Objective Evaluation of Intelligent Medical Systems using a Bayesian Approach to Analysis of ROC Curves

Julian B. Tilbury, Peter W. J. Van Eetvelt, John S. H. Curnow, Emmanuel C. Ifeachor
University of Plymouth, Plymouth, PL4 8AA

Abstract Receiver operating characteristic (ROC) curves are commonly used to quantify the performance of intelligent medical systems. Because of the time and expense in collecting test cases it is beneficial if robust ROC curve analysis can be applied to small numbers of test cases. Using the Bayesian, rather than the Frequentist, approach gives considerable advantages in robustness and understanding. This paper compares the Frequentist and Bayesian approaches to nonparametric ROC analysis, and introduces a robust Bayesian method for parametric ROC analysis.

Keywords: Bayes, ROC, Medical system, Evaluate.

1 Introduction

We present an overview of a research project to investigate the use of Bayesian statistics for ROC curves analysis for intelligent medical system evaluation. The project showed that Bayesian statistics are particularly useful when the sample size is small, which is a problem commonly encountered in evaluating intelligent medical systems. Ideally the more cases used to test an intelligent medical system the better, but the practicalities of the time and cost of collecting each test case, and of defining a Gold standard often limits numbers. Using small numbers of test cases limits the statistical certainty of the conclusions that can be drawn from any analysis. Confidence intervals should always be used, regardless of sample size, but the lower the sample size the more dominant they become in correct interpretation of the results. Existing ROC analysis commonly takes a Frequentist, or classical, approach to ROC curve analysis that does not perform well with small sample sizes. The confidence intervals produced are at best misleading and at worst wrong. It is therefore suggested that those evaluating intelligent medical systems should consider the advantages of switching to Bayesian statistics.

2 The ROC Curve Example

Table 1 gives the data for a hypothetical intelligent medical system evaluation. The columns give the Gold standard diagnosis, the rows the tested system's opinion, right or wrong. Figure 1 shows the ROC curve plotted from this data.

		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 1

The Hit Rates and False Alarm Rates are given by the following ratios, generated by moving a threshold down the table. Point 0 is for the 'Diseased' category alone, point 1 for 'Diseased' plus 'Unknown'.

$$Hit\ Rate_0 = \frac{b_0}{b_0 + b_1 + b_2}$$

$$False\ Alarm\ Rate_0 = \frac{a_0}{a_0 + a_1 + a_2}$$

$$Hit\ Rate_1 = \frac{b_0 + b_1}{b_0 + b_1 + b_2}$$

$$False\ Alarm\ Rate_1 = \frac{a_0 + a_1}{a_0 + a_1 + a_2}$$

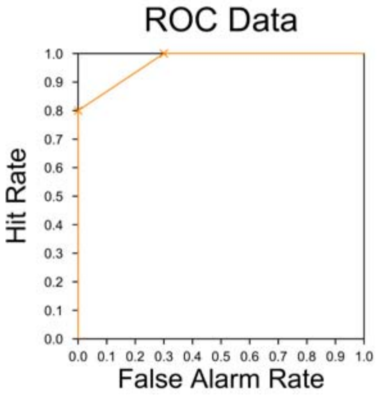

Figure 1

Plotting the ‘curve’ as straight-line segments is the nonparametric approach. Section 6 illustrates the parametric approach where a smooth curve is fitted to the data. Application of the methods to real data can be found in [1] and [2].

3 Frequentist Confidence Intervals

Figure 1 is deceptive in the apparent certainty of the points given the small number of test cases. Confidence intervals should be added. Figure 2 shows confidence intervals (CIs) as calculated using a well-known Frequentist method [3] given by Equations 1 and 2 below:

$$\sigma = \sqrt{\frac{HitRate\ (1 - HitRate)}{DiseasedCases - 1}} \qquad (1)$$

$$\sigma = \sqrt{\frac{FalseAlarmRate\ (1 - FalseAlarmRate)}{HealthyCases - 1}} \qquad (2)$$

Where σ is the standard deviation (Sd). The 95% CI is 1.96 × σ, assuming a Gaussian distribution.

The False Alarm Rate CI of point 0, and the Hit Rate CI of point 1 have zero width. This implies that this False Alarm Rate and this Hit Rate are known with certainty. Given the number of cases, this defies common sense.

Four observations can be made. Firstly, this statistical test is not recommended for such a low sample size. Secondly, this test is not recommended when the Hit Rate (or False Alarm Rate) is close to 0 or 1. Thirdly, even if the recommendations above are ignored, the zero width confidence intervals are absolutely correct within the Frequentist statistical paradigm. The Frequentist paradigm makes an estimate of the Hit Rate of the population by assuming it is the same as the Hit Rate of the sample. For point 1, 5 out of 5 diseased cases are categorised as either ‘Diseased’ or ‘Unknown’, which means *Hit Rate*$_1$ is 1.0. This is now assumed to be the Hit Rate of the population. The confidence interval can now be calculated by seeing how a hypothetical infinite number of samples drawn from this population would be distributed. If the population Hit Rate really was exactly 1.0, then a_0 really would be zero for every sample that was drawn. The confidence interval really would have zero width.

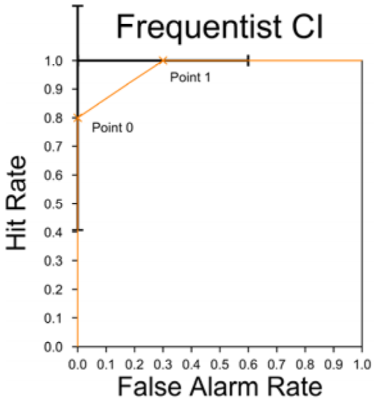

Figure 2

The fourth point is that one of the confidence intervals is outside the ROC graph, which is obviously nonsense. In this case it would be wise to concur with points one and two and not use this method at all! However, sometimes the Frequentist confidence interval is defined as a region that will contain at least 95% of further hypothetical samples, therefore including areas that cannot possible contain a sample is within the definition. Such obfuscation has nothing to commend it.

4 The Bayesian Approach

The Bayesian approach gives a more satisfactory answer. The approach does not estimate the population point with such absolute certainly, but instead uses a probability density function (pdf) of its location. This is called the Bayesian prior. The sample is then used as evidence to update this probabilistic view of the population point’s location using Bayes' law. The simplest distribution to use for the population point is the uniform probability distribution as given by Equation 3 and plotted in Figure 3.

$$p(x) \propto x^0\ (1 - x)^0 \qquad (3)$$

In the case of the False Alarm Rate for point 0, the posterior distribution, or likelihood function, is then give by [2]:

$$p(Far \mid a_0, a_1 + a_2) = \frac{p(a_0, a_1 + a_2 \mid Far)\, p(Far)}{\int_0^1 p(a_0, a_1 + a_2 \mid Far)\, p(Far)\, \mathrm{d}Far} \qquad (4)$$

Where Far is the False Alarm Rate; $p(Far)$ is the probability of the False Alarm Rate; $p(a_0, a_1 + a_2 \mid Far)$ is the probability of obtaining the data a_0, a_1, a_2 for a given value of Far; $p(Far \mid a_0, a_1 + a_2)$ is the posterior probability, the probability of a given value of Far conditional on the data a_0, a_1, a_2.

(NB: the denominator gives the probability of obtaining the data a_0, a_1, a_2 independent of Far by integrating out Far)

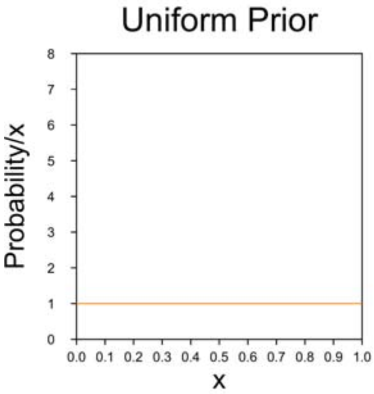

Figure 3

This Bayesian posterior distribution for the False Alarm Rate of point 0 is calculated using Equation 4 and the resultant Bayesian posterior distribution plotted in Figure 4.

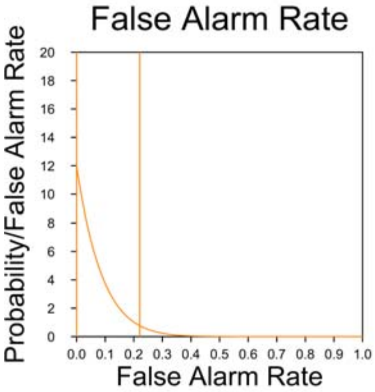

Figure 4

In place of a 95% confidence interval the 95% Bayesian posterior interval is used; it is shown as a vertical line. This method can be applied to the Hit Rate as well, and the pdfs combined.

Figure 5 shows the method applied in both dimensions for both points of the original ROC curve in Figure 1. The 95% posterior intervals have been shown as contours around the posterior pdf. This gives a result more in tune with intuitive understanding of the statistical uncertainties of using such a small sample size.

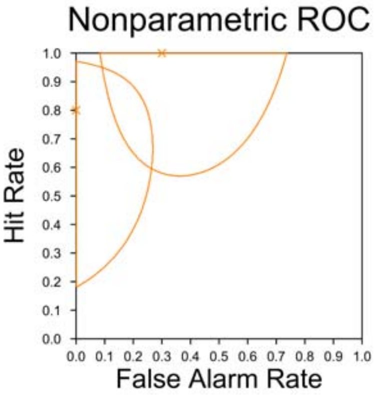

Figure 5

However, it is still dependent on an assumption. The assumption is that the Bayesian prior is uniform, i.e before we saw the evidence we believed that every population False Alarm Rate was equally likely (Equation 3, Figure 3). Alternatively, the ‘uninformative’ Bayesian prior (Equation 5) is a possible candidate.

$$p(x) \propto x^{-1} \ (1-x)^{-1} \qquad (5)$$

This has some appealing mathematical properties [2], but it cannot be plotted as it evaluates to infinity when x is 0 or 1, however, Figure 6 plots the function excluding these two end values. In the case of the data in Table 1, the infinity will also appear in the posterior distribution, and so an uninformative prior is unhelpful in these circumstances.

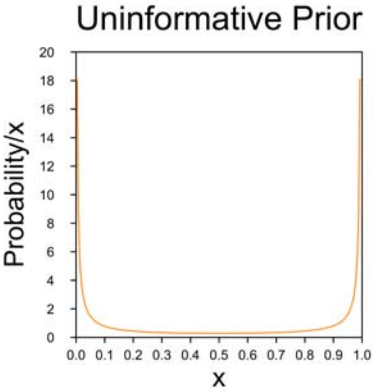

Figure 6

It should be noted that while the Bayesian and Frequentist paradigms are radically different with small sample sizes, the results asymptotically approach each other as the sample size rises. This is also true for different Bayesian priors. Suppose that a_0=80 and a_1+a_2=120. This gives a Frequentist population False Alarm Rate estimate of 0.4. The central limit theorem predicts that the distribution of repeated samples of 200 cases will approximate a Gaussian distribution. Equation 1 can be used to generate the standard deviation of this distribution. All three distributions, the Frequentist, the Bayesian with a uniform prior, and the Bayesian with an uninformative prior, can now be compared by plotting them on the same graph as shown in Figure 7.

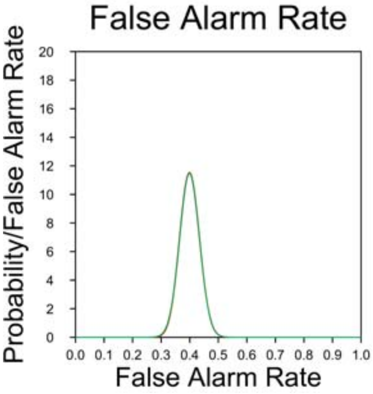

Figure 7

At the resolution of Figure 7 all three distributions appear identical. Though Bayesian statistics and Frequentist statistics are completely different paradigms, this convergence as sample size rises is reassuring. However, it is our opinion that the Bayesian paradigm has significant advantages in common sense interpretation when the sample size is small as demonstrated above.

The example in Figure 5 of the 2-D 95% Bayesian posterior interval was calculated by Dirichlet integrals of a family of equations starting with Equation 4. Full details can be found in [1]. The Dirichlet integrals give an analytic answer from which the graphs were produced directly [1]. The Bayesian approach can also be used for other aspects of ROC analysis, but here the posterior distributions must be approximated using a numerical approach.

5 Bayesian Nonparametric AUC

ROC curves are often quantified by using the Area Under the Curve (AUC) as a summary statistic. The AUC gives the probability that a system can correctly diagnose the diseased case when presented with a random pair of one diseased case and one healthy case [3]. Because this scenario is unlikely in any practical situation it may make more sense to restate the AUC as the probability of correctly diagnosing each case when presented with every possible pair of one healthy and one diseased case from the sample.

Much as it would be desirable to generate the Bayesian posterior distribution for the AUC analytically, our conjecture is that the integration would require a calculation running in time proportional to 3^p, (where p is the number of points). If the ROC graph is quantized into a grid of G×G cells, with the curve approximated by a series of lines joining these cells, the pdf of the AUC can be estimated by an algorithm running in time proportional to G^5 and using memory proportional to G^3 [1].

Figure 8 shows the Bayesian posterior distribution of the AUC of the ROC curve in Figure 1 calculated by this method. In contrast, the Frequentist method calculates the population AUC as equal to the sample AUC of 0.97. The standard deviation can then be calculated using DeLong et al.'s method [4]. This gives a value of 0.0337. Given that Hoeffding [5] proved the distribution is asymptotically Gaussian the 95% confidence interval is therefore ±0.066. This is shown in Figure 9.

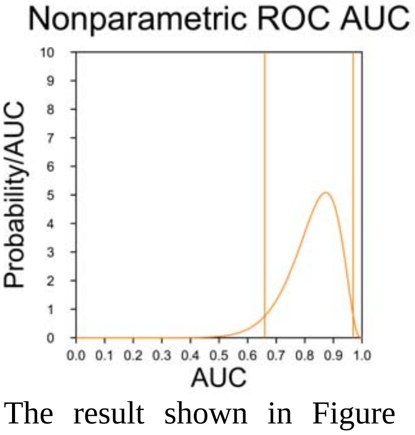

Figure 8

The result shown in Figure 9 is not only radically different from the Bayesian answer it is also obviously nonsense – the upper bound of the 95% confidence interval (1.036) exceeds the maximum possible value of the AUC (1.0)!

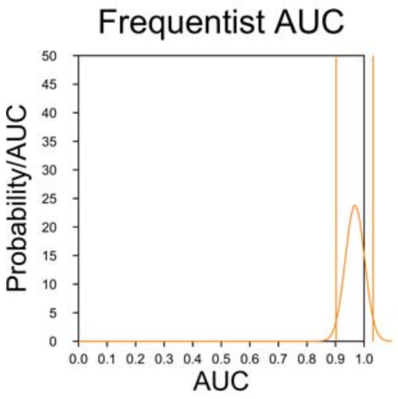

Figure 9

6 Bayesian Parametric ROC

The ROC curve in Figure 1 makes rather literal interpretation of the data points. However, by assuming an underlying parametric model a curve can be fitted. A common model is to assume that the diseased and healthy populations from which the sample is drawn are Gaussian, resulting in a bi-normal ROC curve[1].

The most sophisticated current method of producing a parametric ROC curve is to use a maximum likelihood calculation to find the population curve with the highest probability of generating the sample points [6]. This is actually a Bayesian method, but the confidence interval is then calculated assuming the sample distribution is Gaussian. As illustrated by Figure 7, this is correct for large sample sizes, but is inaccurate for a small sample sizes. However, the biggest flaw of the method is that it is not robust. So called 'degenerate cases', which include the example used here, fail to converge in existing maximum likelihood algorithms. Degenerate cases are commoner the smaller the sample size and the larger the AUC.

A robust algorithm has been developed that can cope with 'degenerate cases' and can also plot the Bayesian posterior interval for the sample [1]. The algorithm quantizes the ROC graph into a G×G grid and runs in time proportional to G^3, but takes memory proportional to G^4. This memory requirement means that the grid size is only 128×128 in the figures so the graininess is visible. Figure 10 shows the maximum likelihood bi-normal ROC curve (to the limits of the underlying grid resolution) of the 'degenerate' data given in Table 1.

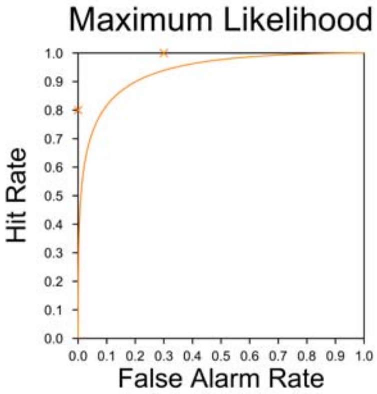

Figure 10

In existing bi-normal parametric ROC analysis ROC curves are quantified by two parameters, α and β, which are the normalised difference of mean and the ratio of the standard deviations of the underlying healthy and diseased Gaussian distributions from which the ROC curve is derived. A point on a 2-D plot of α and β will therefore uniquely specify a ROC curve.

The 95% posterior interval for a bi-normal ROC curve could therefore be illustrated on an (α, β) graph except for the fact that α has a range of ±∞, and β a range of 0 to +∞. In order to plot this graph new parameters were defined:

$$Healthy\ \sigma = \frac{2\sigma_h}{\sigma_d + \sigma_h} \qquad Disease\ \sigma = \frac{2\sigma_d}{\sigma_d + \sigma_h}$$

$$\Delta\mu = \frac{\mu_h \times Disease\ \sigma - \mu_d \times Healthy\ \sigma}{2}$$

Where σ_h is the absolute standard deviation of the healthy population; σ_d is the absolute standard deviation of the diseased population; μ_h is the absolute mean of the healthy population; μ_d is the absolute mean of the disease population.

The pair *Healthy* σ and *Disease* σ are now in the range 0 to 2, but $\Delta\mu$ is in the range ±∞. A

[1] Since monotonic transformations of the underlying data do not affect the shape of the resulting ROC curve, it can be more rigorously stated that the underlying distributions have to be latently Gaussian to give a bi-normal ROC curve.

sigmoid function was therefore used to give a range of ±1:

$$Healthy\ Mean - Disease\ Mean = \frac{1}{1+e^{\Delta\mu}}$$

Figure 11 shows a plot of the 95% posterior interval of *Healthy Mean – Disease Mean* and *Healthy* σ, (or *Disease* σ) of the healthy and diseased populations.

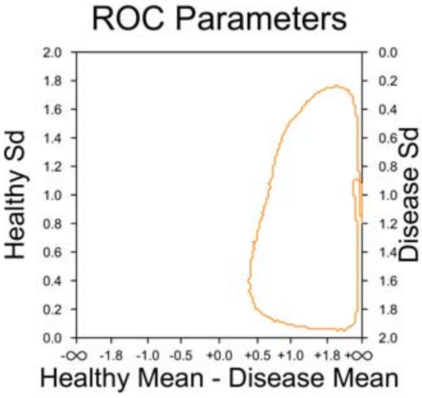

Figure 11

This algorithm can also be used to generate the posterior interval of the parametric AUC. It is shown here in Figure 12.

It is very similar to the distribution of the nonparametric AUC shown in Figure 8; the 95% interval is similar in width, but the peak AUC is slightly larger. This should be expected as the convex curve of a parametric ROC curve encloses a greater area than the straight lines of a nonparametric curve of the same data.

Existing methods fail to produce **any** ROC curve for this data, as it is 'degenerate'.

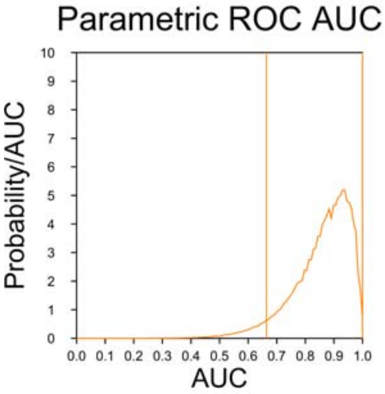

Figure 12

7 Conclusion

A brief overview of novel methods for Bayesian analysis of ROC curves is presented. Full details can be found in [1]. Methods for both nonparametric and parametric ROC curves that are robust and accurate for all sample sizes are described. In contrast existing methods give results that can be counterintuitive, obfuscated, or wrong, or do not produce any results at all. Bayesian methods are robust and accurate at small samples, and produce the same answers as Frequentist methods when the sample size is large. All the algorithms presented here have been tested by extensive Monte Carlo simulations. However, they still have disadvantages. Bayesian analysis requires a prior distribution, and some of the methods presented require considerable computer resources. However, the robustness, particularly at low sample sizes, make Bayesian methods a valuable contribution to the evaluation of intelligent medical systems where sample size is limited by the time and expense involved in collecting test cases.

Acknowledgement

Thanks to Rebecca O'Brien for assisting in preparation of this paper.

References

1. J. B. TILBURY, P. W. J. VAN EETVELT, J. M. GARIBALDI, J. S. H. CURNOW, E. C. IFEACHOR, "Receiver Operating Characteristic Analysis for Intelligent Medical Systems – A New Approach for Finding Confidence Intervals", IEEE Transactions on Biomedical Engineering, 2000, Vol.47, No.7, pp.952-963.
2. J. B. TILBURY, "Evaluation of Intelligent Medical Systems", PhD Thesis, 2002, University of Plymouth, UK.
3. D. M. GREEN, J. SWETS, "Signal detection theory and psychophysics", Wiley, New York, 1966, pp.45-49.
4. E. R. DELONG, D. M. DELONG, D. L. CLARKE-PEARSON, "Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach", Biometrics, 1988, Vol.44, pp.837-845.
5. W. HOEFFDING, "A class of statistics with asymptotically normal distributions", The Annals of Mathematical Statistics, 1948, Vol.19, pp.293-325.
6. D. D. DORFMAN, E. ALF, "Maximum-Likelihood Estimation of Parameters of Signal-Detection Theory – a direct solution", Psychometrike, 1968, Vol.33, pp.117-124.