Nuclear Instruments and Methods in Physics Research A 648 (2011) S297–S301

Contents lists available at ScienceDirect

Nuclear Instruments and Methods in Physics Research A

journal homepage: www.elsevier.com/locate/nima

Recent developments in imaging system assessment methodology, FROC analysis and the search model

Dev P. Chakraborty

University of Pittsburgh, 2103 Noble Ct, Murrysville, PA 15668, USA

ARTICLE INFO

Available online 18 November 2010

Keywords:
Imaging system assessment
Observer performance
ROC
Free-response
FROC
JAFROC

ABSTRACT

A frequent problem in imaging is assessing whether a new imaging system is an improvement over an existing standard. Observer performance methods, in particular the receiver operating characteristic (ROC) paradigm, are widely used in this context. In ROC analysis lesion location information is not used and consequently scoring ambiguities can arise in tasks, such as nodule detection, involving finding localized lesions. This paper reviews progress in the free-response ROC (FROC) paradigm in which the observer marks and rates suspicious regions and the location information is used to determine whether lesions were correctly localized. Reviewed are FROC data analysis, a search model for simulating FROC data, predictions of the model and a method for estimating the parameters. The search model parameters are physically meaningful quantities that can guide system optimization.

1. Introduction

A frequent problem in imaging is assessing whether a new imaging system is an improvement over an existing standard [1]. The imaging system generally consists of several components, e.g., X-ray source, grid, X-ray detector, image processing algorithm, image display and the observer. Fourier measurements like modulation transfer function, signal to noise ratio, etc., are excellent tools for optimization of *parts* of the imaging chain, e.g., detector spatial resolution is optimized by measurements of modulation transfer function. However, the effect on performance of the entire imaging chain, including the observer, requires different methods that fall under the rubric of observer performance methods or "ROC analysis" [2–4]. The receiver operating characteristic (ROC) analysis is widely used in this context but it has limitations that have led to research on alternate paradigms [5–8]. This paper reviews progress in the free-response paradigm [5].

2. ROC

The receiver operating characteristic (ROC) curve is the plot of true positive fraction (TPF) vs. false positive fraction (FPF). A commonly used figure of merit is the area AUC under the ROC curve. AUC measures the ability of the observer/imaging system to correctly classify normal and abnormal images: AUC = 1 for perfect classification ability and 0.5 for chance level classification ability. The ROC curve is usually determined using the ratings method. The observer is shown an image, which could be normal (disease free) or abnormal (disease present), but the observer is "blinded" to this information. The observer reports a subjective confidence level that the image is abnormal. The confidence level is an ordinal variable, e.g., high confidence normal, low confidence normal, equally uncertain normal or abnormal, low confidence abnormal and high confidence abnormal, or the labels 1, 2, 3, 4 and 5 could be used to classify each image according to its confidence level. The ratings of a set of normal and abnormal images are used to calculate AUC [9], an objective measure of performance.

To compare two modalities one obtains AUC for each modality and the modality with the higher AUC is superior. Since the AUCs are subject to sampling variability, the result of the comparison is a p-value for rejecting the null hypothesis that the two modalities are identical. Let α denote the size of the test, i.e., the specified Type I error rate. If the p-value is sufficiently small, and typically one chooses $\alpha = 5\%$ as "small enough", then if $p < \alpha$, the modalities are declared different at the α—significance level. In a multiple-reader multiple-case (MRMC) study a set of observers interpret a common case set in both modalities. The reader and case matching ensure that differences in expertise levels of readers and difficulty levels of cases do not obscure the modality effect that one is interested in detecting. Dorfman–Berbaum–Metz (DBM) MRMC software [10–12] is commonly used to analyze MRMC ROC data.

In ROC data collection the reader assigns a single rating to each image. When the signs of the disease are diffuse then the ROC rating captures the relevant information. An example is interstitial lung disease which is characterized by scarring of lung tissue. When the disease is manifested by the presence of localized lesions, such as lung nodules, pointing to the correction location informs the experimenter that the reader has actually seen the disease. Moreover the location is relevant as it may guide subsequent interventions

E-mail addresses: dpc10@pitt.edu, dpc10ster@gmail.com

0168-9002/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.nima.2010.11.042

(e.g., biopsy). Not collecting location information would introduce ambiguity since the experimenter cannot rule out that the reader missed the lesion and mistook a suspicious normal region for a lesion. For such tasks the ROC rating would represent the answer to the ambiguous question "what is your confidence level that there is at least one nodule somewhere in the image".

3. FROC

In free-response ROC (FROC) data collection the observer reports the locations and confidence levels of regions that are suspicious for disease [5,13]. The unit of data is the *mark-rating pair* where the mark is the location and the rating is the confidence level that the reported region is actually a lesion. The experimenter decides whether a mark is close enough to a real lesion to qualify as a lesion localization (LL) and otherwise the mark is classified as a non-lesion localization (NL) [14]. The FROC curve is defined as the plot of lesion localization fraction (LLF) vs. non-lesion localization fraction (NLF), where the respective denominators are the total number of lesions and the total number of images [15]. Table 1 shows 6-rating FROC data, simulated by a model to be described later, for 50 normal images, and 50 abnormal images with 98 lesions. It illustrates the procedure for calculating the operating points. For example, cumulating the counts in bins 3, 4, 5 and 6 one obtains NLF=(20+5+13+8)/100=0.46 and LLF=(6+5+5+24)/98=0.408. Note that while the total number of potential LLs is known, namely 98, the total number of potential NLs is unknown. The number of true negatives – normal regions that were examined by the observer but correctly rejected as possible lesions – is unknown.

If one assumes that the rating of the highest-rated mark on an image is its ROC-equivalent rating, then one can infer ROC data from FROC data. If the image has no marks then its inferred rating is zero (or any number smaller than the smallest explicit rating, 1 in the present example). In Table 2 this has been done for the normal images and used to determine false positive (FP) counts and FPFs. The values 32 and 48 under Bin 0 are the number of unmarked normal images and the number of unmarked lesions, respectively. The AFROC curve is the plot of LLF vs. FPF and this table illustrates the calculation of AFROC operating points.

Table 3 shows inferred-ROC counts and operating points. On normal images the highest rating is necessarily that of a NL, or zero, if there is no mark, but on abnormal images, the highest rating could be a LL or an NL, whichever is rated higher, or zero, if there is no mark.

3.1. Analysis of MRMC FROC data

Analysis of observer performance data involves specification of a figure of merit quantifying performance and a method for assigning a significance value, or *p*-value, to the observed reader-averaged difference of figures of merit between two modalities. In DBM-MRMC analysis of ROC data one can use the area under the ROC curve as the figure of merit, estimate it using the proper ROC model [16], and the significance testing is performed by DBM analysis of variance [10–12]. In jackknife alternative FROC (JAFROC) analysis of FROC data the figure of merit is the area under the AFROC curve, currently estimated non-parametrically, and the significance testing is performed using DBM analysis of variance—the significant testing procedure is applicable to any scalar figure of merit. Software implementing the analysis is available at www.devchakraborty.com. Since it does not use location information one may suspect that ROC analysis is less precise than FROC and more prone to missing a true modality improvement, i.e., has less statistical power. For lack of statistical power a better algorithm design approach may be abandoned in favor of a suboptimal approach. In simulation studies JAFROC has been shown to have higher statistical power than ROC analysis [17–19].

Table 1
This table illustrates a hypothetical FROC data set and the corresponding FROC operating points. It corresponds to a 6-rating FROC study where 1=very low confidence in presence of lesion and 6=definite lesion.

	Total	Bins and counts					
		Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6
NL	Unknown	9	16	20	5	13	8
LL	98	5	5	6	5	5	24
FROC		Operating points					
		Bins ≥ 6	Bins ≥ 5	Bins ≥ 4	Bins ≥ 3	Bins ≥ 2	Bins ≥ 1
NLF		0.080	0.210	0.260	0.460	0.620	0.710
LLF		0.245	0.296	0.347	0.408	0.459	0.510

Table 2
This table illustrates the calculation of AFROC operating points. The zero bin represents unmarked normal images and unmarked lesions.

	Total	Bins and counts						
		Bin 0	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6
FP	50	32	0	4	8	1	3	2
LL	98	48	5	5	6	5	5	24
AFROC		Operating points						
		Bins ≥ 6	Bins ≥ 5	Bins ≥ 4	Bins ≥ 3	Bins ≥ 2	Bins ≥ 1	Bins ≥ 0
FPF		0.040	0.100	0.120	0.280	0.360	0.360	1
LLF		0.245	0.296	0.347	0.408	0.459	0.510	1