基于欧氏距离分布熵的特征优化研究

鲍 明,管鲁阳,李晓东,田 静
(中国科学院声学研究所,北京 100080)

摘 要: 针对训练样本集的分类特征优化选择问题,改进了样本可分度标准:Kullback-Leiber 距离,并进行了有效性验证.在此基础上定义了欧氏距离分布熵(Distribution Entropy of Euclidian Distance DEED)这一空间分布信息度量参数,同时给出了它的计算方法.提出了"类间互欧氏距离分布熵"(between-class DEED)与"类内自欧氏距离分布熵"(within-class DEED)的分析方法.进一步将其用于样本可分性分析,验证了两者比值愈大,特征样本集可分度愈好这一结论.

关键词: 改进 KL 距离; 欧氏距离; 分布熵; 特征优化
中图分类号: TP391.42 **文献标识码**: A **文章编号**: 0372-2112 (2007) 03-0469-05

A Study on Optimum Classification Character Based on the Distributive Entropy of Euclidian Distance

BAO Ming, GUAN Lu-yang, LI Xiao-dong, TIAN Jing
(Institute of Acoustics, Chinese Academy of Sciences, Beijing 100080 China)

Abstract: An improved Kullback-Leiber distance is presented as a separable criterion for optimizing feature selection problems in pattern classification. A nonlinear parameter, Distributive Entropy of Euclidian Distance (DEED), is introduced and based on which, the ratio of between-class DEED to within-class DEED (JRd) is defined as a criterion for the feature selection. DEED is a nonlinear measure for learning feature space, which gives the congregation and information measure of learning samples space. According to the result of Gaussian data experiments, it is concluded that the larger JRd be, the better separability of learning samples would be.

Key words: improved KL distance; euclidian distance; distributive entropy; character optimization

1 引言

类别可分性准则设计是特征提取、特征选择的重要环节,常用的准则有 Fisher 距离、KL 距离等.在线性分类问题中,Fisher 线性多分类分析(FLDA Fisher Linear Discriminant Analysis)是一种有效的分类分析手段[1].它采用类间离散度矩阵与类内离散度矩阵的行列式的比值作为 Fisher 类别可分性准则.近年来在不同应用中,研究人员针对 Fisher 准则提出了许多改进方法[2],以提高统计分类算法的性能.

对于多维非线性分类问题,目前主要依靠分段线性逼近或利用非线性映射在高维空间获得具有线性可分特征等手段进行分析[3].例如在神经元非线性分类网络中,经非线性映射后在高维空间就可较好地符合 Fisher 准则[4].但因非线性变换已经改变了低维特征空间距离测度,若仅在低维特征空间进行 FLDA 分析,会造成非线性信息的缺失,不能正确地反映特征样本集非线性类别可分性.针对非线性问题,Solmomn Kullback 提出的代表互信息距离测度[5]的 KL 散度(Kullback Divergence),Marill 验证了其可用作为类别可分性准则[6].它采用了非线性标准熵,可适用于非线性分类情况,但在实际应用中存在计算困难等问题.

据此,本文提出了改进的 KL 距离作为样本可分性准则,并验证了它的有效性.进而在欧氏距离统计直方图的基础上,定义了无先验分布模型假设的欧氏距离分布熵(Distributive Entropy of Euclidian Distance DEED),并将其作为样本非线性类别可分性的分析工具.最后在实际运用中验证了这一度量标准的合理性.

2 改进的 KL 距离

KL 距离是 J_D 表征互信息的距离测度,但作为类别可分性准则,它存在表征同类样本自身聚集性能的类内自信息不足的缺点.由此引入自信息参量 H_i,给出如下假设:

收稿日期:2006-11-07;修回日期:2006-09-08

定义:

$$JI = \frac{J_D(1,2)}{H(1:1) + H(2:2)} \tag{1}$$

在两类问题的条件下,其信息散度 J_D 与各类自信息量 $H_i(i=1,2)$ 之和的比值越大,训练样本集的可分性能越好,对多类问题的情况可以类推. 其中:

$$J_D(1,2) = \int f_1(x_1, \cdots, x_k) - f_2(x_1, \cdots, x_k)) \log \frac{f_1(x_1, \cdots, x_k)}{f_2(x_1, \cdots, x_k)} dx_1, \cdots, dx_k \tag{2}$$

$f_1(x_1, \cdots, x_k)$ 和 $f_2(x_1, \cdots, x_k)$ 表示概率分布密度.

$$H(1:1) = - E\{\ln f_1(x)\}$$

为样本集的自信息(population entropy). 该假设验证如下.

在正态分布条件下 JI 的解析表达为:

$$JI = \frac{\frac{1}{2}\mathrm{tr}(\bar{V}_1 - \bar{V}_2)(\bar{V}_2^{-1} - \bar{V}_1^{-1}) + \frac{1}{2}\mathrm{tr}(\bar{V}_1^{-1} + \bar{V}_2^{-1})(\bar{u}_1 - \bar{u}_2)(\bar{u}_1 - \bar{u}_2)'}{\frac{1}{2}n_1 + \frac{1}{2}\ln|\bar{V}_1| + \frac{1}{2}n_1\ln(2\pi) + \frac{1}{2}n_2 + \frac{1}{2}\ln|\bar{V}_2| + \frac{1}{2}n_2\ln(2\pi)} \tag{3}$$

式中 n 为高斯序列维数,$\bar{V}_i$ 为方差矩阵,u_i 为均值向量. 以高斯模型为例:

$$p_i(X) = \frac{1}{(2\pi)^{p/2}|V_i|^{1/2}}\exp\left[-\frac{1}{2}(X - u_i)'V_i^{-1}(X - u_i)\right] \tag{4}$$

依据 Bayesian 准则,以 C_1, C_2 二分类为例:两类数据先验概率,$q_1 = q_2 = 0.5$、在分类正确风险为 0、误判风险相等的假设条件下,随机变量 X 的概率密度为 p_1, p_2 其似然函数为:

$$L(X) = X'(V_2^{-1} - V_1^{-1})X + 2X'(V_1^{-1}u_1 - V_2^{-1}u_2) + u_2'V_2^{-1}u_2 - u_1'V_1^{-1}u_1 - \log\frac{|V_1|}{|V_2|} \tag{5}$$

分类准则为:$X \in C_1$ 若 $L(X) \geq 0$;$X \in C_2$ 若 $L(x < 0)$.

利用随机试验方法得到两类二维高斯序列正确分类率与信息散度 J_D 的关系曲线示于图 1,图中粗实线为等方差的情况. 该图是文献[6]的结果. 图 2 为改进 KL 距离 JI 与样本正确分类概率的关系,图中粗实线为等方差的情况. 由图(1)、(2)可知,J_D 与 JI 均可有效的预估样本集的可分性能.

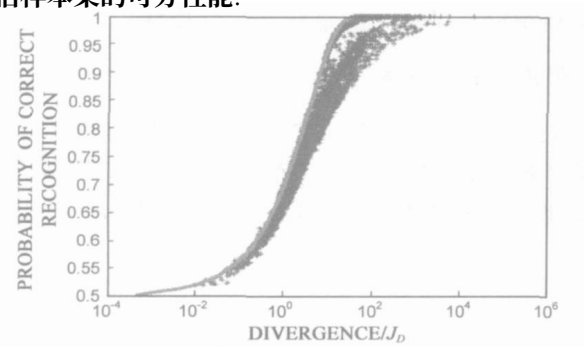

图 1 正确分类率与 J_D 的关系

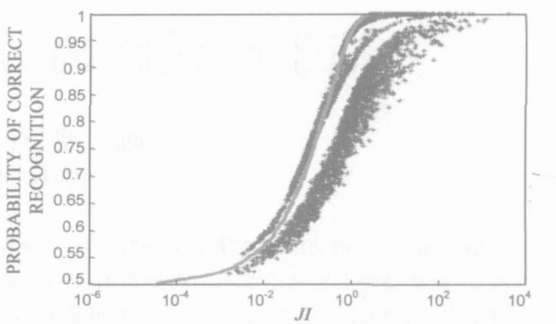

图 2 正确分类率与 JI 关系

3 欧氏距离分布熵及可其可分度分析

3.1 欧氏距离分布熵(DEED)定义

定义样本集中各点相对于欧氏空间某点的欧氏距离直方图的熵为"欧氏距离分布熵(DEED)". 它是一种可反映特征向量集不确定信息的度量参数. 对于多类样本:定义某类样本集各向量中相对于该类样本集均值向量的欧氏距离分布熵为自欧氏距离分布熵(ADEED);定义某类样本集各向量中相对于它类样本集均值向量的欧氏距离分布熵为互欧氏距离分布熵(CDEED).

几何上解释是同类样本集在欧氏空间中的聚集度高,会有助于学习;反之学习过程不易收敛. 由信息熵的严格 ∩型凸函数性质可知,概率空间的概率相等时,熵为最大;概率空间中任何使概率趋等的变动都会使熵增加. 综上推断,当样本集 ADEED 较小时,样本集在欧氏空间聚集度高,ADEED 较大时,样本集在欧氏空间聚集度低.

依据熵函数代数可加性质,针对多类问题,可以获得结论:各类样本的互欧氏距离分布熵之和与各类样本自欧氏距离分布熵之和的比值 $JRd = CDEED/ADEED$ 愈大,其样本集的分类性能愈强.

3.2 欧氏距离分布熵(DEED)计算

设 m 个 n 维特征向量

$$u_i = (x_{i,1}, x_{i,2}, \cdots, x_{i,n})^T, \quad i = 1,2,\cdots,m$$

组成 $n \times m$ 维特征矩阵 $W = [u_1, u_2, \cdots, u_m]$. 对 u_i 进行 L_2 归一化,取

$$u = Eu_i = E(x_{i,1}, x_{i,2}, \cdots, x_{i,n})^T, \quad i = 1,2,\cdots,m$$

(E 为统计平均)为 W 均值向量. 取 W 中各向量对中心向量欧氏距离组成 m 维数组

$$\delta.\ \delta_k(u_k, u) = (u_k - u)^T(u_k - u), \quad k = 1,2,\cdots,m$$

令 $\delta_{\max} = \max(\delta)$, $\delta_{\min} = \min(\delta)$ 则 $\delta_k \in [\delta_{\min}, \delta_{\max}]$. 设定常数 $N(N \ll m)$,取 $\Delta\delta = \frac{(\delta_{\max} - \delta_{\min})}{N}$,得 N 个连续区间 Φ_i, $i = 1,2,\cdots,N$. 数组 δ 中属于区间 Φ_i 的元素数目

为 p_i (i = 1 ,2 , …, N) ,易得 $\sum_{i=1}^{N} p_i = m$. 当 $m \to \infty$条件下,δ中元素属于区间 Φ_i 的概率

$$P_i = \frac{p_i}{m}, i = 1, 2, \cdots, N$$

其中 $\sum_{i=1}^{N} P_i = 1$. 据此可得欧氏距离分布直方图.

给定 $\alpha(0 < \alpha < 1)$,依直方图计算满足置信度为 1 - α的最大欧氏距离 δ_{max}. 剔除置信区间外的样本,重新计算直方图,获修正欧氏距离直方图. 对欧氏距离统计直方图求熵,即得欧氏距离分布熵 $E(P)$:

$$E(P) = -\sum_{i=1}^{N} P_i \log_2 P_i \qquad (6)$$

欧氏距离分布熵 $E(P)$,表征了样本集在欧氏测度条件下的空间分布信息量,但并没有包含其在欧氏空间尺度信息. 取置信度为 1 - α的距离参数δ_{max}的函数 $f(\delta_{max})$,与式(6) 右边相乘,可得包含样本集在欧氏空间聚集度及距离跨度信息的参量. 设 $f(\delta_{max}) = \delta_{max}$,得修正的欧氏距离分布熵 $\mathit{E}(P)$,简称为欧氏距离分布熵.

$$\mathit{E}(P) = -\delta_{max} \sum_{i=1}^{N} P_i \log_2 P_i \qquad (7)$$

在分类问题中,同类样本集的欧氏距离分布熵称为自欧氏距离分布熵. 用他类样本均值为参考点计算的欧氏距离分布熵称为互欧氏距离分布熵.

3.3　可分性标准有效性验证

在二维高斯数据条件下,采用与上文 KL 距离有效性验证相同的方法,获得两类二维高斯序列正确分类率与 FLDA 样本可分度迹判据 $J_t = \mathrm{tr}(\overline{S}_w^{-1}\overline{S}_b)$ 的关系如图 3 所示[1] ,图中粗实线为等方差的情况. 进一步获得两类二维高斯序列正确分类率与基于欧氏距离分布熵的样本可分度判据 JRb 关系如图 4 所示,图中粗实线为等方差的情况. 由图 4 可知,基于 DEED 多分类分析的样本可分度指标 JRd 可以表示样本的可分性能.

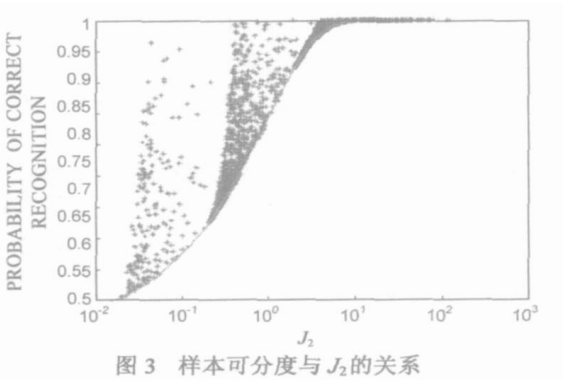

图 3　样本可分度与 J_2的关系

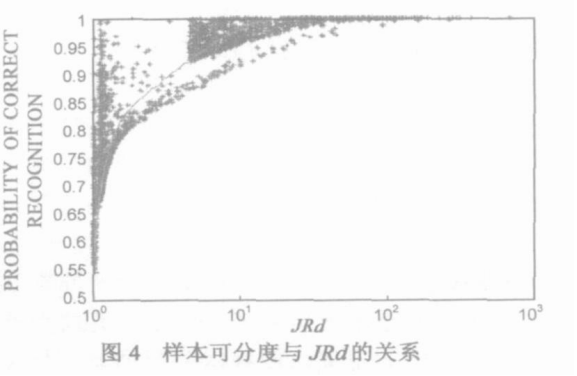

图 4　样本可分度与 JRd的关系

4　两类地面目标噪声实测样本集可分度判别试验

地面目标分类中轮式车与履带车分类是地面目标侦察的关键问题[7]. 本节将通过轮式车与履带车声信号分类,验证准则的有效性.

4.1　试验设计及特征提取

试验数据来自多次野外实验,采样频率为 1000Hz ,样本库中包含 5 种轮式车辆 3250 个样本,9 种履带车辆 4550 样本. 对数据提取了多种分类特征,其中轮式车标记为 W,履带车标记为 T. 采用 m-folder 交叉校验(m = 3) ,重复进行 20 次学习、测试,记录测试均值. 三种特征提取方法为:(1) 采用 25 个在频域非均匀分布的峰值带通滤波器对信号进行分带处理. 各频带能量组成分类特征,获得 25 维非均匀子带能量分类特征. (2) 仿照语音信号处理中 Mel 频率划分方式,设计频带划分映射公式为 $F_{smel} = 900 * \log10(1 + f_n/300)$,在类 Mel 域均分 25 带,按标准 MFCC 特征提取方法计算,获得 25 维改进 MFCC 分类特征. (3) 利用经典“db6 ”小波包对目标噪声信号进行尺度为 5 的小波包分析,获得 32 维小波包特征.

4.2　两类实测样本集三种可分度判别方法的比较

(1) 欧氏距离分布自、互熵及 JRd 的求取

对三种分类特征分别求取欧氏距离直方图、自、互熵及 JRd ,设 N = 256 ,距离修正的置信度为 99 %. 获典型欧氏距离直方图(图(5) (6) (7)) ,DEED 及 JRd 的统计均值示于表 1.

表 1　三种特征各 20 组样本集 DEED 计算均值

	WADEED	W/ T. CDEED	T/ W. CDEED	T. ADEED	ADEED	CDEED	JRd
非均匀子带	3. 8082	7. 3273	6. 9469	3. 4835	7. 2917	14. 2742	1. 9576
改进 MFCC	4. 7927	8. 5399	4. 1159	2. 2804	7. 0731	12. 6558	1. 7893
小波包(db6)	6. 2504	7. 6938	9. 227	6. 3796	12. 63	13. 9442	1. 1041

注:WADEED 代表轮式车自欧氏距离分布自熵,TADEED 代表履带车欧氏距离分布自熵,ADEED 两类自熵和 ,W/ T. CDEED 轮式车相对于履带车的互熵,T/ W. CDEED 履带车相对于轮式车的互熵, CDEED 两类互熵和

表 1 数据表明非均匀子带特征 JRd 最大,五层小波包系数特征 JRd 最小. 由 JRd 可分性标准估计非均

匀子带特征分类性能最优,改进 MFCC 特征分类性能次之,五层小波包系数特征分类性能最弱.

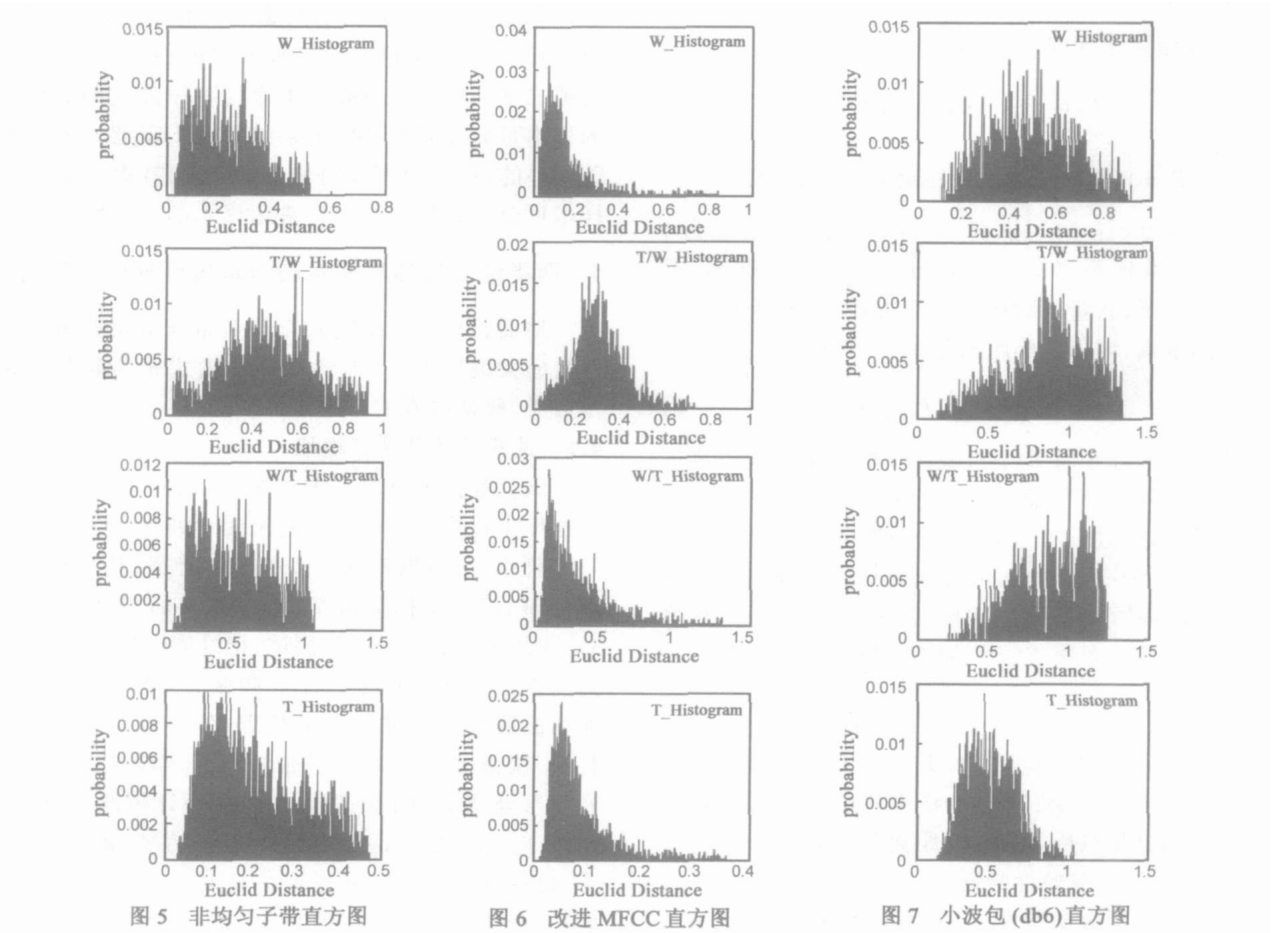

图 5 非均匀子带直方图 图 6 改进 MFCC 直方图 图 7 小波包 (db6) 直方图

(2) 样本集 FLDA 迹距离 J_t 及两分类聚类结果

计算三种分类特征的 FLDA 迹距离 J_t[1],其统计均值示于表 2. 由 J_t 类别可分性准则可知:改进 MFCC 特征分类性能最优,非均匀子带能量特征分类性能次之,五层小波包(db6)系数特征可分性能最弱. 采用 K-means 聚类分析,各特征的聚类正确率统计均值如表 2 所示. 结果表明:非均匀子带能量特征 T、W 两类聚类性能均为最优;改进 MFCC 特征 T 类聚类性能优于五层小波包系数特征,W 类聚类性能弱于五层小波包系数特征. 若以两类聚类平均正确率标准判定,非均匀子带能量特征样本集聚类正确率最高,MFCC 特征样本集聚类性能略高于五层小波系数特征.

表 2 三种特征可分度迹判据 J_t 计算结果及聚类分析结果

	J_t	W 类聚类正确率	T 类聚类正确率	W、T 聚类平均正确率
非均匀子带	1.9347	0.7145	0.9123	0.8134
改进 MFCC	2.3569	0.6316	0.9470	0.7893
小波包(db6)	1.7316	0.7012	0.8748	0.7880

4.3 分类算法验证

对三种分类特征,分别利用 KNN(K nearest neighbor $K=3$,参考样本数为学习样本集的样本数) 分类算法;神经元三层 BP 分类算法,其输入单元为 25,输出单元数为 2,中间隐层单元数 12;基于 RBF(Radial Basis Function) 核的 CSVM(C—Support Vector Classification) 分类算法测试,分类结果示于表 3.

表 3 不同分类算法下三种特征量分类识别正确率

	轮式车 W(%) 正确识别率			履带车 T(%) 正确识别率			两类平均正确识别率(注)		
	KNN	SVM	BP	KNN	SVM	BP	KNN	SVM	BP
非均匀子带	95.7	97.20	97.30	96.90	98.90	97.00	96.30	98.05	97.15
改进 MFCC	96.5	96.60	94.20	96.80	98.00	95.70	96.65	97.30	94.95
小波(db6)	93.4	95.10	96.18	92.20	96.20	88.39	92.80	96.65	92.29

表 3 表明:采用 KNN 分类算法,三种特征分类试验结果符合 J_t 判据估计结论. 利用 CSVM 分类算法及神经元 BP 分类算法测试,三种特征分类试验结果与 JRd 标准估计结论相符. 分析可知,KNN 分类计算中不包含特征空间上的非线性变换,因而 FLDA 的迹类别可分度准则 J_t 估计结果与 KNN 分类结论相符. 利用神经元网络进行分类计算,由于存在特征空间非线性映射,采用基于欧氏距离分布熵的分析方法及可分度准则 JRd 作为样本可分度指标比采用基于 Fisher 分类判别准则的样本可分度指标 J_t 更能反映样本特征的非线性类别可

分性.

5　结论

模式分类理论中有多种类别可分性准则.由于学习问题的复杂性,很难获得一种客观评价标准.欧氏距离分布熵,作为特征空间分类距离的非参数分析方法,采用无先验模型假设的熵标准,获得了包含非线性信息的类别可分性准则指标 JRd,弥补了传统的可分度标准 J_t 等线性准则中非线性信息缺失的不足.并在地面目标识别试验中得到了验证.

文中通过高斯模型计算试验验证了:JRd 愈大,其样本集的分类性能愈强这一结论的正确性.在轮式车与履带车二分类试验中,采用欧氏距离分布熵的分析方法获得了较好的预测结果.

欧氏距离分布熵从学习理论及信息论中均可得到合理的解释,可以在特征优化及智能选择中应用;另外欧氏距离直方图所蕴涵的测度信息也可为学习机的动态学习提供加权参数.分类问题若采用非欧氏距离的其它距离测度,也可用类似的手段进行分析研究.

参考文献:

[1] Richard O Duda, et al. Pattern Classification (Second Edition) [M]. New York: Wiley, 2000. 117 - 124.

[2] Zhang Yanwu, Arthur B Baggeroer. The total variance of a periodogram-based spectral estimate of a stochastic process with spectral uncertainty and its application to classifier design [J]. IEEE Trans Signal Processing, 2005, 53(12): 4556 - 4567.

[3] Covet T M. Geometrical and statistics properties of system of linear inequalities with applications in pattern recognition [J]. IEEE Trans Electronic Computers, 1965, 14: 326 - 334.

[4] Haykin S. Neural Networks a Comprehensive Foundation (Second Edition) [M]. BeiJing: Tsinghua University Press & Prentice Hall, 1999. 199 - 202.

[5] Solomom Kullback. Information Theory and Statistics [M]. New York: John Wily & Sons Inc, 1959. 6 - 7.

[6] T Marill, D M Green. On the effectiveness of receptors in recognition systems [J]. IEEE Trans Information Theory, 1963, 9(1): 11 - 17.

[7] Dan Li, et al. Detection Classification and Tracking of Targets [J]. IEEE Signal Processing Magazine, 2002, 19(2): 17 - 29.

作者简介:

鲍　明　男,1973 年生于湖北武汉,中国科学院声学研究所博士研究生,研究方向为声学信号处理,统计学习及分类等.
E-mail: baoming@mail.ioa.ac.cn

管鲁阳　男,1979 年生于河南南阳,中国科学院声学研究所博士研究生,研究方向为声学信号处理,统计学习及分类等.
E-mail: guanluyang@mail.ioa.ac.cn