自动目标识别算法的识别率比较方法*

何 峻** 付 强
国防科技大学ATR重点实验室, 长沙 410073

摘要 针对模式识别领域中所关注的自动目标识别(automatic target recognition, ATR)算法识别率比较这一共性问题, 首先归纳分析了现有的识别率比较方法并分别指出其局限性. 提出了一种新的基于Bayes分析的识别率比较方法——后验概率比较法. 然后运用该方法分析了ATR算法识别率比较过程中所特别关注的选优和排序这两个典型问题, 证明了应用最大似然原理的合理性. 最后定量分析了比较结果的可信程度与所需的测试样本容量之间的约束关系, 所得到的图表能够有效指导ATR算法评估试验的设计和测试数据采集工作.

关键词 模式识别 算法 评价 不确定分析

近年来, 自动目标识别(automatic target recognition, ATR)技术在SAR图像检测、医学CT诊断、生物特征识别、手写/语音鉴别等多个模式识别的应用领域中取得了长足进步. ATR算法是实现ATR技术的主要研究内容之一, 对ATR算法进行性能评估也一直是模式识别问题的关注热点. 衡量ATR算法性能优劣的一个基本评价指标就是算法对待识别对象的正确识别概率, 简称“识别率”(某些领域习惯上采用等价的“误识率”[1]).

本文所关注的问题可归结为: 给定测试集, 以识别率为指标比较不同ATR算法的性能优劣. 这里将ATR算法识别率的估计和比较作为两个不同问题区分对待. 诚然, 估计问题的目的之一就是为了比较不同ATR算法的识别率, 而且在样本容量趋于无限大时估计的结果就能够实现识别率的比较, 二者没有太大的区别. 但现实中受限于测试样本容量, 对ATR算法识别率的估计和比较还是各有侧重: 识别率估计主要关心估计精度[2]; 而比较则是以识别率高低为依据进行推断.

1 识别率比较方法回顾

总的来说, 现有的识别率比较方法基本上都是采用频率派的经典统计理论来设计假设检验或统计推断问题来实现比较. 本节主要对这些方法进行归纳总结和分析.

1.1 点估计值比较法

最直接的一种方法是先对各ATR算法的识别率 p(或等效的误识率 e)进行点估计, 然后用所得到的估计值 $\hat{p}$ 进行比较.

ATR算法的单次识别结果可用一个二值变量 x 来表示: $x=1$, 表示ATR算法正确识别; $x=0$, 表示错误识别. 测试样本容量为 n 时, 可用序列 x_i ($x_i=1$ or 0, $i=1, 2, \cdots, n$)记录整个识别过程. 若用 X 表示总的正确识别次数, 即

$$X=\sum_{i=1}^{n} x_i \tag{1}$$

则 X 是一个服从二项分布的随机变量 $X\sim B(n, p)$. 识别率的估计值 $\hat{p}$ 的计算式为

2008-02-20 收稿, 2008-03-19 收修改稿
* 武器装备预研重点基金(6140522); 装备预先研究项目(51301050102)资助
** E-mail: hisjune@163.com

$$\hat{p}=\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{X}{n} \tag{2}$$

点估计值比较法的确是一种识别率的比较方法，而且也在被广泛地自觉或不自觉地运用着．但仅使用点估计值难以直接定量计算所得到比较结果的可信程度，也不能分析和计算为达到某个置信度所需的测试样本容量 n.

1.2 区间估计值比较法

点估计值不能给出估计结果的置信度和变动范围，而区间估计显然给出了有关 p 的更多有用信息．因此，利用互不重叠的识别率区间估计值进行比较也就一度成为 SAR ATR 评估中的常规方法[3,4]．

区间估计值比较法所面临的问题主要有：(i) 在样本容量一定的情况下对识别率进行区间估计，置信度和区间长度相互制约．(ii) 识别率的置信区间不是区间数(关于区间数的定义可参考文献[5]或文献[6]的引述)，其“正确性”仅在一定置信度下成立．这意味着：以区间估计值进行 ATR 算法比较，实质上是一种以具有不确定性的“命题”为基础的推理过程．而按照不确定推理的一般原则，多个命题的“逻辑与”的不确定性值应小于等于其中任意一个命题的不确定性值[7]．因此，以区间估计值进行 ATR 算法比较，最终结论的置信度小于等于其中任意一个置信区间的置信度．基于上述原因，为得到一个高置信度的比较结论，区间估计值比较法需要估计出具有很窄宽度和很高置信度的置信区间，而这种高精度的区间估计对测试样本容量的需求通常是很大的[8,9]．

1.3 区间差值法

区间差值法和区间估计值比较法的区别在于，区间差值法计算的是两个 ATR 算法的识别率 p_1 和 p_2 之间差值的置信区间．如果这个区间不包含 0，则认为 p_1 和 p_2 之间存在“显著差异”(significantly different)．若采用近似正态模型，置信度 $1-\alpha$ 下 p_1-p_2 的置信区间[10]为

$$\left[(\hat{p_1}-\hat{p_2})-z_{\alpha/2}\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}},\right.$$
$$\left.(\hat{p_1}-\hat{p_2})+z_{\alpha/2}\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}\right] \tag{3}$$

(3)式中 n_1 和 n_2 分别表示对 ATR 算法 1 和算法 2 进行测试的样本个数．若采用同一批样本进行测试，可令 $n=n_1=n_2$.

区间差值法本质上是一种统计推断方法，其统计意义明显，操作简单，在雷达 ATR 技术的评估中也得到了广泛应用[3,4]．其他的一些识别率比较方法[1]，实质上是区间差值法的变形或简化．区间差值法的主要缺陷在于一次只能比较两个 ATR 算法．如果要进行多个 ATR 算法的识别率比较，需要进行多次两两比较，然后将这些两两比较的结果合成总的比较结果．而多个假设检验结果的合成将导致总的比较结果的置信度降低．

1.4 R&S法

R&S 法是 Gibbons[11] 针对总体排序选优而提出的一种方法．R&S 法并不是要确定 m 个总体的成败率 p_i，而是要从 m 个具有二项分布的总体中选择具有最大 p 值的那个总体．在确定了比较结果的显著性水平 α 之后，决定所需最小测试样本容量 $n_{\min}$ 的是最优识别率 $p_{[m]}$ 和次优识别率 $p_{[m-1]}$ 之间的差值 $\delta^*=p_{[m]}-p_{[m-1]}$．R&S 法操作起来非常简单：首先，确定测试集(按照文献[11]给出的数据表)并按照(2)式算出各 ATR 算法识别率的点估计值 $\hat{p_i}$ $(1\leqslant i\leqslant m)$；然后，按 $\hat{p_i}$ 的大小进行排序；最后，选择 $\hat{p_i}$ 最大的 ATR 算法作为最优算法.

R&S 法作为一种比较 ATR 算法识别率的方法，其缺陷是明显的．它只能从 m 个算法中选出最优(识别率最高)算法；当 $m>2$ 时 R&S 法就不能有效地对所有算法进行排序.

1.5 Wald 序贯检验法

作为序贯分析的创始人[12]，Wald 提出了一种序贯的假设检验方法．Wald 序贯检验法有多个具体应用形式，与识别率比较问题密切相关的是对两个二项分布总体均值的检验方法．不同于前面的比较方法中度量“均值差”的分析思路，Wald 考察两个算法识别率的“效率”比．这里“效率”(efficiency)被定义为 $k=p/(1-p)$，而算法 1 和算法 2

的效率比为[13]

$$u=\frac{k_2}{k_1}=\frac{\frac{p_2}{1-p_2}}{\frac{p_1}{1-p_1}}=\frac{p_2(1-p_1)}{p_1(1-p_2)} \tag{4}$$

Wald序贯检验过程中，如果发现 $u<1$，说明ATR算法1比2优异(识别率高)；$u=1$，说明二者等同；如果 $u>1$，则说明算法2更优异．具体的Wald序贯检验的过程可参考文献[13]，这里不再赘述．

Wald序贯检验法采取序贯检验策略，其优点有：当两个实际算法的识别率 p_1 和 p_2 之间的差异比预期的灵敏度 δ大时，Wald序贯检验法实际所使用的测试样本数 n 能够大大减少；并且该方法可以同时设置假设检验的两类错误容限，而前面介绍的几种比较方法只能设置第一类错误概率——显著性水平．将Wald序贯检验法直接用于ATR算法识别率比较的局限性也是明显的：首先，Wald序贯检验法只能比较两个算法的识别率高低；其次，它只能预测所需测试样本容量的期望值．针对第一点不足，Catlin[10]提出了改进的Wald序贯检验的识别率比较方法——Wald MSRB法．

1.6 Wald MSRB法

Wald MSRB法是Catlin针对SAR图像ATR背景问题所提出的一种识别率比较方法．前面介绍的Wald序贯检验法和MSRB法[14](modified sequentially rejective Bonferroni procedure)是Wald MSRB方法的两个基础．概括地讲，Wald MSRB法就是用MSRB法来处理 m 个ATR算法两两比较的Wald序贯检验法结果，并从中选出具有最高识别率的ATR算法．

Wald MSRB法的核心在于*Bonferroni*不等式的应用，实际上是一种采用了Wald序贯检验作为两两比较手段的MRSB法．其优点主要集中在测试样本容量上．相对于同等的显著性要求，Wald MSRB法实际所需的测试样本容量大幅降低．Wald MSRB法的不足可以概况为以下二个方面：(i)只能选出最优的算法而不是对多个算法按识别率高低进行排序；(ii)文献[10]以实际应用和仿真手段说明了该方法对测试样本容量的需求有较大下降，却仍无法预测需要的测试样本容量，只能以R&S法所提供的测试样本容量 $n_{\min}$ 作为所需的测试样本上限．

1.7 存在问题

通过上述论述不难发现，目前的各种ATR算法识别率比较方法存在以下两个主要问题．

1.7.1 所需测试样本容量仍然偏大 现有的这些识别率比较方法都是用频率派的统计观点来处理识别率的比较问题，因此决定所需的测试样本容量的因素主要有3个：(i)结论的显著性水平 α；(ii)所要比较的ATR算法数目 m；(iii)这些算法识别率 $p_i(1\leqslant i\leqslant m)$ 之间的差异程度(灵敏度) δ．当对结论的显著性水平 α、ATR算法数目 m 和灵敏度 δ都有比较高的要求时，所需的测试样本容量将非常大(可参考文献[1，10]的分析结果)．如何降低测试样本容量，进而有效减少数据采集代价是一个极具现实意义的问题．

1.7.2 比较过程缺乏推理模型支持 尽管现有的比较方法都将识别率作为常量，但受限于测试样本容量，所得到的结论都将带有一定的不确定性．而某些比较方法中还需要综合多个具有不确定性的结论(命题)来得到最终的比较结果(合成命题)．从这个意义上说，ATR算法识别率比较的本质就是不确定推理问题．从前面的论述中不难看出，目前的各种比较方法在推理模型上的考虑都还比较欠缺．

2 后验概率比较法

2.1 需求分析

前一节简要回顾了现有ATR算法识别率比较方法，并对它们的特点和局限性进行了分析和讨论．通过上述讨论，总结出对ATR算法识别率比较方法所希望具有的功能和特性如下：

(1)能够充分利用先验信息；

(2)能够根据停止法则实施中止，减少测试时间和样本采集代价；

(3)允许ATR算法测试被强制打断，但仍然能够利用已有的测试结果给出结论；

(4)比较结果具有明确的统计意义；

(5)有效实现多个ATR算法的比较(包括选优

和排序);

(6) 具备半定制能力，即能够根据具体的评测目的来设计具体的流程方案.

2.2 基本思想及操作流程

针对上述几点需求，本文提出了一种适用于多个ATR算法识别率比较的新方法——后验概率法(posterior probability procedure)，实现了所期望的功能需求. 后验概率法的基本思想可概括如下:

基于Bayes分析的理论体系，将ATR算法的识别率 p 作为随机变量来处理. 多次的测试和先验信息通过Bayes公式实现融合，而所得到 p 的后验概率分布包含了ATR算法识别率的全部(不确定性)信息. 为实现ATR算法识别率的比较(如选优或者排序)，根据具体的比较问题抽象出特定事件，计算这些事件的发生概率，以此作为识别率比较问题的推理判据. 后验概率法的逻辑基础是概率推理，核心问题在于针对各种事件的后验概率计算.

图1给出后验概率法的操作流程图，下面结合图1阐述其具体步骤.

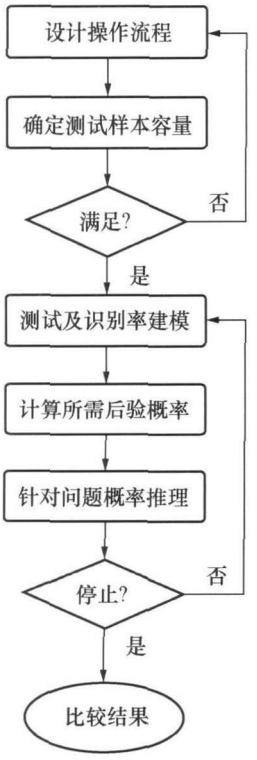

图1 后验概率法操作流程

步骤**1** 设计操作流程

这一步主要是根据具体的ATR算法比较问题，设计出符合实际需求的操作流程，主要包含比较内容、停止法则及检验步长3个方面内容. 其中，比较内容指针对具体问题而所需要知道的特定事件(后验)概率. 例如，为选出最优ATR算法，决策者需要知道“ATR算法 i 识别率最优”这 m 个特定事件的概率. 停止法则指的是停止整个操作的条件，如比较结果的置信度大于0.95、所用的测试样本容量 $n>1000$ 等；而检验步长指的是对 k 个样本进行测试后再依照停止法则决定是否中止.

步骤**2** 确定测试样本容量

这一步主要是根据步骤1所设计出的具体操作流程确定测试样本容量，并建立测试集. 如果所需要的测试样本容量过大，则需要调整步骤1所给出的设计结果(例如适当降低对比较结果置信度的要求). 有关测试样本容量的分析将在第4节做详细讨论.

步骤**3** 测试及识别率建模

这一步首先按照步骤1中确定的步长 k，选择 k 个新样本对 m 个ATR算法分别进行测试. 然后根据具体的测试结果和事先确定的先验信息，依据贝叶斯理论建立每个ATR算法识别率的概率分布模型 $P_i(1\leqslant i\leqslant m)$. 本文推荐采用β分布来描述ATR算法识别率，其建模过程和部分重要结论如下:

根据Bayes公式，经过 n 个样本测试后的ATR算法识别率 P 的后验概率密度函数可表述为

$$\pi(p\mid x,n)=\frac{\pi(p)f(x,n\mid p)}{\int_{\Theta}\pi(p)f(x,n\mid p)\mathrm{d}p} \tag{5}$$

(5)式中 $\pi(p)$ 表示识别率 P 的先验分布；$f(x,\ n\mid p)$ 为似然函数，表示样本量 n 时ATR算法正确识别的次数为 x 的概率；Θ 为 P 的值域. 显然，识别率的后验概率分 $\pi(p\mid x,\ n)$ 由似然函数 $f(x,\ n\mid p)$ 和先验信息 $\pi(p)$ 共同决定.

假设 $X\sim B(n,\ p)$ 可得

$$f(x,n\mid p)=\binom{n}{x}p^{x}(1-p)^{n-x} \tag{6}$$

再由ATR算法的识别结果 x 为0—1变量，因此对ATR算法的测试可以看作成败型试验. 成败型试验的先验分布有各种形式的假设，其中以β分布最为常见，即假设先验信息π(p)为β概率密度函数

$$\pi(p)=\mathrm{betapdf}(p;a,b)=\frac{1}{\mathrm{Beta}(a,b)}p^{a-1}(1-p)^{b-1}\quad(0<p<1) \tag{7}$$

(7)式中Beta(a，b)表示参数为(a，b)的Beta函数.

将(6)式和(7)式带入(5)式进行积分运算可得识别率 p 的后验概率密度函数为

$$\pi(p\mid x,n)=\frac{1}{\mathrm{Beta}(x+a,n-x+b)}p^{x+a-1}\cdot(1-p)^{n-x+b-1}\quad(0<p<1) \tag{8}$$

很明显，ATR算法识别率的后验概率同样服从β分布. 采用β分布作为先验分布的意义就在于如同预先已经做了 $a+b$ 次识别测试，其中正确识别 a 次；再加上实际测试时的 n 次识别测试，等效样本容量 $n+a+b$，等效正确识别 $x+a$ 次.

步骤**4** 计算所需后验概率

如前所述，特定事件的后验概率是进行ATR算法识别率比较的推理判据. 下面针对ATR算法比较中典型的选优问题为例，说明如何计算所需的后验概率.

设共有 m 个ATR算法参与比较，记ATR算法 $i(1\leqslant i\leqslant m)$ 的识别率为 P_i. 当执行到步骤4时，已经测试了 n 个样本，ATR算法 i 正确识别的样本数为 x_i. 那么，ATR算法 i 为最优算法(识别率 p_i 最大)的概率 Pb_i 可表述为

$$Pb_i=P\{P_i>P_j\mid j\neq i,j=1,2,\cdots,m\} \tag{9}$$

一般可以认为ATR算法 i 和ATR算法 j 的识别过程是相互独立的，故 m 个ATR算法识别率的联合概率密度函数为

$$\pi(p_1,p_2,\cdots,p_m)=\prod_i\pi(p_i\mid x_i,n) \tag{10}$$

采用(8)式的β分布模型描述识别率，则(10)式为

$$\pi(p_1,p_2,\cdots,p_m)=\prod_i\mathrm{betapdf}(p_i;x_i+a_i,n-x_i+b_i) \tag{11}$$

由(11)式，(9)式可进一步写作

$$Pb_i=\int_{P_i>P_j\,(j\neq i)}\cdots\int\prod_i^m\mathrm{betapdf}(p_i;x_i+a_i,n-x_i+b_i)\cdot\mathrm{d}p_1\,\mathrm{d}p_2\cdots\mathrm{d}p_m=\int_0^1\left\{\left[\prod_{j\neq i}^m\mathrm{betacdf}(p_i;x_j+a_j,n-x_j+b_j)\right]\cdot\mathrm{betapdf}(p_i;x_i+a_i,n-x_i+b_i)\right\}\mathrm{d}p_i \tag{12}$$

(12)式中betacdf(p；a，b)表示β分布函数，即

$$\mathrm{betacdf}(p;a,b)=\int_0^p\mathrm{betapdf}(t;a,b)\mathrm{d}t \tag{13}$$

由于β概率密度函数和β分布函数都是比较常见的函数，可以借助工具软件(如Matlab)用数值积分法按(12)式计算出 m 个 Pb_i.

步骤**5** 针对问题概率推理

针对比较问题，以步骤4所得到特定事件(命题)的后验概率为基础，针对问题进行概率推理. 然后再对照停止法则，决定是否中止检验. 上例中的具体问题是选出最优的ATR算法，因而推理过程是简单而直接的，即选择具有最大的后验概率 P_i 的ATR算法 i 作为最优算法. 如果停止法则是优选结论的置信度不小于0.9，即 $Pb_{\max}=\max\{Pb_i\}\geqslant 0.9$，则根据具体的 $Pb_{\max}$ 值决定操作是继续检验(转入步骤3)还是给出最终的比较结果.

2.3 停止法则

通过对后验概率法操作流程的直观的分析不难发现，后验概率法的这种序贯检验过程由于外界条件限制(如测试样本使用完)而被强制中止后，仍然可以根据已有的测试结果对ATR算法的识别率进行比较，只是此时所得结论的置信度不足以满足预期要求. 相比之下，如果在Wald MSRB法中提前中止检验，则不能给出有效的检验(比较)结果. 造成这种巨大差别的本质原因在于：本文所提出的后验概率法是基于Bayes理论框架的，而以往的那些识别率比较方法则是基于频率统计学派思想的.

总的来说[12]，“频率学派本质上是原始方案的奴隶(包括所要采用的停止法则的选择). ……对方案作任何改变都破坏给出有效的频率派结论的可能性. 另一方面，Bayes 派可以比预期的停止得早，或比预期的继续得长，并仍然能得出有效的 Bayes 派的结论.” 对于停止法则原理的深入讨论这里不再展开，有兴趣的读者可以参考文献[12]. 这里强调指出：本节所提的后验概率法中虽然包含了停止法则，但以后验概率为判据的比较结论却与具体的停止法则无关，即所得结论的置信度仅与测试结果和先验信息有关.

3 最大似然原理的应用

上节所提出的后验概率法实际上是采用多元联合概率密度函数进行序贯 Bayes 分析，将 ATR 算法识别率比较的问题转换为计算特定事件的后验概率进行不确定推理，并以后验概率值来定量度量所得结论的不确定程度. 本节运用该方法，结合 ATR 算法识别率比较时所特别关注的两个典型问题——选优和排序，证明识别率比较过程中的应用最大似然原理的合理性.

3.1 选优中的最大似然法

以识别率为指标对 ATR 算法进行选优，如果严格按照后验概率法的操作流程，在每次序贯检验中都需要对 m 个参评的 ATR 算法按(12)式计算其为最优算法的后验概率 P_i，选出 P_i 最大的算法 i 作为当前的最优算法，并对照具体的停止法则决定是否中止检验. 在实际的 ATR 算法评估中，往往只需要先将这 m 个算法按照后验分布的均值 $E[\pi(p_i)]=(x_i+a_i)/(n+a_i+b_i)$进行排序，计算最大均值 $E[\pi(p_{(m)})]=\max\limits_{1\leqslant i\leqslant m}\{E[\pi(p_i)]\}$所对应的算法为最优这一事件的概率 $Pb_{(m)}$. 如果 $Pb_{(m)}>1/2$，则不需要计算其余的 $m-1$ 个 Pb_i. 这与第 2 节中的点估计值比较法的做法类似，其中所蕴涵的深层原因可用如下定理解释.

定理 1 设 m 个随机变量 P_i 的联合概率分布为 $f(p_1,p_2,\cdots,p_m)=\prod\limits_i^m \text{betapdf}(p_i;x_i+a_i,n-x_i+b_i)$. Q 为样本空间 $S=[0,1]\times[0,1]\times\cdots\times[0,1]$ 中 $f(p_1,p_2,\cdots,p_m)$的极大值点. 定义事件 $H_0=\{\exists i\neq j,P_i=P_k\}$和 $H_i=\{P_i>P_j\mid j\neq i\}$ $(i,j=1,2,\cdots,m)$，并将这些事件按照$(x_i+a_i-1)/(n+a_i+b_i-2)$的大小进行排序，记排列结果为 H_0，$H_{(1)}$，$H_{(2)}$，…，$H_{(m)}$，满足

$$\frac{x_{(1)}+a_{(1)}-1}{n+a_{(1)}+b_{(1)}-2}\leqslant\frac{x_{(2)}+a_{(2)}-1}{n+a_{(2)}+b_{(2)}-2}\leqslant\cdots\leqslant\frac{x_{(m)}+a_{(m)}-1}{n+a_{(m)}+b_{(m)}-2}$$

有如下命题成立

(1) $\sum\limits_{i=1}^m P\{H_i\}=1$，$\sum\limits_{i=1}^m P\{H_{(i)}\}=1$；

(2) Q 唯一存在，若

$$\frac{x_{(m-1)}+a_{(m-1)}-1}{n+a_{(m-1)}+b_{(m-1)}-2}<\frac{x_{(m)}+a_{(m)}-1}{n+a_{(m)}+b_{(m)}-2}$$

则 $Q\in H_{(m)}$且$\lim\limits_{n\to\infty}P\{H_{(m)}\}=1$.

证明 详见文后的附录 1.

定理 1 的意义在于：

(1) 当采用上节所提出的后验概率法进行以识别率为指标的 m 个 ATR 算法的选优，按照(12)式计算每个算法最优的后验概率 Pb_i，定理 1 保证有 $\sum Pb_i=1$. 所以某个$Pb_i>0.5$ 就可以断定它是最大的后验概率值.

(2) 考虑到 betapdf$(p;x_i+a_i,n-x_i+b_i)$的均值为 $E[P_i]=(x_i+a_i)/(n-x_i+b_i)$，在 n 较大时与 $(x_i+a_i-1)/(n+a_i+b_i-2)$ 相当贴近，因此实际上可以选择识别率后验概率均值最大的算法作为最优算法，定理 1 保证这种选择结果包含了最大的后验概率事件 Q，并且选择正确的可能性 $P\{H_{(m)}\}$ 随测试样本容量 n 的增大依概率收敛到 1.

在 Bayes 分析中，当对先验信息一无所知时常假设先验分布为值域范围内的均匀分布，称为无信息先验. 对于这里的 ATR 算法识别率比较问题而言，无信息先验是$\pi(p)$在(0，1)上的均匀分布，也即假定 $a_i=b_i=1$. 此时排序量$(x_i+a_i-1)/(n+a_i+b_i-2)$退化成 x_i/n，即识别率点估计值. 所以通常所采用的以识别率点估计值(测试值)来选择最优算法的做法，实际上正是遵循着最大似然原理，

而选择正确的可能性(置信度)在样本容量无限大时趋近于1.

3.2 排序中的最大似然法

同样，以识别率为指标对ATR算法进行排序，如果严格按照后验概率法的操作流程，在每次序贯检验中都需要计算 m 个参评的ATR算法的所有 $m!$ 种排列 $k_1k_2\cdots k_m$ 结果成立的后验概率 $Pb_{k_1k_2\cdots k_m}$. 然后选出 $Pb_{k_1k_2\cdots k_m}$ 最大的排列作为当前的排序结果，再对照具体的停止法则决定是否中止检验. 根据上述定义

$$Pb_{k_1k_2\cdots k_m}=\int_{P_{k_1}>P_{k_2}>\cdots>P_{k_m}}\cdots\int\prod_i^m \mathrm{betapdf}(p_i;x_i+a_i,n-x_i+b_i)\mathrm{d}p_1\mathrm{d}p_2\cdots\mathrm{d}p_m \tag{14}$$

但在实际计算时可以根据条件概率公式将(14)式改写成

$$\begin{aligned}&Pb_{k_1k_2\cdots k_m}=\\&P\{P_{k_1}>P_{k_2}>\cdots>P_{k_m}\}=\\&P\{P_{k_1}>P_{t_1}\mid t_1=k_2,k_3,\cdots,k_m\}\cdot\\&P\{P_{k_2}>P_{k_3}>\cdots>P_{k_m}\}=\cdots=\\&P\{P_{k_1}>P_{t_1}\mid t_1=k_2,k_3,\cdots,k_m\}\cdot\\&P\{P_{k_2}>P_{t_2}\mid t_2=k_3,\cdots,k_m\}\cdot\cdots\cdot\\&P\{P_{k_{m-1}}>P_{t_{m-1}}\mid t_{m-1}=k_m\}=\\&\prod_i^{m-1}\left\{\int_0^1\left[\prod_{t_i}\mathrm{betacdf}(p_{t_i};x_{t_i}+a_{t_i},n-x_{t_i}+b_{t_i})\right]\cdot\right.\\&\left.\mathrm{betapdf}(p_{k_i};x_{k_i}+a_{k_i},n-x_{k_i}+b_{k_i})\mathrm{d}p_{k_i}\right\}\end{aligned} \tag{15}$$

(15)式中 $t_i=k_{i+1},k_{i+2},\cdots,k_m$. 具体计算时可以先用数值积分的方法分别算出 $m-1$ 个条件概率 $\{\cdot\}$ 值，然后求积得到 $Pb_{k_1k_2\cdots k_m}$.

与选优问题类似，实际的ATR算法评估中往往是先按照这 m 个算法后验分布的均值 $E[\pi(p_i)]=(x_i+a_i)/(n+a_i+b_i)$ 进行排序，然后计算这一排序所对应的后验概率 $Pb_{(m)(m-1)\cdots(2)(1)}$. 如果 $Pb_{(m)(m-1)\cdots(2)(1)}>1/2$，不需要计算其余的 $m!-1$ 种排列所对应的后验概率.

定理2 设 m 个随机变量 P_i 的联合概率分布为 $f(p_1,p_2,\cdots,p_m)=\prod_i^m\mathrm{betapdf}(p_i;x_i+a_i,n-x_i+b_i)$. Q 为样本空间 $S=[0,1]\times[0,1]\times\cdots\times[0,1]$ 中 $f(p_1,p_2,\cdots,p_m)$ 的极大值点. $k_1k_2\cdots k_m$ 为1，2，…，m 的一个排列. $\forall i\neq j(i,j=1,2,\cdots,m)$，有

$$(x_i+a_i-1)/(n+a_i+b_i-2)\neq(x_j+a_j-1)/(n+a_j+b_j-2).$$

特别地，将按 $(x_i+a_i-1)/(n+a_i+b_i-2)$ 从大到小顺序的排列记为 $(m)(m-1)\cdots(2)(1)$. 定义事件 $H_0=\{\exists k_i\neq k_j,P_{k_i}=P_{k_j}\}$ 和 $H_{k_1k_2\cdots k_m}=\{P_{k_1}>P_{k_2}>\cdots>P_{k_m}\}$. 有如下命题成立

(1) $\sum_{k_1k_2\cdots k_m}^{m!}P\{H_{k_1k_2\cdots k_m}\}=1$;

(2) Q 唯一存在;

(3) $Q\in H_{(m)(m-1)\cdots(2)(1)}$ 且 $\lim_{n\to\infty}P\{H_{(m)(m-1)\cdots(2)(1)}\}=1$.

证明 证明方法与定理1类似，略.

定理2中的已知条件“$\forall i\neq j(i,j=1,2,\cdots,m),(x_i+a_i-1)/(n+a_i+b_i-2)\neq(x_j+a_j-1)/(n+a_j+b_j-2)$”也即假设这 m 个ATR算法经 n 个样本测试后所得到的后验概率 $\mathrm{betapdf}(p;x_i+a_i,n-x_i+b_i)$ 各不相同. 如果这个条件不满足可以再测试一些新的测试样本，或者直接将所有具有相同的后验概率算法“合并”为一个ATR算法(因为从最终测试结果看，这些算法的识别率完全相同)参与排序.

定理2的意义在于

(1) 按照(14)式或(15)式计算 $m!$ 种排列可能性的后验概率 $Pb_{k_1k_2\cdots k_m}$，定理2保证有 $\sum_{k_1k_2\cdots k_m}Pb_{k_1k_2\cdots k_m}=1$. 所以某种排序结果的 $Pb_{k_1k_2\cdots k_m}>0.5$ 就可以断定它具有最大后验概率值.

(2) n 较大时 $\mathrm{betapdf}(p;x_i+a_i,n-x_i+b_i)$ 的均值 $E[P_i]=(x_i+a_i)/(n-x_i+b_i)$ 与 $(x_i+a_i-1)/(n+a_i+b_i-2)$ 相当贴近，因此实际上可按识别率后验概率的均值从大到小进行排序，定理2保证：这种排序方法是一种最大似然方法，排序正确的可能性 $P\{H_{(m)(m-1)\cdots(2)(1)}\}$ 随测试样本容量 n 的增大依概率收敛到1.

显然，通常采用的以识别率点估计值(测试值)来对 m 个ATR算法进行排序的做法实际上正是应用到最大似然原理，而这种经验做法结果的置信度在样本容量无限大时趋近1.

4 测试样本容量分析

预先的测试样本容量定量分析能够有效地指导评估试验的设计和数据采集工作. 本节针对这一需求围绕样本容量与比较结果的置信度之间的相互约束关系展开研究，分两种典型情况展开. ATR算法评估的一种典型情况是测试集的样本容量是一定的. 这种情况下往往需要预先分析一定数量的测试样本所能够保证的识别率比较结果的置信度到底有多大，在此基础上才能设计合理的评估试验和操作流程. ATR算法评估的另一种典型情况是对所得比较结论提出一定的置信度要求. 对于这种情况下更应该分析所需的测试样本容量，预计数据采集代价.

通过前面几节的分析不难发现，以识别率为指标的ATR算法比较结果的置信度 Pb 与样本容量 n、参评的ATR算法数 m、具体的测试结果 x_i(通常用等效的识别率测试值 $p_i = x_i/n$ 等效，$i = 1, 2, \cdots, m$)及先验信息等有关. 多数ATR算法评估中不考虑识别率的先验信息，因此下面仅对无信息先验的情况进行分析.

4.1 两个算法的比较

当参评的算法数 $m=2$ 时，ATR算法的选优和排序是等价的，此时 $Pb_{(2)} = Pb_{(2)(1)}$. 结合大多数模式识别领域中ATR算法的实际效果，将 $p_{(2)} = x_2/n$ 的取值范围限定在0.65—0.95之间，并用差异 $\delta = p_{(2)} - p_{(1)}$ 来标定 $p_{(1)}$ 的取值.

4.1.1 测试样本容量一定 取典型测试样本容量 $n = 100, 500, 1000, 3000$，令 $x_{(1)} = p_{(1)} \times n$，$x_{(2)} = p_{(2)} \times n$，$a_{(1)} = b_{(1)} = a_{(2)} = b_{(2)} = 1$，用(12)式计算 $Pb_{(2)}$，具体计算结果见图2.

从图2不难看出：

(1) 当两个算法识别率测试值的差异 $\delta > 0.10$ 后，即使样本容量 n 仅为100，选择 $p_{(2)}$ 作为最优算法的置信度 $Pb_{(2)}$ 也将大于0.9；而当两个算法识

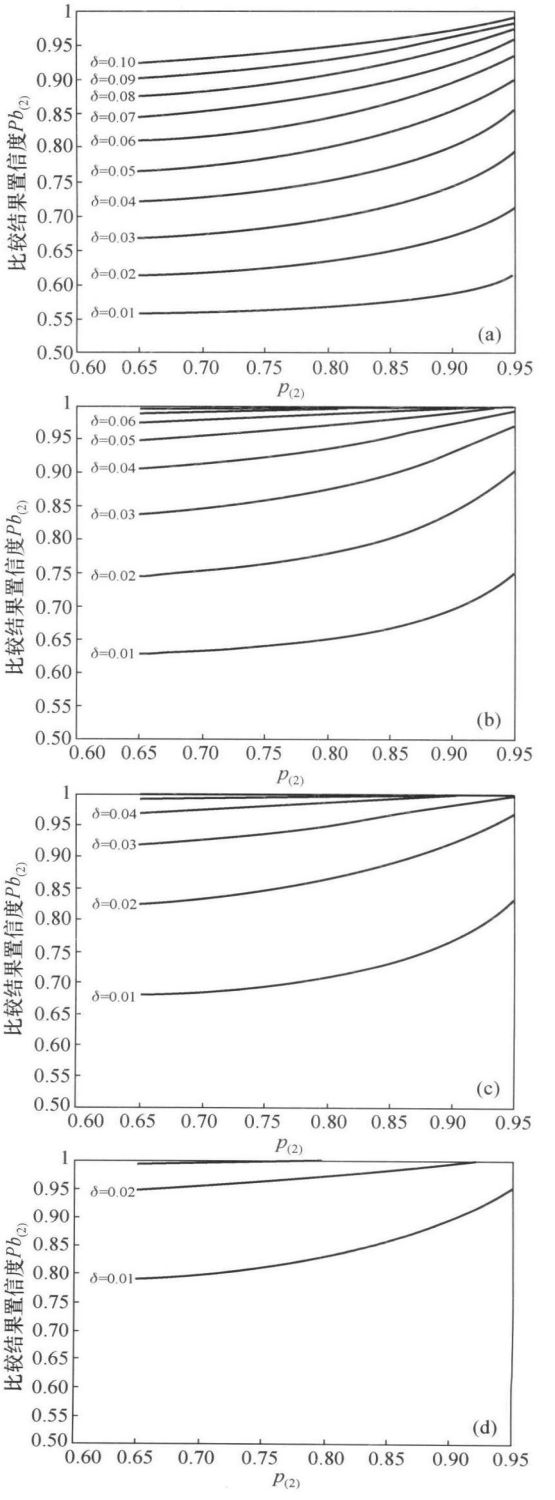

图2 样本容量一定时的比较结果置信度($m=2$)
(a) 样本容量 $n=100$；(b) 样本容量 $n=500$；(c) 样本容量 $n=1000$；(d) 样本容量 $n=3000$

别率测试值的差异 $\delta \leqslant 0.01$ 时，即使样本容量 n 高达3000，$p_{(2)}$ 也低于0.8．这说明 δ 对 $Pb_{(2)}$ 的影响很大，当 δ 比较明显时（>0.05），即使中等规模的测试样本容量（>500）也能使得比较结果的非常可信（置信度>0.95）．

（2）相同测试样本容量条件下，$p_{(1)}$ 和 $p_{(2)}$ 自身的取值越大，比较结果的置信度也越大，在 δ 较小时更是如此．关于这一点，下面的分析中体现的更加明显．

4.1.2 结论置信度要求一定　保持 $x_{(1)}=p_{(1)}\times n$，$x_{(2)}=p_{(2)}\times n$，$a_{(1)}=b_{(1)}=a_{(2)}=b_{(2)}=1$ 并用(12)式计算 $Pb_{(2)}$，设定典型置信度要求 $Pb_{(2)}=0.90$，0.95，用对分法搜索 n 值使得 $Pb_{(2)}$ 等于置信度要求，计算结果见图3.

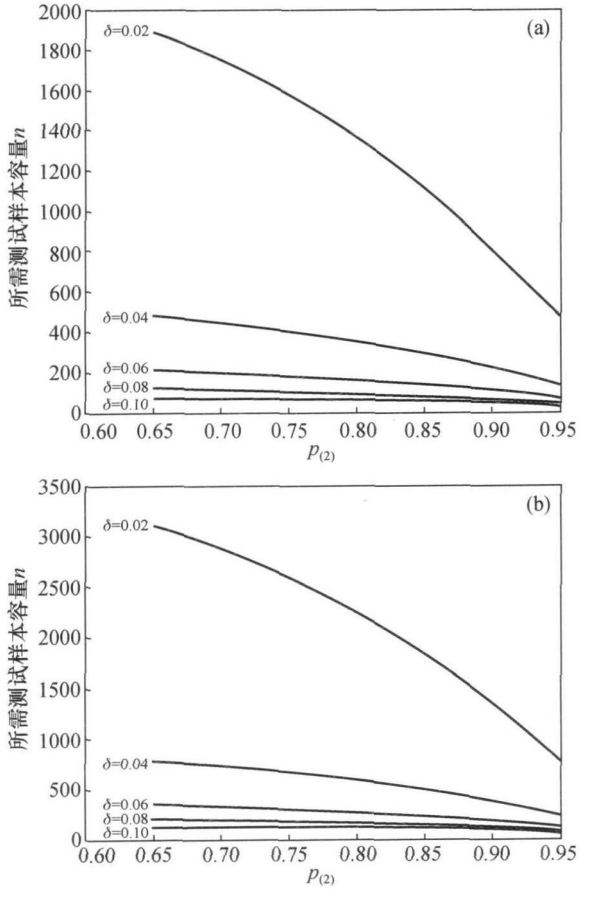

图3　比较结果置信度一定的样本容量（$m=2$）
（a）比较结果置信度 $Pb_{(2)}=0.90$；（b）比较结果置信度 $Pb_{(2)}=0.95$

图3给出了比较结果置信度一定时，各种具体的 $p_{(1)}$ 和 $p_{(2)}$ 情况下所需要的最小测试样本容量．从图3不难看出：（1）相同置信度要求下，δ 在很大程度上决定了所需的 n．δ 较大时，即使要求 $Pb_{(2)}$ 较高，所需的测试样本容量也不大．例如，$\delta>0.06$，比较结果置信度要求为0.95，所需的样本容量不到500；（2）δ 较小时，对 n 的需求与 $p_{(1)}$ 和 $p_{(2)}$ 自身取值密切相关．图3表明：$p_{(1)}$ 和 $p_{(2)}$ 的取值越大，所需测试样本容量 n 的越小．

不难对两个ATR算法的识别率比较做出如下结论：使用几百个左右的“中样本”测试集来比较ATR算法，识别率测试值5个百分点以上的差异才能说明比较结果具有较高的可信度；而使用具有几千个样本的“大样本”测试集来比较ATR算法，识别率测试值2—3个百分点左右的差异就能使得比较结果较为可信度．

4.2　多个算法的比较

参照前面对两个ATR算法识别率比较问题的分析结果，下面的分析中将多个算法识别率之间差异的取值限定在0.02—0.10之间，结合大多数模式识别背景中ATR算法的实际效果，定量分析计算了 $m=3$，4时部分典型情况下的比较结果置信度（样本容量一定）和所需测试样本容量（结果置信度一定），详细的计算结果见附录2的表1—表4.

通过对这些定量计算结果的分析发现：在ATR算法选优问题中，最优和次优测试值的差值 $\delta^*=p_{(m)}-p_{(m-1)}$ 主要决定了 $Pb_{(m)}$ 和所需样本容量 n；而在ATR算法排序问题中，m 个 $p_{(i)}$ 之间的最小差异 $\delta_{\min}=\min\{p_{(i)}-p_{(j)}\}$（$i\neq j$，$i$，$j=1$，2，…，$m$）在很大程度上决定了 $P_{(m)(m-1)\cdots(2)(1)}$ 和所需样本容量 n.

5　结束语

本文针对模式识别领域中所关注的一个共性问题——ATR算法识别率的比较，在归纳分析现有的各种算法识别率比较方法局限性的基础上，基于Bayes分析的理论框架提出了一种新的识别率比较方法——后验概率比较法，实现了所期望的功能和特性．结合ATR算法识别率比较过程中所特别关注的两个典型问题——以识别率为评价指标的算法

选优和排序，运用后验概率法分析并以定理形式证明了应用最大似然原理的合理性. 还根据目前大多数模式识别背景中ATR算法的实际效果，定量分析了采用后验概率法所需的测试样本容量与比较结果的置信度之间的约束关系，所得图表能够有效指导评估试验设计和数据采集工作. 所取得的这些研究成果对多个应用领域中模式识别算法的比较都具有参考价值和指导意义.

附录 1

对本文3.1节中定理1的证明:

(1) 任取样本空间 S 中的点 $X=(p_1, p_2, \cdots, p_m)$，由 H_0 和 H_1, H_2, …, H_m 的定义易知，$\exists k$ 使得 $X\in H_k$，则由 X 的任意性得 $S\subset H_0\cup H_1\cup H_2\cup\cdots\cup H_m$. S 为样本空间，故

$$S=H_0\cup H_1\cup H_2\cup\cdots\cup H_m.$$

若 $X\in H_k(0\leqslant k\leqslant m)$，由 H_0 和 H_1, H_2, …, H_m 的定义得 $X\notin H_u(u\neq k)$. 由 k 的任意性，得

$$H_i\cap H_j=\phi, i\neq j, i, j=0, 1, 2, \cdots, m.$$

根据概率划分的定义，命题 H_0, H_1, H_2, …, H_m 是样本空间 S 的划分，所以 $P\{H_0\}+\sum_{i=1}^{m}P\{H_i\}=1$. 由于 $H_{(1)}$, $H_{(2)}$, …, $H_{(m)}$ 仅是 H_1, H_2, …, H_m 的重新排序，H_0, $H_{(1)}$, $H_{(2)}$, …, $H_{(m)}$ 也是 S 的划分，故也有 $P\{H_0\}+\sum_{i=1}^{m}P\{H_{(i)}\}=1$. 注意到 $f(p_1, p_2, \cdots, p_m)$ 在 S 上连续，因而有 $P\{H_0\}=0$. 所以，$\sum_{i=1}^{m}P\{H_i\}=1$ 和 $\sum_{i=1}^{m}P\{H_{(i)}\}=1$ 成立.

(2) 由 $f(p_1, p_2, \cdots, p_m)=\prod_i \text{betapdf}(p_i; x_i+a_i, n-x_i+b_i)$ 可知，$\forall i(i=1, 2, \cdots, m)$，$p_i$ 的边缘分布 $f_i=\text{betapdf}(p_i; x_i+a_i, n-x_i+b_i)$. 由β分布的性质，$f_i$ 在(0, 1)上为单峰连续函数，其极大值点存在且唯一. 所以，$f(p_1, p_2, \cdots, p_m)$ 在 S 上的极大值点 Q 存在且唯一. 令

$$\frac{\partial f(p_1, p_2, \cdots, p_m)}{\partial p_1\partial p_2\cdots\partial p_m}=\prod_i\frac{1}{\text{Beta}(a_i, b_i)}p_i^{x_i+a_i-2}(1-p)_i^{x_i+b_i-2}[(n+a_i+b_i-2)p_i-(x_i+a_i-1)]=0$$

解得

$$Q=\left(\frac{x_1+a_1-1}{n+a_1+b_1-2}, \frac{x_2+a_2-1}{n+a_2+b_2-2}, \cdots, \frac{x_m+a_m-1}{n+a_m+b_m-2}\right).$$

若 $\frac{x_{(m-1)}+a_{(m-1)}-1}{n+a_{(m-1)}+b_{(m-1)}-2}<\frac{x_{(m)}+a_{(m)}-1}{n+a_{(m)}+b_{(m)}-2}$，那么 $\forall j\neq(m)$，有 $\frac{x_{(m)}+a_{(m)}-1}{n+a_{(m)}+b_{(m)}-2}>\frac{x_j+a_j-1}{n+a_j+b_j-2}$，所以

$$Q\in H_{(m)}=\{p_{(m)}>p_j\,|\,j\neq(m), j=1, 2, \cdots, m)\}.$$

边缘分布 $f_i=\text{betapdf}(p_i; x_i+a_i, n-x_i+b_i)$ 的方差为

$$Var[f_i]=\frac{(x_i+a_i)(n-x_i+b_i)}{(n+a_i+b_i)^2(n+a_i+b_i+1)}$$

显然 $\lim_{n\to\infty}Var[f_i]=\infty$，而由前面的分析知 f_i 的单峰值在 $q_i=(x_i+a_i-1)/(n_i+a_i+b_i-2)$ 处取得，所以 $\forall\Delta p>0$，$\lim_{n\to\infty}P\{q_i-\Delta p\leqslant P_i\leqslant q_i+\Delta p\}=1$. 因此对任意一个包含 $Q=(q_1, q_2, \cdots, q_m)$ 的事件 H_Q，只要其“体积”非空 $\left(\iint_{H_Q}\cdots\int\mathrm{d}p_1\mathrm{d}p_2\cdots\mathrm{d}p_m>0\right)$，就有 $\lim_{n\to\infty}P\{H_Q\}=1$. 由于已证 $Q\in H_{(m)}$，当

$$\frac{x_{(m-1)}+a_{(m-1)}-1}{n+a_{(m-1)}+b_{(m-1)}-2}<\frac{x_{(m)}+a_{(m)}-1}{n+a_{(m)}+b_{(m)}-2}$$

时 $\iint_{H_{(m)}}\cdots\int\mathrm{d}p_1\mathrm{d}p_2\cdots\mathrm{d}p_m>0$，故有 $\lim_{n\to\infty}P\{H_{(m)}\}=1$.

证毕

附录 2

表 1 样本容量一定时的选优和排序结果置信度($m=3$)

$p_{(3)}$	$p_{(2)}$	$p_{(1)}$	选优结果 $Pb_{(3)}$ 置信度					排序结果 $Pb_{(3)(2)(1)}$ 置信度				
			$n=100$	300	500	1000	3000	$n=100$	300	500	1000	3000
0.7	0.68	0.66	0.52	0.64	0.72	0.82	0.95	0.32	0.45	0.54	0.68	0.91
		0.64	0.56	0.68	0.75	0.83	0.95	0.41	0.58	0.68	0.81	0.95
		0.62	0.59	0.70	0.75	0.83	0.95	0.48	0.65	0.73	0.83	0.95
		0.60	0.60	0.70	0.75	0.83	0.95	0.53	0.69	0.75	0.83	0.95
	0.66	0.64	0.64	0.82	0.90	0.97	1.00	0.40	0.57	0.67	0.80	0.95
		0.62	0.68	0.84	0.91	0.97	1.00	0.49	0.71	0.82	0.94	1.00
		0.60	0.70	0.85	0.91	0.97	1.00	0.57	0.80	0.89	0.97	1.00
	0.64	0.62	0.75	0.93	0.98	1.00	1.00	0.46	0.64	0.73	0.82	0.95
		0.60	0.78	0.94	0.98	1.00	1.00	0.56	0.79	0.88	0.97	1.00
	0.62	0.60	0.84	0.98	1.00	1.00	1.00	0.52	0.68	0.74	0.82	0.94
0.80	0.78	0.76	0.54	0.68	0.76	0.86	0.97	0.34	0.49	0.58	0.73	0.94
		0.74	0.58	0.71	0.78	0.86	0.97	0.43	0.62	0.72	0.85	0.97
		0.72	0.61	0.72	0.78	0.86	0.97	0.51	0.69	0.77	0.86	0.97
		0.70	0.62	0.73	0.78	0.86	0.97	0.56	0.72	0.78	0.86	0.97
	0.76	0.74	0.68	0.86	0.93	0.98	1.00	0.43	0.61	0.71	0.84	0.96
		0.72	0.71	0.88	0.94	0.98	1.00	0.53	0.76	0.87	0.96	1.00
		0.70	0.73	0.88	0.94	0.98	1.00	0.61	0.84	0.92	0.98	1.00
	0.74	0.72	0.79	0.95	0.99	1.00	1.00	0.49	0.67	0.75	0.84	0.96
		0.70	0.82	0.96	0.99	1.00	1.00	0.60	0.83	0.91	0.98	1.00
	0.72	0.70	0.87	0.99	1.00	1.00	1.00	0.54	0.70	0.76	0.84	0.96
0.90	0.88	0.86	0.60	0.75	0.83	0.92	0.99	0.39	0.58	0.69	0.84	0.98
		0.84	0.64	0.78	0.84	0.92	0.99	0.50	0.72	0.81	0.92	0.99
		0.82	0.66	0.78	0.84	0.92	0.99	0.58	0.77	0.84	0.92	0.99
		0.80	0.67	0.78	0.84	0.92	0.99	0.62	0.78	0.84	0.92	0.99
	0.86	0.84	0.75	0.92	0.97	1.00	1.00	0.49	0.70	0.79	0.89	0.98
		0.82	0.78	0.93	0.97	1.00	1.00	0.61	0.85	0.93	0.99	1.00
		0.80	0.80	0.93	0.97	1.00	1.00	0.69	0.91	0.97	1.00	1.00
	0.84	0.82	0.86	0.98	1.00	1.00	1.00	0.56	0.73	0.80	0.88	0.98
		0.80	0.88	0.99	1.00	1.00	1.00	0.67	0.88	0.95	0.99	1.00
	0.82	0.80	0.93	1.00	1.00	1.00	1.00	0.59	0.73	0.79	0.87	0.98

表 2 样本容量一定时的选优和排序结果置信度($m=4$)

$p_{(4)}$	$p_{(3)}$	$p_{(2)}$	$p_{(1)}$	选优结果 $Pb_{(4)}$ 置信度					排序结果 $Pb_{(4)(3)(2)(1)}$ 置信度				
				$n=100$	300	500	1000	3000	$n=100$	300	500	1000	3000
0.70	0.68	0.66	0.64	0.48	0.63	0.71	0.82	0.95	0.15	0.28	0.38	0.55	0.86
			0.62	0.50	0.64	0.72	0.82	0.95	0.20	0.37	0.48	0.66	0.90
			0.60	0.51	0.64	0.72	0.82	0.95	0.24	0.42	0.52	0.68	0.91
		0.64	0.62	0.54	0.68	0.75	0.83	0.95	0.21	0.39	0.50	0.66	0.90
			0.60	0.55	0.68	0.75	0.83	0.95	0.27	0.48	0.61	0.78	0.95
	0.66	0.64	0.62	0.61	0.82	0.90	0.97	1.00	0.19	0.36	0.47	0.65	0.90
			0.60	0.63	0.82	0.90	0.97	1.00	0.25	0.47	0.60	0.77	0.95
		0.62	0.60	0.66	0.84	0.91	0.97	1.00	0.26	0.47	0.60	0.77	0.94
	0.64	0.62	0.60	0.73	0.93	0.98	1.00	1.00	0.23	0.41	0.51	0.66	0.89

续表

$p_{(4)}$	$p_{(3)}$	$p_{(2)}$	$p_{(1)}$	选优结果 $Pb_{(4)}$ 置信度					排序结果 $Pb_{(4)(3)(2)(1)}$ 置信度				
				$n=100$	300	500	1000	3000	$n=100$	300	500	1000	3000
0.80	0.78	0.76	0.74	0.51	0.67	0.75	0.86	0.97	0.17	0.32	0.43	0.62	0.90
			0.72	0.53	0.68	0.75	0.86	0.97	0.23	0.41	0.54	0.72	0.94
			0.70	0.54	0.68	0.76	0.86	0.97	0.27	0.46	0.57	0.73	0.94
		0.74	0.72	0.57	0.71	0.78	0.86	0.97	0.24	0.43	0.55	0.71	0.93
			0.70	0.58	0.71	0.78	0.86	0.97	0.30	0.53	0.67	0.83	0.97
	0.76	0.74	0.72	0.65	0.86	0.93	0.98	1.00	0.22	0.40	0.52	0.70	0.92
			0.70	0.67	0.86	0.93	0.98	1.00	0.28	0.52	0.65	0.82	0.96
		0.72	0.70	0.70	0.88	0.94	0.98	1.00	0.29	0.52	0.65	0.81	0.96
	0.74	0.72	0.70	0.77	0.95	0.99	1.00	1.00	0.25	0.44	0.54	0.70	0.92
0.90	0.88	0.86	0.84	0.58	0.75	0.83	0.92	0.99	0.22	0.42	0.55	0.75	0.97
			0.82	0.59	0.75	0.83	0.92	0.99	0.29	0.52	0.66	0.83	0.98
			0.80	0.59	0.75	0.83	0.92	0.99	0.33	0.56	0.68	0.84	0.98
		0.84	0.82	0.63	0.78	0.84	0.92	0.99	0.30	0.52	0.65	0.81	0.97
			0.80	0.64	0.78	0.84	0.92	0.99	0.37	0.64	0.77	0.91	0.99
	0.86	0.84	0.82	0.74	0.92	0.97	1.00	1.00	0.27	0.49	0.62	0.78	0.97
			0.80	0.75	0.92	0.97	1.00	1.00	0.35	0.62	0.75	0.88	0.98
		0.82	0.80	0.78	0.93	0.97	1.00	1.00	0.35	0.61	0.73	0.86	0.98
	0.84	0.82	0.80	0.85	0.98	1.00	1.00	1.00	0.30	0.51	0.61	0.77	0.96

表 3 比较结果置信度一定时所需测试样本容量($m=3$)

$p_{(3)}$	$p_{(2)}$	$p_{(1)}$	选优结果置信度		排序结果置信度	
			$Pb_{(3)}=0.90$	0.95	$Pb_{(3)(2)(1)}=0.90$	0.95
0.7	0.68	0.66	1792.55	2902.88	2902.34	4156.28
		0.64	1758.00	2894.66	1834.69	2915.30
		0.62	1757.76	2894.66	1759.11	2894.74
		0.60	1757.76	2894.66	1757.78	2894.65
	0.66	0.64	500.90	771.00	1929.12	3091.50
		0.62	457.16	739.20	748.00	1070.91
		0.60	448.95	737.04	535.45	795.90
	0.64	0.62	242.61	365.72	1915.87	3154.27
		0.60	217.08	340.30	557.42	839.85
	0.62	0.60	145.18	216.76	1954.63	3219.21
0.80	0.78	0.76	1392.65	2252.74	2289.08	3278.39
		0.74	1364.43	2245.86	1433.42	2266.36
		0.72	1364.19	2245.86	1365.80	2245.87
		0.70	1364.19	2245.86	1364.20	2245.83
	0.76	0.74	396.85	609.81	1583.07	2545.90
		0.72	362.02	583.90	608.25	871.04
		0.70	355.04	581.93	432.59	638.40
	0.74	0.72	195.56	294.27	1620.30	2667.54
		0.70	175.26	273.73	468.98	711.59
	0.72	0.70	118.92	177.17	1692.14	2786.51

续表

$p_{(3)}$	$p_{(2)}$	$p_{(1)}$	选优结果置信度		排序结果置信度	
			$Pb_{(3)}=0.90$	0.95	$Pb_{(3)(2)(1)}=0.90$	0.95
0.90	0.88	0.86	826.68	1332.48	1409.80	2020.11
		0.84	807.89	1327.43	864.20	1348.21
		0.82	807.65	1327.43	809.95	1327.54
		0.80	807.65	1327.43	807.68	1327.44
	0.86	0.84	247.90	378.74	1068.62	1729.83
		0.82	226.08	361.67	402.34	576.96
		0.80	221.08	360.02	282.57	410.88
	0.84	0.82	127.47	190.49	1161.09	1910.72
		0.80	114.75	177.23	333.38	513.22
	0.82	0.80	80.47	118.88	1265.86	2083.68

表4 比较结果置信度一定时所需测试样本容量($m=4$)

$p_{(4)}$	$p_{(3)}$	$p_{(2)}$	$p_{(1)}$	选优结果置信度		排序结果置信度	
				$Pb_{(4)}=0.90$	0.95	$Pb_{(4)(3)(2)(1)}=0.90$	0.95
0.70	0.68	0.66	0.64	1792.71	2902.88	3653.93	4970.61
			0.62	1792.55	2902.88	2914.08	4158.83
			0.60	1792.55	2902.88	2902.36	4156.26
		0.64	0.62	1758.00	2894.66	2987.08	4273.43
			0.60	1758.00	2894.66	1907.94	2940.48
	0.66	0.64	0.62	505.69	772.43	3079.46	4401.72
			0.60	501.26	771.04	1995.59	3111.17
		0.62	0.60	457.95	739.27	2063.58	3243.56
	0.64	0.62	0.60	249.86	369.66	3142.33	4499.92
0.80	0.78	0.76	0.74	1392.81	2252.74	2927.07	3983.62
			0.72	1392.65	2252.74	2301.58	3281.44
			0.70	1392.65	2252.74	2289.12	3278.39
		0.74	0.72	1364.43	2245.86	2425.88	3475.37
			0.70	1364.43	2245.86	1508.86	2297.66
	0.76	0.74	0.72	400.99	611.11	2570.13	3675.87
			0.70	397.21	609.85	1646.59	2565.94
		0.72	0.70	362.79	583.97	1768.02	2801.34
	0.74	0.72	0.70	201.71	297.72	2689.00	3850.92
0.90	0.88	0.86	0.84	826.86	1332.51	1871.18	2552.18
			0.82	826.70	1332.48	1423.88	2024.54
			0.80	826.68	1332.48	1409.89	2020.11
		0.84	0.82	807.89	1327.43	1597.54	2303.77
			0.80	807.89	1327.43	940.07	1392.45
	0.86	0.84	0.82	250.97	379.83	1794.46	2568.96
			0.80	248.24	378.79	1125.27	1749.91
		0.82	0.80	226.79	361.76	1302.44	2089.38
	0.84	0.82	0.80	131.86	193.12	1969.24	2820.56

参 考 文 献

1 Guyon I, Makhoul J, Schwartz R, et al. What size test set gives good error rate estimates. IEEE Trans on Pattern Analysis and Machine Intelligence, 1998, 20(1): 52—64

2 Ross T D. Confidence intervals for ATR performance metrics. In: Zelnio EG, eds. Algorithms for Synthetic Aperture Radar Imagery VIII. Orlando, USA, 2001. Bellingham: Society of Photo-Optical Instrumentation Engineers, 2001, Proc SPIE. 4382: 318—329

3 Alsing SG. Evaluation of competing classifiers. Doctor thesis, Air Force Inst of Tech, Wright-Patterson AFB, OH, School of

Engineering, 2000
4 Bassham CB. Automatic target recognition classification system evaluation methodology. Doctor thesis, Air Force Inst of Tech, Wright-Patterson AFB, OH, School of Engineering and Management, 2002
5 Moore RE. Method and Application of Interval Analysis. London: Prentice-Hall, 1979, 9—13
6 吴 江, 黄登仕. 区间数排序方法研究综述. 系统工程, 2004, 22(8): 1—4
7 徐 扬. 不确定性推理. 成都: 西南交通大学出版社, 1994, 16—20
8 Joseph L, Wolfson DB, Berger R. Sample size calculation for binomial proportions via highest posterior density intervals. The Statistician, 1995, 44(2): 143—154
9 He J, Zhao HZ, Fu Q. Sample size analysis for confidence interval estimation of performance metrics in ATR evaluation. In: IEEE 2007 Radar Conference. Waltham, USA, 2007. Piscataway: Inst of Electrical and Electronics Engineers Inc, 2007, 585—589
10 Catlin AE, Bauer Jr KW, Mykytka EF, et al. System comparison procedures for automatic target recognition system. Naval Research Logistics, 1999, 46: 357—371
11 Gibbons JD, Olkin I, Sobel M. Selecting and Ordering Populations: A New Statistical Methodology. New York: Wiley, 1977, 105
12 Berger JO 著. 统计决策论及贝叶斯分析. 贾乃光译. 北京: 中国统计出版社, 1998, 482, 560—561
13 Wald A. Sequential Analysis. New York: John Wiley & Sons, 1947, 109—116
14 Shaffer JP. Modified sequentially rejective multiple test procedures. Journal of the American Statistical Association, 1986, 81(395): 826—831