 $\{(x,t)\mid a\le x\le b,\alpha\le t\le\beta\}$ 上的二元函数($[a,b]$，$[\alpha,\beta]$ 可以无限)，固定 $t\in[\alpha,\beta]$，$f(x,t)$ 是 x 的勒贝格可积函数. 如果关于勒贝格测度对几乎所有 x，函数 $f(x,t)$ 对 t 有偏导数，并存在勒贝格可积函数 $F(x)(x\in[a,b])$，使得对 $t\in[\alpha,\beta]$ 及充分小的 $|h|$，有

$$\left|\frac{f(x,t+h)-f(x,t)}{h}\right|\le F(x)\ a.e.\ 于[a,b]$$

(附 30)

那么 $I(t)=\int_a^b f(x,t)\mathrm{d}x$ 在$[\alpha,\beta]$上具有导函数,并且

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_a^b f(x,t)\mathrm{d}x=\int_a^b \frac{\partial}{\partial t}f(x,t)\mathrm{d}x \qquad (附\ 31)$$

我们以 $\partial\varepsilon_{il}/\partial\boldsymbol{\Sigma}$ 为例来说明问题.为方便记 $h_{i,l}$ 为 h. 注意到 h 是 $\boldsymbol{M}_i,\boldsymbol{M}_l,\boldsymbol{\Sigma}$ 的连续函数. $\Delta h=h(\boldsymbol{X},\boldsymbol{M}_i,\boldsymbol{M}_l,\boldsymbol{\Sigma}+\Delta\boldsymbol{\Sigma})-h(\boldsymbol{X},\boldsymbol{M}_i,\boldsymbol{M}_l,\boldsymbol{\Sigma})$,$\Delta h$ 可任意地小,只要 $\Delta\boldsymbol{\Sigma}$ 依范数$\to 0$.

$$\left|\frac{\mathrm{e}^{\mathrm{j}\omega h(\boldsymbol{\Sigma}+\Delta\boldsymbol{\Sigma})}}{\mathrm{j}\omega}-\frac{\mathrm{e}^{\mathrm{j}\omega h(\boldsymbol{\Sigma})}}{\mathrm{j}\omega}\right|=\left|\frac{\mathrm{e}^{\mathrm{j}\omega(h+\Delta h)}}{\mathrm{j}\omega}-\frac{\mathrm{e}^{\mathrm{j}\omega h}}{\mathrm{j}\omega}\right|$$

$$=\frac{|\mathrm{e}^{\mathrm{j}\omega\Delta h}-1|}{|\omega|}=\frac{\left|2\sin\frac{\Delta h}{2}\omega\right|}{|\omega|}\leq|\Delta h| \qquad (附\ 32)$$

当$|\Delta h|\leq 1$时,控制函数可取为$|p_i(\boldsymbol{X})|$,由引理 1 知道 $\partial\varepsilon_{il}/\partial h$ 存在,又因为 $\partial\varepsilon_{il}/\partial\boldsymbol{\Sigma}=\partial\varepsilon_{il}/\partial h\cdot\partial h/\partial\boldsymbol{\Sigma}$,所以 $\partial\varepsilon_{il}/\partial\boldsymbol{\Sigma}$ 存在.同理可证 $\partial\varepsilon_{il}/\partial\boldsymbol{M}_i,\partial\varepsilon_{il}/\partial\boldsymbol{M}_l$ 存在.所以 $\partial\varepsilon/\partial\boldsymbol{M}_i,\partial\varepsilon/\partial\boldsymbol{\Sigma},i=1,2,\cdots,c$ 也都存在.综合上述推导容易得到如下定理.

定理 3. $\boldsymbol{Y}(\boldsymbol{X})=[p(\omega_1/\boldsymbol{X}),p(\omega_2/\boldsymbol{X}),\cdots,p(\omega_c/\boldsymbol{X})]^{\mathrm{T}}$ 是贝叶斯线性分类器下的最优特征.

若令 $\boldsymbol{\Sigma}=\boldsymbol{I}$ 且 $p(\omega_1)=p(\omega_2)=\cdots=p(\omega_c)$,则(附 27)中 $h_{i,l}$ 变为

$$h_{il}(\boldsymbol{X},\boldsymbol{M}_i,\boldsymbol{M}_j,\boldsymbol{\Sigma})=(\boldsymbol{M}_i-\boldsymbol{M}_l)^{\mathrm{T}}\boldsymbol{Y}(\boldsymbol{X})+\frac{1}{2}(\boldsymbol{M}_i^{\mathrm{T}}\boldsymbol{M}_i-\boldsymbol{M}_l^{\mathrm{T}}\boldsymbol{M}_l)\quad i,\ l=1,2,\cdots,c,l\neq i \qquad (附\ 33)$$

易知

$$h_{il}<0\Leftrightarrow 2(\boldsymbol{M}_i-\boldsymbol{M}_l)^{\mathrm{T}}\boldsymbol{Y}(\boldsymbol{X})+(\boldsymbol{M}_i^{\mathrm{T}}\boldsymbol{M}_i-\boldsymbol{M}_l^{\mathrm{T}}\boldsymbol{M}_l)<0$$
$$\Leftrightarrow|\boldsymbol{Y}(\boldsymbol{X})-\boldsymbol{M}_i|-|\boldsymbol{Y}(\boldsymbol{X})-\boldsymbol{M}_l|<0 \qquad (附\ 34)$$

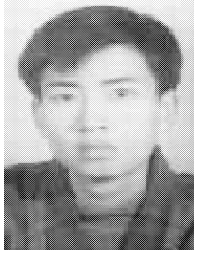
式中$|\cdot|$为欧氏距离.这时我们得到最小距离分类器:

$$\boldsymbol{X}\in\omega_i \text{ 如果 } |\boldsymbol{Y}(\boldsymbol{X})-\boldsymbol{M}_i|=\min_l|\boldsymbol{Y}(\boldsymbol{X})-\boldsymbol{M}_l| \qquad (附\ 35)$$

它将待判样本 $\boldsymbol{X}$ 判为与类均值距离最小的那一类.

推论 2. 最小距离分类器下的最优特征是 $\boldsymbol{Y}(\boldsymbol{X})=[p(\omega_1/\boldsymbol{X}),p(\omega_2/\boldsymbol{X}),\cdots,p(\omega_c/\boldsymbol{X})]^{\mathrm{T}}$的线性函数.

因为最小距离分类器误判率具有(附 19)形式,但并不满足(附 15),所以由定理 1 知最优特征是 $\boldsymbol{Y}(\boldsymbol{X})=[p(\omega_1/\boldsymbol{X}),p(\omega_2/\boldsymbol{X}),\cdots,p(\omega_c/\boldsymbol{X})]^{\mathrm{T}}$的线性函数.



LI Zhao-Yang, born in 1973, Ph. D.. Research field: pattern recognition, human computer interface.

WANG Yuan-Quan, born in 1973, Ph. D. candidate, research interests include pattern recognition, medical image analysis.

XIA De-Shen, born in 1941, Ph. D.. Research field: pattern recognition, remote sensing information system, medical image analysis and computer vision.