ROC curves and the χ^2 test

Andrew P. Bradley

Cooperative Research Centre for Sensor Signal and Information Processing,

Dept. of Electrical and Computer Engineering,

The University of Queensland, QLD 4072. AUSTRALIA.

bradley@elec.uq.edu.au

Abstract

In this paper we review the Receiver Operating Characteristic (ROC) curve, and the χ^2 test statistic, in relation to the analysis of a confusion matrix. We then show how these two methods are related, and propose an extension to the ROC curve so that it shows contours of χ^2 values. These contours can be used to provide further insight into the appropriate setting of the decision threshold for a particular application.

Keywords — Misclassification Cost, Receiver Operating Characteristic (ROC), χ^2 test, Neyman-Pearson method.

1 Introduction

In many pattern recognition schemes classification performance is measured in terms of the overall probability of error (Fukunaga, 1990). However, sometimes it is useful to treat the errors from each class separately, and associate the idea of a *misclassification cost* with each

class (Weiss & Kulikowski, 1991). In this case, classification performance is often specified by an upper bound on the probability of error for the class with the highest misclassification cost. An example of this is in the case of screening for cervical cancer, here the cost associated with wrongly predicting an abnormal slide as normal is high (the abnormality may go untreated). Whilst, there is a lesser cost associated with wrongly predicting a normal slide as abnormal (the slide has to be screened again by a Cytologist). Here, the performance of an automated cervical screener will be specified primarily in terms of the first type of error.

This paper deals with the frequently occurring problem of designing and evaluating classification systems in domains with different numbers of examples from each class (prior probabilities) and different misclassification costs. It tackles the problem of estimating from the data just how difficult a desired operational point would be to obtain in practice. This operational point being defined in terms of misclassifications from one class as compared to misclassifications from the others. This work is related to estimating the probability of error on a given data set, an area which has seen much interest, particularly in relation to feature set selection (Devijver & Kittler, 1982; Fukunaga, 1990). Here however, we are interested in varying a decision threshold, or misclassification costs, in order to aid the selection of an appropriate operational point. To do this we propose to combine two well known concepts from statistical pattern recognition, the Receiver Operating Characteristic (ROC) curve, and the χ^2 test statistic. The proposed ROC curve with χ^2 contours then provides additional information to aid with the selection and evaluation of a classifier's operational point.

First some background to binary classification problems is given in Section 2, then Sections 3 and 4 describe the ROC curve and the χ^2 test in relation to the analysis of a classifier's confusion matrix. Section 5 then shows a relationship between these two methods that enables us to visualise the amount of "work" involved in attempting to meet a specified operational point. Finally Section 6 discusses the material presented and some of the advantages of using this type of analysis.

2 Binary Classification Problems

For binary classification problems there are two types of errors, called *false positives* and *false negatives*. A false positive is a classification of positive given to an example that is actually negative, and a false negative is the negative classification of an example that is actually positive. They are defined as follows:

$$P(\text{False Positive}) = P(\text{Classify Positive} \mid \text{Negative}). \quad (1)$$

$$P(\text{False Negative}) = P(\text{Classify Negative} \mid \text{Positive}), \quad (2)$$

In general, the probability of a false positive, (denoted α), is referred to as the *level of significance* of a test, and the probability of a true positive, (denoted, $1 - \beta$), the *power* of a test. In signal detection theory and in particular, radar theory, the plot of α versus $1 - \beta$ as the decision threshold is varied, is known as the "Receiver Operating Characteristic" or ROC curve (Selin, 1965). In this domain it is used to measure how well a receiver can detect signal from noise, but in general it measures the ability to classify positive from negative examples. The ROC curve has also become common in medical fields (Sherwood, Bartels & Wied, 1976) and is often used as part of the Neyman-Pearson method (Fukunaga, 1990). A simple example showing the effect of varying the decision threshold on the probability density functions for two classes is shown in Figure 1. The resulting ROC curve is then shown in Figure 2, the three threshold points of Figure 1 being marked. On a ROC curve a decision threshold that produces low overall error probabilities will be high in the upper left hand corner of the chart.

3 The ROC Curve

On a ROC chart, the diagonal line from (0,0) to (1,1) indicates the condition where $\alpha = 1 - \beta$. This is the line of no discrimination between the classes and is the locus that a random classification scheme *i.e.,* one that does not use any of the input attributes, would follow. This can be shown as follows. Given C_n negative cases, C_p positive cases, and a probability that

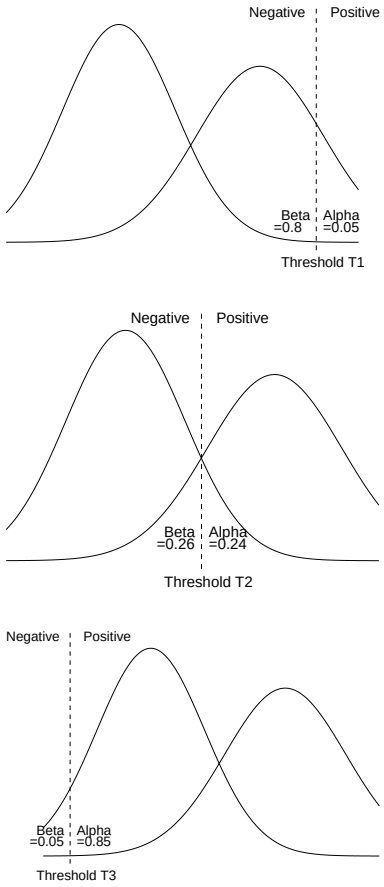

Figure 1: Three decision thresholds and their associated error rates.

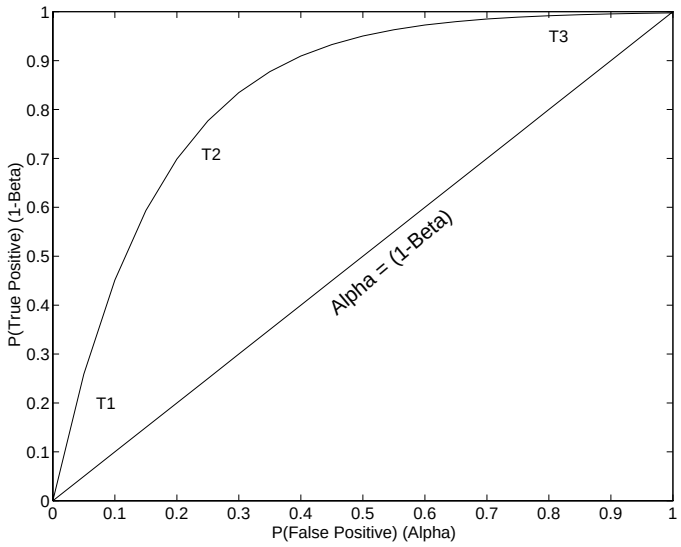

Figure 2: The ROC curve produced by changing the decision threshold.

we predict, at random, the class of an example as positive is P_p and negative is P_n, where $P_p + P_n = 1$. The expected number of true positives, $E(T_p)$, true negatives, $E(T_n)$, false positives, $E(F_p)$, and false negatives, $E(F_n)$, are given by

$$E(T_p) = C_p \cdot P_p \quad (3)$$

$$E(T_n) = C_n \cdot P_n \quad (4)$$

$$E(F_p) = C_n \cdot P_p \quad (5)$$

$$E(F_n) = C_p \cdot P_n \quad (6)$$

The ROC curve is then a plot of the *probability* of a false positive against the *probability* of a true positive. For a random classification scheme, P_p may be varied in the range $[0, 1]$, giving the locus

$$\alpha = 1 - \beta, \quad (7)$$

where $\alpha = P_p$ and $(1 - \beta) = P_p$. It should be noted that this is also the case for classifiers that predict one of the classes all of the time, in which case $P_p = 0$ or $P_p = 1$.

This means that the diagonal line from (0,0) to (1,1) can be thought of as the locus of the *expected* values of estimates of $1 - \beta$ and α. In other words, it is the ROC curve of a classifier which randomly assigns examples to classes with $P_p = 1 - P_n$, for P_p in [0,1]. The ROC curve obtained from a classification scheme that actually uses the input attributes is then the locus of the *observed* estimates of $1 - \beta$ and α for different decision thresholds.

4 The χ^2 Test

This view of the ROC curve as the difference between observed and expected values, offers a direct comparison to the χ^2 test. The χ^2 test is used as a measure of the independence between two variables (Everitt, 1992), in our case these variables are the true class and the predicted class of the examples. The χ^2 test is based upon the difference between the observed, O, and

expected, E, class frequencies for each cell, c, in a contingency table.

$$\chi^2 = \sum_c \frac{(O-E)^2}{E} \tag{8}$$

The χ^2 test can be used on the confusion matrix produced by a classification scheme (Ingelfinger et al., 1994,Chapter 8). A confusion matrix is a form of contingency table showing the differences between the true and predicted classes for a set of labelled examples, this is shown in Table 1.

	Predicted Class		
True Class	$^-$ve	$^+$ve	
$^-$ve	T_n	F_p	C_n
$^+$ve	F_n	T_p	C_p
	R_n	R_p	N

Table 1: A confusion matrix.

In Table 1, T_p, T_n, F_p, and F_n are counts of the numbers of true positives, true negatives, false positives and false negatives respectively, C_n and C_p are the number of *true* negative and positive examples, R_n and R_p are the number of *predicted* negative and positive examples, and N is the total number of examples. Although the confusion matrix shows all of the information about the classifier's performance, more meaningful measures can be extracted from the confusion matrix to illustrate certain performance criteria, for example:

$$\text{Accuracy (1}-\text{Error)} = \frac{(T_p + T_n)}{(C_p + C_n)} \tag{9}$$

$$\text{Sensitivity or Power } (1-\beta) = \frac{T_p}{C_p} \tag{10}$$

$$\text{Specificity } (1-\alpha) = \frac{T_n}{C_n} \tag{11}$$

$$\text{Positive predictive value} = \frac{T_p}{R_p} \tag{12}$$

$$\text{Negative predictive value} = \frac{T_n}{R_n} \tag{13}$$

The χ^2 value is another parameter that can be extracted from the confusion matrix. It is usually used to test the hypothesis that the confusion matrix was obtained by random selection

of examples from each class *i.e.,* without using the input attributes at all, this is often called the *null* hypothesis. In this case true and predicted classes will be independent, giving a low χ^2 value. In effect, the χ^2 value measures the (squared) difference between the observed and expected values of T_n, T_p, F_n, and F_p. If these populations are multinomial and the observed frequencies are not too small (say < 5), in which case Fisher's exact test is preferable (Sterling & Pollack, 1968), then the χ^2 statistic defined in Equation 8 will approximately follow a χ^2 distribution. If the value of the χ^2 test exceeds certain bounds the evidence to reject the *null* hypothesis is called significant ($\chi^2 > 3.84$; at the 5% level), highly significant ($\chi^2 > 6.63$; 1%) or very highly significant ($\chi^2 > 7.88$; 0.5%). In this paper, we propose to use the actual χ^2 value as a guide to selecting an operational point on the ROC curve.

5 The ROC Curve with χ^2 Contours

It is interesting to look at the similarities between the χ^2 statistic and the distance from the diagonal line ($\alpha = 1 - \beta$) for a point on a ROC curve. The χ^2 value is given from Equation (8) as

$$\chi^2 = \frac{(E(T_p) - T_p)^2}{E(T_p)} + \frac{(E(F_p) - F_p)^2}{E(F_p)} + \frac{(E(T_n) - T_n)^2}{E(T_n)} + \frac{(E(F_n) - F_n)^2}{E(F_n)}. \tag{14}$$

Because a 2×2 contingency table only has one degree of freedom

$$(E(T_p) - T_p) = (E(F_p) - F_p) = (E(T_n) - T_n) = (E(F_n) - F_n) = \varepsilon \ , \tag{15}$$

and so

$$\chi^2 = \varepsilon^2 \cdot (\frac{1}{E(T_p)} + \frac{1}{E(F_p)} + \frac{1}{E(T_n)} + \frac{1}{E(F_n)}). \tag{16}$$

Comparing this to the distance, d, from the observed to the expected values of the probabilities of false positives (α) and true positives ($1 - \beta$), we have

$$d \ = \ \sqrt{(\frac{E(T_p)}{C_p} - \frac{T_p}{C_p})^2 + (\frac{E(F_p)}{C_n} - \frac{F_p}{C_n})^2} \tag{17}$$

$$= \sqrt{\frac{\varepsilon^2}{C_p^2} + \frac{\varepsilon^2}{C_n^2}}. \quad (18)$$

So, we can see that they are both minimized when the observed value is equal to the expected value and are both maximised when $1 - \beta = 1$ and $\alpha = 0$, *i.e.,* a point in the top left hand corner of the ROC chart.

The χ^2 value is therefore not only dependent on d but also upon where in $(\alpha, 1 - \beta)$ space the observed values of α and $1 - \beta$ are. This is illustrated from Equations (16) and (18) by

$$\chi^2 = d^2 \cdot \frac{C_n^2 C_p^2}{C_n^2 + C_p^2} \cdot \left(\frac{1}{E(T_p)} + \frac{1}{E(F_p)} + \frac{1}{E(T_n)} + \frac{1}{E(F_n)}\right). \quad (19)$$

This relationship is best viewed graphically. Figures 3 and 4 show the χ^2 contours for two example distributions and illustrates how the χ^2 contours vary with different class distributions. The contours indicate that performance increases are best obtained in an approximately radial distance from the point of *null* classification. Here, the point of *null* classification is defined as in Equations (3) and (5), where

$$P_p = \frac{C_p}{(C_n + C_p)}, \quad (20)$$

and

$$P_n = \frac{C_n}{(C_n + C_p)}. \quad (21)$$

The χ^2 contours can be drawn onto the ROC chart, for given values of C_n and C_p, from Equation (16), where

$$E(T_n) = \frac{C_n(C_n \cdot (1 - \alpha) + \beta \cdot C_p)}{(C_n + C_p)}, \quad (22)$$

$$E(F_p) = \frac{C_n(C_p \cdot (1 - \beta) + \alpha \cdot C_n)}{(C_n + C_p)}, \quad (23)$$

$$E(F_n) = \frac{C_p(C_n \cdot (1 - \alpha) + \beta \cdot C_p)}{(C_n + C_p)}, \quad (24)$$

$$E(T_p) = \frac{C_p(C_p \cdot (1 - \beta) + \alpha \cdot C_n)}{(C_n + C_p)}. \quad (25)$$

In this case we evaluated the χ^2 values in one half (the upper triangular half) of a 40 by 40 matrix and drew only the required contours. These contours could also be found by evaluating

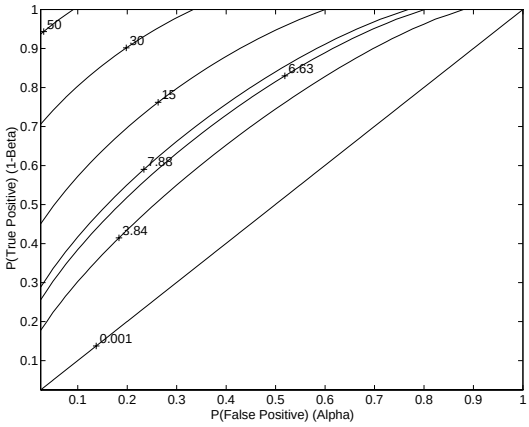

Figure 3: χ^2 contours for class distributions $C_n = C_p = 30$.

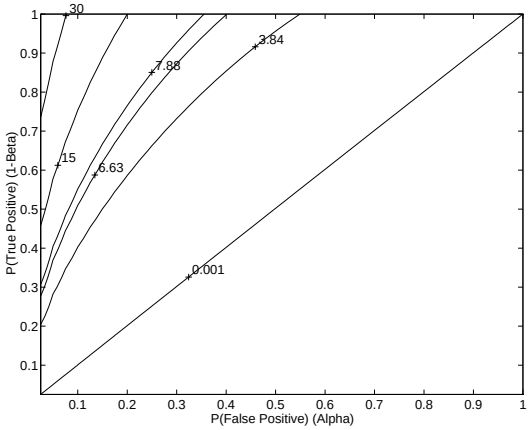

Figure 4: χ^2 contours for class distributions $C_n = 55$, $C_p = 5$.

the roots of Equation (16) for a particular χ^2 value, holding, say α, constant and numerically solving this as a function of $(1 - \beta)$ (Press et al., 1992,Chapter 9).

Remark: the χ^2 value is dependent upon the number of examples in the contingency table, so that if you increase T_p, F_p, T_n, and F_n by a factor of 10 then the χ^2 value also increases by a factor of 10. However, in any one particular case, it is the shape of the contours that are important.

Accuracy contours, as defined in Equation (9), are also included in Figures 5 and 6 to show how in cases with uneven class distributions it becomes "easier" to obtain seemingly high classification accuracies. In the case of distributions $C_n = 55$ and $C_p = 5$ a classification accuracy of 86% can be obtained by predicting the classes at random. Predicting negative all

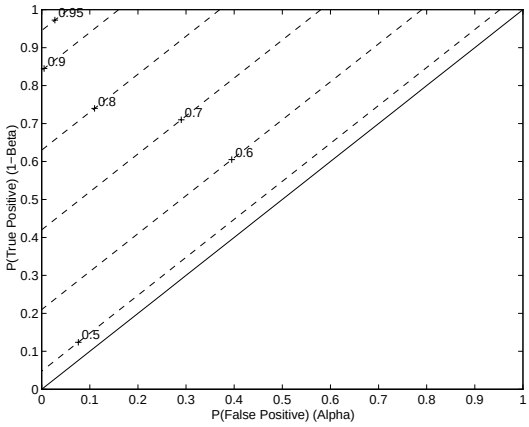

Figure 5: Overall accuracy contours for class distributions $C_n = C_p = 30$.

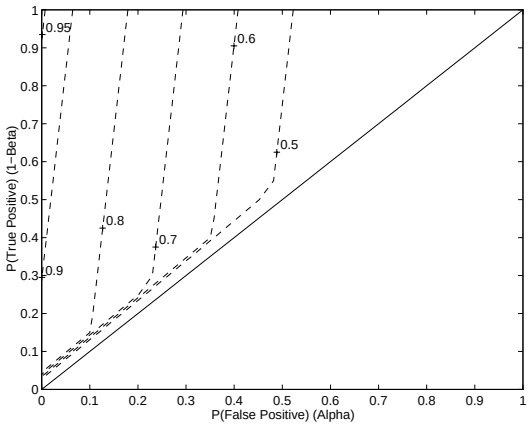

Figure 6: Overall accuracy contours for class distributions $C_n = 55$, $C_p = 5$.

the time leads to an overall accuracy of 92%, however, both these cases have very small χ^2 values and so are revealed as "dumb" classifiers. These "dumb" classifiers may also have been revealed by using the performance measures listed in Equations (10) to (13). So, we can see that on a ROC chart there is more to selecting an operating point than picking a point "near the top left hand corner."

6 Discussion and Conclusions

What we are proposing in this paper is that the value of the χ^2 statistic should be used as a measure of the amount of "work" a classification scheme is doing. Therefore, on a particular problem domain, operational points with larger χ^2 values will be harder to obtain. The χ^2 contours help us to quickly and easily to find the area of optimum classification, as they show where the highest gains can be made in relation to the class distributions of our sample. Figures 4 and 6 show that accuracy, on its own, is an insufficient measure of classifier performance, as "dumb" classifiers, *i.e.,* that are doing little actual "work," can have a seemingly high accuracy.

For two points (or classification schemes) that have the same accuracy, but different χ^2 values, the point (or classifier) with the higher χ^2 value is preferable as it will, in general, have better performance criteria as measured by Equations (10) to (13), or, in our terminology, the classification scheme with the highest χ^2 value is doing more "work." However, it should be noted that in domains with large differences in the prior probabilities of each class, the point of highest accuracy may not always be the point of highest χ^2 value. This reiterates the fact that the optimum operational point for a particular application is obtained by specifying one of the error probabilities and then minimizing the other, as in the Neyman-Pearson method, or by minimising the misclassification cost:

$$\text{Cost} = \sum_c F_c \cdot C_c, \tag{26}$$

where F_c and C_c are the number of false classifications and their associated cost, for each class,

c. The χ^2 contours relate not only to accuracy but also to distance from the *null* classification line ($\alpha = 1 - \beta$) on the ROC chart. They add further insight into the setting of operational points for classifiers using a ROC curve.

It should be noted that the idea of setting a decision threshold and producing a ROC curve, is quite general. In statistical analysis it relates to the positioning of a decision boundary, so as to minimise the Baye's risk (Fukunaga, 1990); Decision trees can use cost-sensitive construction or pruning (Knoll, Nakhaeizadeh & Tausend, 1994); neural networks can varying the class threshold of their output units (usually set to 0.5 for log-sigmoid activation functions) (Twomey & Smith, 1993); and the K Nearest Neighbour algorithm can vary its class decision threshold (usually set at $> \frac{k}{2}$). Plotting the ROC curve for a discriminant function, or the K Nearest Neighbours, using resubstitution (Weiss & Kulikowski, 1991), can then be used as an estimate of the ROC curve of the (optimal) Baye's error rate classifier. This ROC curve can then be used in a comparative study with proposed classification schemes, measuring perhaps the area under the ROC curve (Sherwood, Bartels & Wied, 1976) for evaluation purposes.

To conclude, we have shown a relationship between the χ^2 test on a confusion matrix and the distance of an observed point on a ROC curve from the line of no discrimination (where $\alpha = 1 - \beta$). To aid in the selection of decision thresholds we have proposed the use of χ^2 contours drawn on a ROC chart and this has been shown to be of use particularly in cases of uneven class distributions.

Acknowledgements

The Author is grateful to Paul Jackway and Brian Lovell for discussions that motivated this work. Thanks are also due to the anonymous referees for helpful comments on earlier drafts of this paper. This work is supported by the Cooperative Research Centre for Sensor Signal and Information Processing.

References

Devijver, P. A. & J. Kittler (1982), *Pattern Recognition: A statistical approach*, Prentice Hall, London.

Everitt, B. S. (1992), *The Analysis of Contingency Tables*, Monographs on Statistics and Applied Probability 45, Chapman and Hall, London, Second Edition.

Fukunaga, K. (1990), *Introduction to Statistical Pattern Recognition*, Academic Press, San Diego, CA, Second Edition.

Ingelfinger, J. A., F. Mosteller, L. A. Thibodeau & J. H. Ware (1994), *Biostatistics in Clinical Medicine*, McGraw-Hill, New York, Third Edition.

Knoll, U., G. Nakhaeizadeh & B. Tausend (1994), "Cost Sensitive Pruning of Decision Trees ," *Machine Learning: Proceedings of ECML-94*, 383–386.

Press, W. H., S. A. Teukolsky, W. T. Vetterling & B. P. Flannery (1992), *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, Second Edition.

Selin, I. (1965), *Detection Theory*, Princeton University Press.

Sherwood, E. M., P. H. Bartels & G. L. Wied (1976), "Feature selection in cell image analysis: use of the ROC curve," *Acta Cytol.* Vol. 20, No 3, 255–261.

Sterling, T. D. & S. V. Pollack (1968), *Introduction to statistical data processing*, Prentice Hall, 287–303.

Twomey, J. M. & A. E. Smith (1993), "Power Curves for Pattern Classification Networks," *Proceedings of IEEE International Conference on Neural Networks*, San Francisco, California, 950–955.

Weiss, S. M. & C. A. Kulikowski (1991), *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems*, Morgan Kaufmann, San Mateo.