ROC ANALYSIS IN THE EVALUATION OF INTELLIGENT MEDICAL SYSTEMS

Emmanuel C. Ifeachor and Brahim Hamadicharef

Signal Processing and Multimedia Communications Research
School of Computing, Communications and Electronics
University of Plymouth, Drake Circus
Plymouth PL4 8AA, Devon, U.K.
E.Ifeachor@plymouth.ac.uk

Abstract

A large number of intelligent medical systems exist, but few are in routine clinical use. This is due, in part, to a lack of a robust objective method to quantify the performance of such systems. Potentially, ROC analysis could form a basis for a robust and objective evaluation of intelligent medical systems, but existing methods of ROC analysis require large sample sizes to be statistically valid. However, evaluation of intelligent medical systems often involve a small number of cases (because of time and cost of collecting 'gold standards') and so confidence bounds are required for ROC indices of performance. In this paper we present a new method for generating the probability density functions (pdfs) and confidence bounds for ROC points which is robust and accurate for any sample size. The method is generic and is particularly suited for evaluating the performance of systems where sample sizes are small. We illustrate the use of the method by applying it to assess the performance of two medical systems taken from the literature. The method has been implemented in C and in MATLAB.

Key words - intelligent medical systems, ROCanalysis, confidence bounds, probability density function.

1 Introduction

A large number of intelligent medical systems (including medical expert systems, neural networks, classifiers, knowledge discovery and data mining systems) have been, and are being, developed to practically aid the busy clinician and to improve patient care in areas such as diagnosis,

prognosis, decision support and screening, but few such systems are in routine clinical use. This is partly because of uncertainty about safety, performance and/or clinical effectiveness of such systems. The lack of a robust and objective method for evaluating intelligent medical systems has contributed significantly to the uncertainty. A robust method should allow us to objectively compare an intelligent medical system to human experts and to make accurate predictions of the risks of introducing the system into practice to ensure it is safe and will perform at an acceptable level. In medicine and healthcare, where safety is critical, it is important to have an objective method of evaluating the performance of intelligent medical systems if such systems are to be widely accepted in clinical practice.

Potentially, analysis of receiver operating characteristic (ROC) curves could form a basis for a robust and objective evaluation of intelligent medical systems. This should allow, for example, the diagnosis given by an intelligent system to be compared to a ”bench mark” diagnosis or ”gold standard” provided by clinical experts. Statistical uncertainties in performance must also be quantified so that the risk of deploying the system can be calculated. However, existing methods of ROC analysis require large sample sizes to be statistically valid. Medical intelligent system evaluations are expensive (e.g. in terms of time and resources) and for this reason they are usually carried out with a small number of cases. Thus, in practice the number of cases for which expert opinion can be obtained is severely constrained. The importance of sample size on the performance of intelligent medical systems is recognised, but the problems of small sample size which is inherent in intelligent medical systems has not been hitherto addressed.

In this paper, we will first introduce ROC curve analysis. We will then introduce a method that arose from a critical investigation into the potential role of ROC analysis as a basis for objective evaluation of intelligent medical systems which is robust and accurate for any sample size. The method allows visual analysis of confidence surfaces for an intelligent medical system and may be used to predict the risks associated with deploying such a system in clinical practice. The method is generic and may be applied to other complex diagnostic or classification problems. The practical use of the method is illustrated by retrospective analysis of two existing systems reported in the literature. In future, it will form part of a framework for robust objective

evaluation of the performance of intelligent medical systems developed within the BIOPATTERN project and the associated software made widely accessible.

2 Basics of ROC analysis

Taking the situation where there are two types of events, diseased and healthy, it is hoped to distinguish between them by measuring a characteristic property of these events, on an

ordinal, interval or ratio scale. Fig. 1 gives a hypothetical example of the relative frequency with which the two types of events give different values of the measured property.

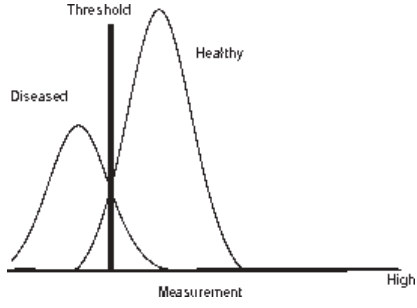

Figure 1: The underlying model for ROC curves

Table 1: Single threshold contingency

		Gold Standard	
		Diseased	Healthy
Test	Diseased	True Positive	False Positive
	Healthy	False Negative	True Negative

To distinguish between the two types of events, a threshold is chosen so that events with a measurement greater the threshold are labelled as 'healthy', and events with a measurement less than the threshold are labelled as 'diseased'. Since the two distributions overlap, no threshold value will completely separate them. Table 1 shows the 2 x 2 contingency of the actual type of an event, against its test classification according to the threshold. This table assumes there is a standard by which the actual type of the event is known. The test can then be characterised by two ratios:

$$Hit\ Rate = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$False\ Alarm\ Rate = \frac{FalsePositive}{FalsePositive + TrueNegative}$$

		Gold Standard	
		Diseased	Healthy
Test	'Diseased'	$b_0 = 4$	$a_0 = 0$
	'Unknown'	$b_1 = 1$	$a_1 = 3$
	'Healthy'	$b_2 = 0$	$a_2 = 7$

Table 2: Hypothetical intelligent medical system

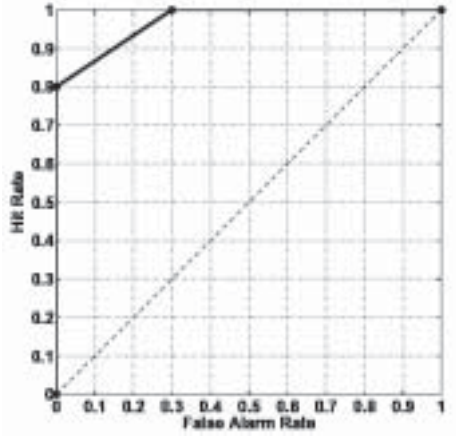

Figure 2: Typical ROC curve

The Hit Rate is the fraction of the total diseased cases the system gets right and the False Alarm Rate is the fraction of the total healthy cases the system gets wrong, i.e. misclassifies as diseased.

If multiple thresholds are used, for example, to categorise events into 'definitely diseased', 'possibly diseased', 'possibly healthy' and 'definitely healthy', the contingency table can be expanded to a 2 x n + 1 table, where n is the number of thresholds. For example for two threshold we obtain the 2 x 3 table, as shown in Table 2 and thus 2 pairs of Hit Rate and False Alarm Rate can be calculated (by moving a threshold down the table). Point 0 is for the 'Diseased' category alone, point 1 is for the 'Diseased' plus 'Unknown'. The resulting ROC curve is shown in Fig. 2.

With a sufficiently large number of thresholds changing in small discrete steps, as might be obtained from a Neural Network (NNet) for example, a plot of Hit rate (along the y axis) against False Alarm Rate (along x axis) for each threshold gives a ROC curve. The resulting ROC curve for a multi threshold plot is shown in Fig. 2. The ROC curve shows the trade off between correctly detecting a diseased case, and mistaking healthy case for a diseased. If the two underlying population distributions are well separated, the curve will immediately rise to the top corner (0.0, 1.0), and then proceed

horizontally. If the distributions tend to overlap, so that diseased case and healthy case cannot be distinguished by the measurement, the curve will approach the diagonal (dotted line from 0.0, 0.0 to 1.0, 1.0).

3 New approach to ROC analysis

There are a number of limitations of ROC analysis when applied to intelligent medical systems. The area under the curve (AUC) is a useful index which is often used to quantify performance of the system. However, it may be clinically more meaningful to consider points on the ROC curve than the use of AUC because intelligent systems are more likely to be used to make decisions about the status of an individual. Another limitation of existing ROC analysis arises when the number of points is small, as it is often the case in medicine because of reasons such as cost and time. In the above cases, confidence bounds are needed for the indices of performance. To address this we have introduced a robust and accurate method which calculates the probability density function (pdf) for each point on the ROC curve for any sample size and generates confidence bounds as contours on the point. The method is based upon asking the following question for every possible point on the surface of the ROC graph (see [1], [2] for full details): *"If this point represents the true Hit Rate and False Alarm Rate of the population, what would be the probability of getting the sample actually obtained?"*

If that question can be answered for every point on the graph, the relative probability of every point can then be calculated, and normalised such that the total relative probability sums to 1. This generates the normalised relative probability surface for the true Hit Rate and False Alarm Rate. By dividing the surface into a fine grid, and integrating the expression for the relative probability of every point over each square of the grid, the surface can then be presented as a 3D mesh, or contour lines can be drawn to enclose an arbitrary percentage of the probability, e.g. 95% of the probability, which gives the 95% confidence interval for the location of the true Hit Rate and False Alarm Rate.

Consider the full situation in which y is the Hit Rate of the population, given as a probability; x is the False Alarm Rate of the population, given as a probability; f is the frequency of disease events in the population, given as

a probability; $b0$ is the number of True Positive (TP) in the sample, $a0$ is the number of False Positive (FP) in the sample, $b1$ is the number of False Negative (FN) in the sample, $a1$ is the number of True Negatives (TN) in the sample. Then, P, the probability of a ROC point being at the location (x, y), is given by:

$$PtProb_{xy} = \frac{x^{a_0}(1-x)^{a_1}}{\int_0^1 x^{a_0}(1-x)^{a_1}dx} \frac{y^{b_0}(1-y)^{b_1}}{\int_0^1 y^{b_0}(1-y)^{b_1}dy} \quad (1)$$

Beta function are used to substitute for the integrals (See Appendix in [1] for details) obtaining

$$PtProb_{xy} = \frac{x^{a_0}(1-x)^{a_1}}{\frac{a_0!a_1!}{(a_0+a_1+1)!}} \frac{y^{b_0}(1-y)^{b_1}}{\frac{b_0!b_1!}{(b_0+b_1+1)!}} \quad (2)$$

To represent the surface, it is divided into a fine grid and the probability of each quantised grid square is calculated by integrating the probability at a point, over the area of each grid square. The integral over the area is equal to the product of the two one-dimensional (1-D) integrals along the Hit Rate and False Alarm Rate axes. Therefore, two vectors, **X** and **Y**, each with i elements, are defined to hold the 1-D integrals

$$\mathbf{X}_i = \frac{\int_{(i-1)/n}^{i/n} x^{a_0}(1-x)^{a_1}dx}{\frac{a_0!a_1!}{(a_0+a_1+1)!}} \quad (3)$$

and

$$\mathbf{Y}_i = \frac{\int_{(i-1)/n}^{i/n} y^{b_0}(1-y)^{b_1}dy}{\frac{b_0!b_1!}{(b_0+b_1+1)!}} \quad (4)$$

For all i from $i = 0$ to $i = n$. The pdf, quantised as a fine grid, is therefore the product of the two vectors

$$Surface = \mathbf{X} \cdot \mathbf{Y}^T \quad (5)$$

the integration of (3) and (4) give both (6) and (7).

Generalized mathematical equations for generating the pdfs and confidence bounds for multiple ROC points have been derived (see [1] and [2]). A C-language implementation is described in [2] and a MATLAB

implementation has recently been developed. The MATLAB implementation will be developed into a toolkit and made accessible in the near future.

4 Illustrative Examples

Two examples taken from the literature are used to illustrate the use and benefit of the new method.

The first example is taken from Swets [3] in which a radiological example was presented to illustrate the use of ROC analysis and the Area Under the Curve (AUC) as a measure of accuracy. A study had previously been carried out, in which six radiologists were asked to examine and classify 118 mammograms (58 malignant, 60 benign) into one of five categories, according to the likelihood that the lesion was malignant. The radiologists first diagnosed the mammograms unaided (denoted as 'standard'), and then used two diagnostic aids (denoted as 'enhanced'). The raw data for the pooled categorisations were given in the paper, allowing the ROC graph for the standard and enhanced diagnoses to be reproduced as shown in Fig. 3. One should note that the validity of pooling the data from six experts to produce one ROC curve is debatable, however, the data does provide a useful example of a ROC curve with a high sample size.

The second example is taken from Adlassnig and Scheithauer [4], in which an expert system, known as CADIAG- 2/PANCREAS, for the differential diagnosis of ten different types of pancreatic disease, is described. The performance of the system was compared to an histologically or clinically confirmed 'Gold-Standard' diagnosis. There were 47 patient records available in which one or more of a subset of six pancreatic diseases had been diagnosed. Four patients had dual diagnoses, giving a total of 51 diagnoses of one of six diseases. A series of ROC graphs were presented, illustrating the performance of the CADIAG-2 system in the differential diagnosis of specific diseases, both using what was described as a limited set of patient data and with the 'full' set of available patient data. Two of Adlassnig and Scheithauers ROC curves, (See Fig. 4), illustrate the evaluations of 8 diagnoses of acute pancreatitis from the 51 cases compared to the 'Gold-Standard', using 'limited' patient data and 'full' patient data respectively. Although the raw data were not given, they can be reconstructed from the ROC graphs, because the number of cases was small. The data are combined

here and reproduced as Fig. 4. However, originally 11 categories were used, which because of the small number of cases often left zero cases in a category which gives an ambiguous reconstruction. Thus here only five categories have been reconstructed. It should also be noted that this makes the confidence boundaries wider and so this re-analysis is slightly unfair to Adlassnig and Scheithauer, but still illustrates the point about sample sizes.

$$\mathbf{X}_i = (a_0 + a_1 + 1)! \sum_{k=0}^{a_1} \frac{(\frac{i}{n})^{a_0+a_1+1-k}(1-\frac{i}{n})^k - (\frac{i-1}{n})^{a_0+a_1+1-k}(1-\frac{i-1}{n})^k}{k!(a_0+a_1+1-k)!} \tag{6}$$

$$\mathbf{Y}_i = (b_0 + b_1 + 1)! \sum_{k=0}^{b_1} \frac{(\frac{i}{n})^{b_0+b_1+1-k}(1-\frac{i}{n})^k - (\frac{i-1}{n})^{b_0+b_1+1-k}(1-\frac{i-1}{n})^k}{k!(b_0+b_1+1-k)!} \tag{7}$$

For all i from $i = 0$ to $i = n$

From the raw data the 95% confidence boundary of each point on the four curves was calculated using tools described in [2] with a grid of 256 x 256 and are shown in Fig. 5, and Fig. 6. From Fig. 6, it is obvious that there is considerable uncertainty in the results due to the limited number of cases used. In their analysis, Adlassnig and Scheithauer state that accuracy was always increased by adding the 'full' patient data, in accordance with anticipation. While the ROC curves presented in Fig. 4 appear to support this commonsense conclusion, the large confidence boundaries shown in Fig. 6 suggest that this conclusion was probably premature given the data.

Fig. 5 and 6 clearly illustrate the difference that sample size makes to the confidence that can be placed in the location of each point. While the ROC curve of the mammogram diagnoses in Fig. 3 do not look as accurate as the ROC curve for the diagnosis of acute pancreatitis in Fig. 4, examination of the confidence boundaries in Fig. 5 and 6 shows that the 708 (six opinions of 118) mammograph cases are sufficient to give good confidence of the location of the ROC points, while the 51 pancreatic cases give a much larger posterior boundary.

In particular, it can be seen by consideration of pairwise points in Fig. 5, that the points are outside of each others con- fidence boundaries in all cases, and that the posterior boundaries are mutually exclusive in one case. In contrast, in Fig. 6, there is a high degree of overlap in posterior boundaries in all cases. In particular, the points with False Alarm Rate of 0.093 lie within each others confidence boundaries, and the 'limited' data point with False Alarm Rate

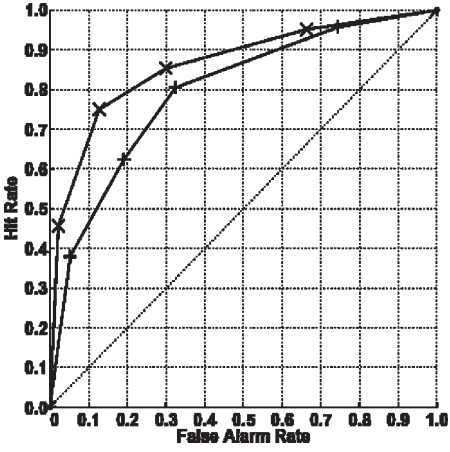

Figure 3: ROC curve of diagnosis of 708 mammograms (from Swets [3])

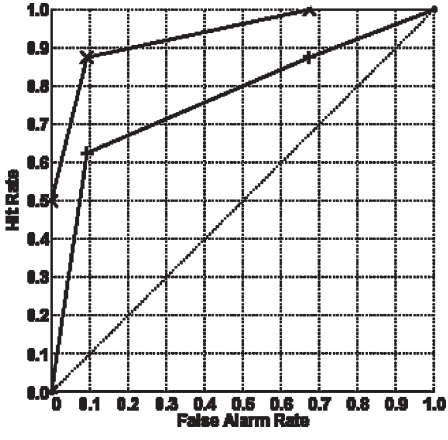

Figure 4: ROC curve of diagnosis of acute pancreatic from 51 cases (from Adlassnig & Scheithauer [4]

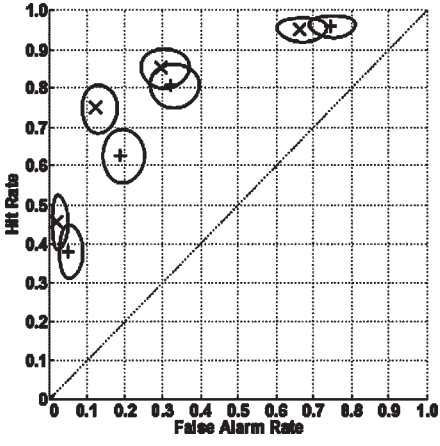

Figure 5: 95% Confidence boundary of ROC points in Fig. 3

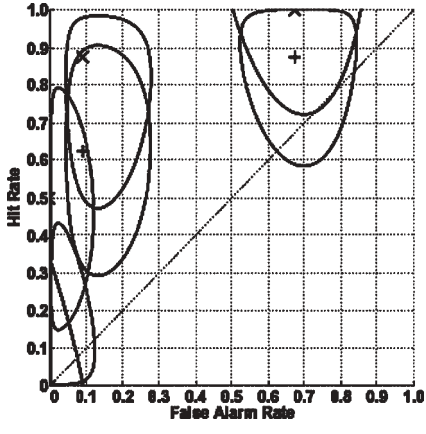

Figure 6: 95% Confidence boundary of ROC points in Fig. 4

of 0.638 lies well within the confidence boundary of the 'full' point.

5 Conclusions

In this paper we have presented a new method for the evaluation of intelligent medical systems using ROC analysis. The method is robust and valid for any sample size, removing one major limiting factor of standard ROC analysis. Potentially, the method could be extended to determine risks associated with deployment of intelligent medical systems in clinical practice in area such as brain diseases [5] and cancer [6]. A Bayesian

formulation of the method can be found in [7]. In future, it will form part of a framework for robust objective evaluation of the performance of intelligent medical systems within the BIOPATTERN project and the associated software will be made widely accessible.

Acknowledgments

We acknowledge the financial support of the European Commission (The BIOPATTERN Project, Contract No. 508803) for part of this work. The contributions and assistance of Dr J Tilbury and Dr V Stabovskaya are gratefully acknowledged.

References

[1] J. Tilbury, P. Van-Eetvelt, J. Garibaldi, J. Curnow, and E. Ifeachor, "Receiver Operator Characteristic Analysis for Intelligent Medical Systems - A New Approach for Finding Confidence Intervals," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 952–963, 2000.

[2] J. B. Tilbury, "Evaluation of Intelligent Medical Systems," Ph.D. thesis, Department of Communications and Electronic Engineering (DCEE), University of Plymouth, Drake Circus, Plymouth PL4 8AA, Devon, United Kingdom, September 2002. [Online]. Available: http://www.tech.plym.ac.uk/spmc/people/jtilbury/ pdf/JulianTilburyPhD.pdf

[3] J. A. Swets, "Measuring the accuracy of diagnostic systems," *Science*, vol. 204, no. 4857, pp. 1285–1293, June 1988.

[4] K. P. Adlassnig and W. Scheithauer, "Performance evaluation of medical expert systems using ROC curves," *Computers and Biomedical Research*, vol. 22, no. 4, pp. 297–313, 1989.

[5] G. T. Henderson, E. C. Ifeachor, H. S. K. Wimalartna, E. M. Allen, and N. R. Hudson, "Electroencephalogrambased methods for routine detection of dementia," *Proceedings* of the 4th International Workshop on Biosignal *interpretation, Como, Italy*, pp. 319–322, June 24-26 2002.

[6] M. Metz, W.-D. Groch, J. Curnow, E. Ifeachor, and P. Kersey, "Computer aided lesion border detection applied to macroscopic images," *Proceedings of the 4th International* Conference on Neural Networks and Expert Systems in Medicine and Healthcare (NNESMED'01), *Milos Island, Greece*, June 2001.

[7] J. Tilbury, P. Van-Eetvelt, J. Curnow, and E. Ifeachor, "Objective evaluation of intelligent medical systems using a bayesian approach to analysis of ROC curves," Proceedings of the 1st International Conference on Computational Intelligence in Medicine and Healthcare *(CIMED'03), Sheffield, United Kingdom, July, 2003*, 2003.