

# ROC 分析的基本原理

宇传华 徐勇勇

ROC 是受试者工作特征 (Receiver Operating Characteristic) 或相对工作特征 (Relative Operating Characteristic) 的缩写。ROC 分析五十年代起源于统计决策理论, 后来应用于雷达信号观察能力的评价, 六十年代中期大量成功地用于实验心理学和心理物理学研究<sup>[1]</sup>。Lusted 首次提出了 ROC 分析可用于医学决策评价。自从八十年代起, 该方法广泛用于医学诊断性能的评价<sup>[2]</sup>, 如用于诊断放射学、实验室医学、癌症的筛选和精神病的诊断, 尤其是医学影像诊断准确性的评价。

## 1 其它准确性评价指标的局限性

对于二分类总体, 如对照与病例 (或无病与有病, 正常与异常, 噪声与信号), 诊断试验结果分别写成阴性和阳性, 其资料可列成表 1 的四格表形式<sup>[3]</sup>, 这时可计算出正确百分率、灵敏度、特异度等指标 (计算公式见表 1), 这几个指标均可不同程度反映诊断的准确性。正确百分率是病例正确诊断为阳性与对照正确诊断为阴性的例数和占总例数的百分比。它很大程度上依赖患病率, 如患病率为 5%, 完全无价值地诊断所有样本为阴性, 也可有 95% 的正确百分率; 其次, 它没有揭示假阴性和假阳性错误诊断的频率, 相同的正确百分率可能有十分不同的假阴性和假阳性。第三, 它也受诊断阈值的限制。更好的方法是计算灵敏度和特异度, 它们的值越高, 诊断性能越好。灵敏度是病例被正确诊断为阳性的比例, 也叫真阳性率 (true positive fraction/rate, 简称 TPF)。特异度是对照被正确诊断为阴性的比例, 也叫真阴性率, (1-特异度) 为假阳性率 (false positive fraction/rate, 简称 FPF)。应用这对指标最明显的问题是: 比较两个诊断系统时, 可能出现一个诊断系统的灵敏度高, 而另一个特异度高, 无法判断哪一个诊断系统更好。此时可将灵敏度和特异度结合, 改变诊断阈值, 获得多对灵敏度和 (1-特异度), 即 TPF 和 FPF 值, 绘制 ROC 曲线, 作 ROC 分析来解决这一问题。

表 1 诊断资料 2×2 四格表

诊 断 结 果	“金标准”		合 计
	病例	对照	
阳 性	a(真阳性)	b(假阳性)	a+b
阴 性	c(假阴性)	d(真阴性)	c+d
合计	a+c	b+d	a+b+c+d=N

$$\text{正确百分率} = \frac{a + d}{N} \times 100\%$$

$$\text{灵敏度} = \text{真阳性率 (TPF)} = \frac{a}{a + c}$$

$$\text{特异度} = \frac{d}{b + d} = [1 - \text{假阳性率 (FPF)}] = 1 - \frac{b}{b + d}$$

	疾病 (+) Disease(+)	疾病 (-) Disease(-)	合计 Total
试验(+) Test(+)	真阳性 True positives	假阳性 False positives	试验阳性数 Test positives
试验(-) Test(-)	假阴性 False negatives	真阴性 True negatives	试验阴性数 Test negatives
合计 Total	实际病人数 Disease positives	实际非病人数 Disease negatives	所有受试者 All subjects

## 2 ROC 分析资料收集与整理

一个诊断系统获得的原始资料可记录成连续性和离散性两种形式。

连续性资料常见于临床检验，它是利用仪器、设备等测量的数据。为了叙述的方便，假设某诊断系统的对照组有 4 个受试者，其检验结果为 8.29, 6.24, 4.61, 1.77，病例组有 5 个受试者，其检验结果为 15.90, 13.35, 12.87, 10.22, 5.01。将这 9 个数据从大到小排列，以前 8 个数据的每一个作为诊断阈值(或截断点)，大于等于该阈值为阳性，小于该阈值为阴性，则对于每一个诊断阈值，都可整理成类似表 1 的四格表。由此可得出以下 8 个 ROC 工作点 (FPF,TPF) =(0/4,1/5), (0/4,2/5), (0/4,3/5), (0/4,4/5), (1/4,4/5), (2/4,4/5), (2/4,5/5), (3/4,5/5)。实际工作中，两个组的样本量往往很大，可将资料整理成频数表，以组中值为诊断阈值，获得频数表的 (组数-1) 个 ROC 工作点。也可根据临床需要取若干个诊断阈值。

离散性资料常见于医学影像诊断和心理学评价。常将诊断结果划分为 5 或 6 类。如果分 5 类，可将受试对象按肯定正常、可能正常、异常可疑、可能异常、肯定异常进行分类，分别记为 1、2、3、4、5。例如：有 109 份 CT 影像，其中有 58 份正常，51 份异常，一影像工作者对这些 CT 影像进行分类，其结果见表 2。

表 2 109 份 CT 影像分类

实 际 情 况	诊 断 分 类					合 计
	1	2	3	4	5 *	
正 常	33	6	6	11	2	58
异 常	3	2	2	11	33	51

\*: 1,2,...,5 分别表示肯定正常,可能正常,...,肯定异常

对于 5 分类资料，可从分类高到低以前 4 个分类 (即分类 5,4,3,2) 为置信阈计算 ROC 工作点。该类及以上例数的和为阳性，该类以下例数的和为阴性，对每一个置信阈都可整理出一个类似表 1 的四格表。如以表 2 分类 4 为置信阈时，则正常组阳性例数为 2+11=13，其假阳性率 FPF=13/58=0.2241；异常组阳性例数为 33+11=44，其真阳性率 TPF=44/51=0.8627。同样，以表 2 分类 5、3、2 为置信阈计算的 ROC 工作点分别为 (FPF,TPF) =(0.0345,0.6471)、(0.3296,0.9020)、(0.4310,0.942)。

假设正常组和异常组均服从正态分布，则 5 分类等级资料可按图 1 解释。图中左边钟型曲线为正常组，右边为异常组，同时采用 4 个置信阈将实验结果分为 5 类，置信阈越严格，即分类越高，阳性决策的置信度越高。置信阈越宽松，即分类越低，阳性决策的置信度越低。连续性资料是采用诊断阈值进行判断的，也可类似理解。

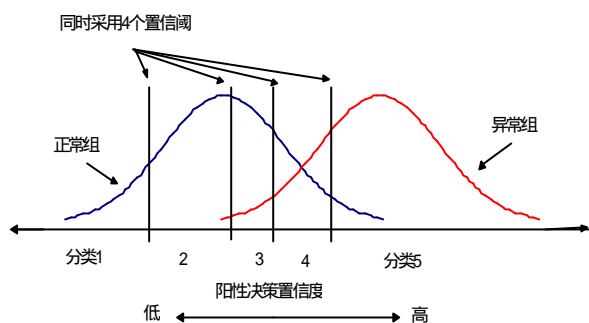


图 1 5 分类等级资料图解

### 3 ROC 曲线的构建

以假阳性率 (FPF) 为横轴，真阳性率 (TPF) 为纵轴，横轴与纵轴长度相等，形成正方形，在图中将 ROC 工作点标出，用直线连接各相邻两点构建未光滑 ROC 曲线。构建光滑曲线需要假设对照组和病例组服从某种分布，用一曲线拟合技术估计参数，直接用参数产生。也可采用肉眼目测构建光滑曲线。注意：无论资料类型如何，曲线一定通过 (0,0) 和 (1,1) 这两点，这两点分别相当于灵敏度为 0 而特异度为 1 和灵敏度为 1 而特异度为 0。理论上，完善的诊断有  $TPF=1$ ， $FPF=0$ ，图中表现为 ROC 曲线从原点垂直上升到图的左上角，然后水平到达右上角；完全无价值的诊断有  $TPF=FPF$ ，是一条从原点到右上角的对角线；一般 ROC 曲线位于正方形的上三角。表 2 资料对应的 ROC 曲线见图 2。

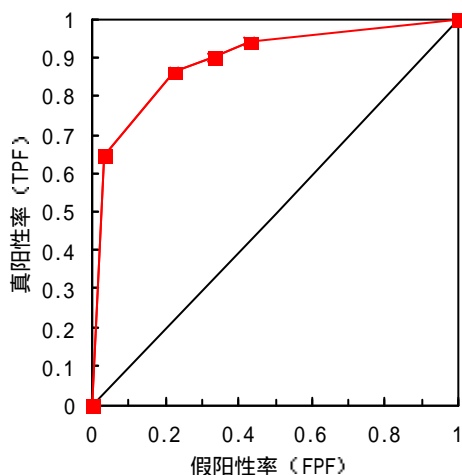


图 2 未光滑的 ROC 曲线示意图

ROC 曲线对诊断的准确性提供了直观的视觉印象，描述了相反两种状态间诊断系统的判别能力。曲线上的每一点代表了随着病例诊断阈值（或置信阈）变化的灵敏度与特异度的折衷。严格的标准产生较低的灵敏度和较高的特异度，ROC 点位于曲线的左下方；宽松的标准产生较高的灵敏度和较低的特异度，ROC 点位于曲线的右上方。如果比较的诊断系统曲线不交叉，则较高的 ROC 曲线具有较好的诊断性能。如果曲线交叉，可在考虑费用与效益的前提下，结合统计学检验判断其诊断性能。

### 4 ROC 分析的准确性评价指标

一般用“ROC 曲线下面积”反映诊断系统的准确性。理论上，这一指标取值范围为 0.5 至 1，完全无价值的诊断为 0.5；完善的诊断为 1。该指标及其标准误的计算目前有非参数、

半参数和参数方法。其中得到广泛应用的方法有：Wilcoxon 非参数法和最大似然估计“双正态”参数法（具体方法参见其它文献）。

## 5 其它有关问题

### 5.1 ROC 工作点的可信区间

对于每个 ROC 工作点，可根据二项分布标准误计算公式  $S_p = \sqrt{\frac{p(1-p)}{n}}$  计算 FPF 和 TPF 的标准误。对于 FPF， $p=FPF$ ， $n=$ 对照组总例数；对于 TPF， $p=TPF$ ， $n=$ 病例组总例数。如表 2 资料分类 4 的工作点(FPF,TPF)=(0.2241,0.8627) 的标准误为(0.0548,0.0023)，其 95% 的可信区间为 $[(0.2241 \pm 1.96 \times 0.0548), (0.8627 \pm 1.96 \times 0.0023)]$ 。

### 5.2 “金标准”

对于诊断系统的准确性评价,首先应知道受试者(人、动物或影像等)的真实情况,即哪些是对照组,哪些是病例组。划分它们的标准就是“金标准”。常见的“金标准”有:跟踪随访、活组织检查、尸体解剖等。尽管“金标准”不需要十全十美,但是它们应比评价的诊断系统更可靠,且与评价的诊断系统无关。

### 5.3 最佳工作点的选择

阳性似然比为真阳性率与假阳性率之比,即等于(TPF/FPF);约登(Youden)指数为真阳性率与假阳性率之差,即等于(TPF-FPF)。一般选择阳性似然比<sup>[4]</sup>或约登指数<sup>[5]</sup>最大者为最佳工作点。

也有人采用  $C/B \times [(1-P)/P]$  计算最佳工作点的斜率,表达式中 C、B 和 P 分别表示花费、收益和患病率。在假定对病例组实施治疗,而对对照组不治疗的前提下,这一表达式表示治疗疾病的花费(Cost)和收益(Benefit)之比与(1-患病率)和患病率(Prevalence)之比的乘积。在 ROC 曲线上,从(FPF,TPF)=(0,0)到(1,1),工作点斜率从大到小单调改变。从表达式可以看出,如果疾病治疗花费多,收益少,或患病率低,则斜率大,最佳工作点接近(0,0),确保了假阳性的减少。如果疾病治疗花费少,收益多,或患病率高,则斜率小,最佳工作点接近(1,1),确保了假阴性的减少。

## 参考文献

- 1 Metz CE, Herman BA, Shen JH. Maximum-likelihood estimation of receiver operating characteristic(ROC) curves from continuously-distributed data. Stat Med,1998, in press.
- 2 Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 1989;24:234-245.
- 3 章扬熙. 临床流行病学(第六讲):诊断试验研究(续). 中华流行病学杂志 1997,18(3):177.
- 4 Feinstein AR. Clinical epidemiology: architecture of clinical research. WB Saunders Company, 1985,601-632.
- 5 刘杰,林一帆,张沥等. 图象自动分析检测 MG7 抗原表达预测胃癌高危价值探讨. 中华预防医学杂志 1996;30(5):286.
- 6 Dwyer AJ. In pursuit of a piece of the ROC. Radiology 1997;202:621-625.