Updated 2/3/11

The ROC task

Receiver operating characteristic (ROC) methodology is a widely used method for comparing the performances of two or more imaging modalities. Radiologists rate each patient image for probability of disease, e.g., 1 = definitely not diseased, …, 3 = equivocal, …, 6 = definitely diseased. The truth is known to the person running the study but not the radiologists. The most common study design is where all radiologists interpret all images in all modalities, the multiple-reader multiple-case (MRMC) study design. DBM-MRMC software is available on two websites that analyze MRMC-ROC data, and if the p-value is less than 5% one concludes that there is a statistically significant performance difference between at least two modalities. Generally the area under the ROC curve (AUC) is used as a measure of performance or figure of merit.

ROC is a binary paradigm: the patient either does or does not have disease, i.e., truth is binary. The radiologist's task is to state whether the patient does or does not have disease, i.e., the response is binary. The resulting 2 x 2 truth-response table defines good decisions (true positives, true negatives) and bad decisions (false positives and false negatives). The performance measure rewards good decisions and penalizes bad decisions.

ROC task definitions:

TP ≡ true positive ≡ diseased patient diagnosed as diseased
TN ≡ true negative ≡ non-diseased patient diagnosed as non-diseased
FP ≡ false positive ≡ non-diseased patient diagnosed as diseased
FN ≡ false negative ≡ non-diseased patient diagnosed as diseased

TPF = true positive fraction ≡ # TP patients divided by total number of diseased patients ($0 \le TPF \le 1$).
TNF = true negative fraction ≡ # TN patients divided by total number of non-diseased patients ($0 \le TNF \le 1$).
FPF = false positive fraction ≡ # FP patients divided by total number of non-diseased patients ($0 \le FPF \le 1$) ($0 \le FPF \le 1$).
FNF = false negative fraction ≡ # FN patients divided by total number of diseased patients ($0 \le FNF \le 1$).

Obviously TNF = 1 – FPF and FNF = 1 – TPF.

The ROC curve is the plot of TPF along the y-axis vs. FPF as the confidence level is varied. The ROC curve is contained within the unit square. AUC = area under the ROC curve and $0 \le AUC \le 1$. AUC is the probability that an abnormal image is rated higher than a normal image.

The free-response task

In some diagnostic examinations the truth is multi-focal, i.e., there could be one or more diseased regions (lesions) in the patient image and a location is associated with each focal lesion. While interpreting the image the radiologist may find one or more regions that are suspicious for disease. The truth-response table is more complex that the 2 x 2 ROC table. The number of lesions is known to the person running the study, but the number and locations of suspected lesions is a-priori unknown.

In the free-response paradigm the radiologist does not know a priori how many lesions may be present in an image, if any, and therefore must search the image for lesions and mark regions that are suspicious for disease. The basic unit of data is a mark-rating pair. The mark is the indicated location of the suspicious region and the rating is the radiologist's estimate of the probability of disease, or confidence level; e.g., 1 ≡ less that 2% probability of disease, 2 ≡ 3 to 5%, probability of disease, 3 ≡ 6 – 30% probability, 4 ≡ 31 to 70%, 5 ≡ 71 to 94% probability and 6 ≡ greater than 95% probability.

Free-response task definitions:

LL ≡ lesion localization, i.e., a lesion marked to within an agreed upon accuracy.
NL ≡ non-lesion localization, i.e., the mark is not close to any lesion.

LLF ≡ lesion localization fraction ≡ # LL divided by total number of lesions ($0 \le LLF \le 1$).
NLF ≡ non-lesion localization fraction ≡ # NL divided by total number of images ($0 \le NLF$); note the lack of an upper bound.

The FROC (free-response ROC) curve is the plot of LLF vs. NLF. It can extend indefinitely to the right, but the ordinate is limited to unity or less.

The rating of the highest rated mark on the image is often used to infer the ROC rating for that image. An inferred TP rating could be the rating of a LL or a NL, whichever happened to be higher, on an abnormal image (diseased patient). An inferred FP rating is always the rating of the highest rated NL on a normal image (non-diseased patient). Using inferred ROC quantities one can define TPF, FPF and an inferred ROC curve.

AFROC ≡ alternative FROC curve is the plot of LLF vs. FPF. Like the ROC it is contained within the unit square. A plausible figure of merit is the area θ under the AFROC curve. If all abnormal images had exactly one lesion θ, is the probability that a lesion is rated higher than a normal image ($0 \le \theta \le 1$).

JAFROC

JAFROC (jackknife AFROC) is a method for analyzing free-response MRMC data. The JAFROC figure of merit is the trapezoidal estimate of θ. It is defined by

$$\theta = \frac{1}{N_N N_L} \sum_{i=1}^{N_N} \sum_{j=1}^{N_L} \psi(X_i, Y_j)$$

$$\psi(X_i, Y_j) = \begin{bmatrix} 1.0 \; if \; Y_j > X_i \\ 0.5 \; if \; Y_j = X_i \\ 0.0 \; if \; Y_j < X_i \end{bmatrix}$$

Here N_N is the number of disease-free cases, N_L is the total number of lesions, X_i is the rating of the highest rated mark on the ith disease-free case and Y_j is the rating of the jth lesion. Unmarked disease-free cases and unmarked lesions are assigned the -2000 rating. The non-lesion localization marks on

diseased cases are not used. The expression in the 2004 Medical Physics paper which involves "weights" is not used.

Statistical power

JAFROC is more sensitive at detecting differences between the performances of modalities than ROC, i.e. it has higher statistical power. Statistical power is an important consideration in observer performance studies as it determines the probability of detecting a true difference between modalities while controlling the probability of detecting non-existent differences.

Comment

Users of JAFROC are sometimes surprised that a modality (for example, A) on which the observer marks some of the lesions without marking any normal image does not yield a perfect figure of merit ($\theta = 1$) and conversely a modality (for example, B) on which the observer marks some of the normal images and does not mark any of the lesions does not yield a zero figure of merit ($\theta = 0$). In fact it is observed that $0 < \theta_B < \theta_A < 1$. The observer who marks only lesions is obviously better than the observer who marks only normal images. If the observer marked every lesion and did not mark any normal image, the figure of merit would be unity and if the observer marked every normal image and did not mark any lesion, the figure of merit would be zero.