Bradley, Andrew
(2013)
ROC curve equivalence using the Kolmogorov-Smirnov test.
Pattern Recognition Letters, *34*(5), pp. 470-475.

https://doi.org/10.1016/j.patrec.2012.12.021

ROC Curve Equivalence using the Kolmogorov-Smirnov Test

Andrew P. Bradley*

The University of Queensland, School of Information Technology and Electrical Engineering, St Lucia, QLD 4072, Australia.

Abstract

This paper describes a simple, non-parametric and generic test of the equivalence of Receiver Operating Characteristic (ROC) curves based on a modified Kolmogorov-Smirnov (KS) test. The test is described in relation to the commonly used techniques such as the Area Under the ROC curve (AUC) and the Neyman-Pearson method. We first review how the KS test is used to test the null hypotheses that the class labels predicted by a classifier are no better than random. We then propose an interval mapping technique that allows us to use two KS tests to test the null hypothesis that two classifiers have ROC curves that are equivalent. We demonstrate that this test discriminates different ROC curves both when one curve dominates another and when the curves cross and so are not discriminated by AUC. The interval mapping technique is then used to demonstrate that, although AUC has its limitations, it can be a model-independent and coherent measure of classifier performance.

Keywords: ROC curves, KS-test, AUC, Specificity, Sensitivity, Coherence,

*Tel.: +61 7 3365 3284; fax: +61 7 3365 4999.

Email address: `bradley@itee.uq.edu.au` (Andrew P. Bradley)

1. Introduction

The Receiver Operating Characteristic (ROC) curve is the graph of a classifier's true positive rate (TPR) against false positive rate (FPR) at various *operating points* as a decision threshold or misclassification cost is varied (Fawcett, 2006; Swets et al., 2000). Over the past fifteen years ROC analysis has become established as an important tool for classifier evaluation (Bradley, 1997). This is especially the case in biomedical applications where TPR and FPR can be directly related to the clinically meaningful measures of *sensitivity* and *specificity*. However, current tests for the equivalence of two or more ROC curves are limited in that they either: require domain specific knowledge, do not work in a wide variety of situations, are based on Normal assumptions, or are computationally expensive. Therefore, this paper proposes a simple, non-parametric and general purpose test of ROC curve equivalence based on a modified Kolmogorov-Smirnov (KS) test.

Receiver operating characteristic curves are traditionally used to answer two questions about classifier performance (Bradley and Longstaff, 2004):

1. Does a classifier have better performance than random labelling?
2. Does one classifier have better performance than another?

There are two common methods to test the *null* hypothesis that the predicted class labels produced by a classifier are no better than random. For a single operating point, all binary classifiers produce results that can be presented in a confusion matrix. A confusion matrix is a form of *contingency table* showing the number of true positive and true negative instances on the leading diagonal and the number of false positive and false negative instances in the off-diagonals. Therefore, a χ^2 test (Press et al., 2007, Section 14.4.1) can be used to test the independence of the true and predicted class labels. We reject the null hypothesis only when there is sufficient evidence that the predicted class labels are dependent on the true class labels. Alternatively, we can utilise information from a number of operating points to test the null hypothesis that the Area Under the ROC curve (AUC) is equal to 0.5 (Bradley, 1997; Bradley and Longstaff, 2004). When estimated empirically, AUC is equivalent to the Wilcoxon-Mann-Whitney test of ranks (Fawcett, 2006). Therefore, an AUC of 0.5 implies that the probability that a classifier will rank (score) a randomly chosen positive instance higher than a randomly chosen negative instance is $P(s_p > s_n) = 0.5$. Here $s_k = m(\mathbf{x})$ is the "score" produced by a classifier for an instance of class $k \in \{p, n\}$ using the feature vector $\mathbf{x}$. Again, we only reject the null hypothesis when there is sufficient evidence that the classifier can correctly rank positive and negative instances. The relationship between ROC curves and the χ^2 test is explored in (Bradley, 1996).

There are typically three ways to test the null hypothesis that two classifiers are equivalent; by comparing:

1. An appropriate measure of classifier performance, such as accuracy or error rate, extracted from the confusion matrix obtained at an individual operating point (Bradley, 1997);
2. The TPR, FPR pair at an individual operating point (Bradley and Longstaff, 2004); or

3. The AUC measured over all, or a sub-set of, operating points on the ROC curve (Bradley, 1997; Landgrebe et al., 2006).

Comparing classifiers based on a single measure of performance can be problematic as the choice of the "best" measure is dependent upon the application domain, class prior probabilities and operating point (Landgrebe et al., 2006). In addition, extracting a single measure from a confusion matrix does not capture the implicit trade-off between positive and negative classifications (Bradley, 1997). Comparing classifiers when both TPR and FPR differ makes it unclear whether the observed differences are due to classifier performance or just different operating points. That is, are these just different operating points on equivalent ROC curves? Comparing TPR or FPR *individually* has the advantage that it effectively implements the Neyman-Pearson method (Bradley, 1997). That is, for a specific FPR, do the classifiers have the same TPR? (or vice-versa). However, again, the FPR or TPR at which to perform the comparison is application dependant. Therefore, because of these issues AUC has gained popularity as a single measure of classifier performance that is extracted from the whole ROC curve. The AUC is independent of prior class probabilities and misclassification costs and has a probabilistic interpretation through its equivalence to the Wilcoxon-Mann-Whitney test of ranks (Fawcett, 2006).

Recently, a number of problems with AUC have been highlighted in the literature. One of the most significant issues is that, as AUC estimates $P(s_p > s_n)$, it's statistical interpretation relies on an implicit alternative (Berrar and Flach, 2012). This probability of correct ranking only has meaning when the evaluation of the classifier is undertaken on a test set consisting of both positive and negative instances. In practice, end-users are primarily concerned with a classifier's performance on a single instance of unknown class. Therefore, error rate or TPR and FPR having meaning; how that instance is ranked against a hypothetical alternative does not (Hilden, 1991). This issue is related to the fact that AUC is estimated from the whole ROC curve and so averages performance over all possible operating points. This is especially problematic when the differences between two ROC curves occur only over a small range of operating points. Classic examples of this problem occur when two different, but crossing, ROC curves have a similar AUC or when an AUC of 0.5 is obtained from a classifier that is clearly not performing random labelling (Hilden, 1991). These issues have recently been described and referred to as the *early retrieval problem* and the *fallacy of the undistributed middle* respectively (Berrar and Flach, 2012). Therefore, unless one classifier *dominates* another over all operating points, AUC will not be a sensitive test of the equivalence of their ROC curves (Drummond and Holte, 2006; Hand, 2009). Here, dominate is taken to mean that one classifier has a higher TPR for all FPR, a condition that appears to occur rarely in practice (Bradley, 1997; Hand, 2009).

It has been argued that it is "fundamentally incoherent" to compare different classifier types using AUC as they effectively use different misclassification costs to generate the ROC curve (Hand, 2009; Hand and Anagnostopoulos, 2012). Again, there is an issue of calculating AUC over the whole curve, using inappropriate misclassification cost ratios ranging from 0 to ∞. The proposed H measure, an extension of that proposed in (Hand, 2005), has two clear advantages: misclassification costs are the same between classifiers and are limited in range. However, from a Neyman-Pearson perspective, an end-user wants to determine whether a specific classifier, at a specified sensitivity or specificity, is better than another (classifier). It is not important to an end-user that in order to get to these operating points one classifier had to use different cost ratios to another. Therefore, in general for two ROC curves to be equivalent there must be no operating points, anywhere on the curve, that have significantly different performance (TPR or FPR). Of course, equivalent ROC curves have an equivalent AUC, but as the issues with crossing ROC curves demonstrate: AUC is a necessary, but not sufficient, condition for ROC equivalence.

A number of alternatives to ROC curves have been developed, including cost curves (Drummond and Holte, 2006), frequency-scaled and expected-utility ROC curves (Hilden, 1991). However, ROC curves are a well-used and well-understood methodology and so we must be careful not to reject them because of issues with their most commonly applied single number summary (AUC) (Hilden, 1991; Berrar and Flach, 2012). Therefore, this paper proposes an improved test of equivalence between two empirical ROC curves.

A number of alternatives to AUC have been proposed, such as the H and *diagnosticity* measures (Hand, 2009; Hilden, 1991) and probability cost PC(+) (Drummond and Holte, 2006). However, these are all designed to be a meaningful *measure* of classifier performance (or utility), rather than a test of ROC equivalence. That is, they are an estimate of how well a classifier will perform, on average, over an appropriate range of misclassification costs and prior probabilities. Note, AUC is a measure of the *ranking* performance of a classifier only (Flach et al., 2011; Berrar and Flach, 2012).

The question of ROC equivalence has previously been tackled by Campbell (1994), Venkatraman and Begg (1996) and Antoch et al. (2010) . However, the first two of these these methods are computationally complex as they involve bootstrap estimates and permutations respectively. The last two do not allow the results of the test to be mapped back to the ROC curves to highlight where the curves differ from each other. Therefore, this paper describes a simple technique, based on on a modified KS test, that finds the corresponding points on two ROC curves that are the most dissimilar. If there is no such point found anywhere on the curve, at the specified level of significance, then the ROC curves are deemed to be statistically equivalent.

The paper is organised as follows: first we discuss the well-known KS test and demonstrate how it can be used to test the null hypothesis that the observed performance of a classifier is no better than random. Next we go on to propose an interval mapping technique whereby two KS tests are used to compare the TPR and FPR of competing classifiers at all operating points. We illustrate the efficacy of this technique with examples where one ROC curve dominates another and where two crossing ROC curves have an equivalent AUC. Finally, the interval mapping technique is used to highlight the conditions under which AUC is a coherent measure of classifier performance.

2. Preliminaries

2.1. ROC Curves

The empirical ROC curve is the plot of $1 - F_n(s)$ versus $1 - F_p(s)$ on a test set of instances with known class membership (Hilden, 1991; Campbell, 1994; Hand, 2009). Here $F_k(s)$ is the cumulative density function (CDF) of the classifier scores $s = m(\mathbf{x})$ for each class $k \in \{n, p\}$. An instance is classified as positive if the given score s is greater than some decision threshold ($s > t$) and negative otherwise. We denote the prior probability of class k in the data set as π_k, where $\pi_n + \pi_p = 1$.

2.2. The KS test

The KS test is defined as (Hand, 2005):

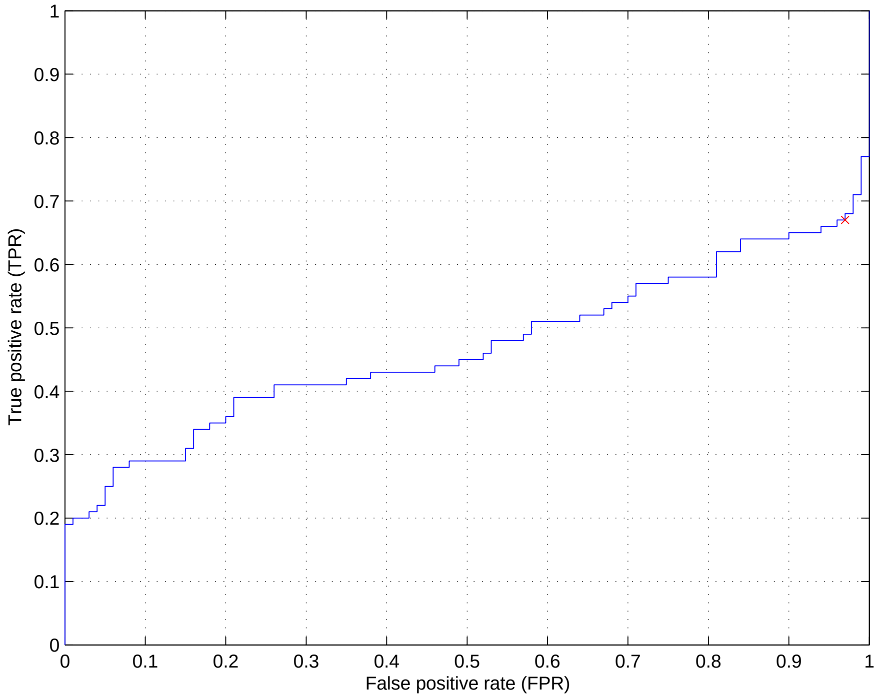
$$D = \max_s |F_n(s) - F_p(s)| \tag{1}$$

The KS statistic, D, can be used to test null Hypothesis that the negative and positive CDFs are equivalent (Press et al., 2007, Section 14.3.3). That is, that the classifier gives, on average, identical scores to instances of both classes. Whilst this behaviour is indicative of a classifier that randomly allocates instances to each class, the KS statistic is not a meaningful measure of classifier performance (Hand, 2005). Specifically, D only relates to the validity of the null hypothesis for that classifier and requires modification before it can be used to compare differences in D between classifiers (Krzanowski and Hand, 2011). The KS statistic does, however, indicate the furthest point on ROC curve from the diagonal (0,0) to (1,1) (Campbell, 1994), which is the expected ROC curve for a classifier that labels instances randomly (Bradley, 1996).

2.2.1. Example

Figure 1 illustrates an example where a ROC curve, with an AUC ≈ 0.5, is obtained from a classifier that scores 100 positive instances with the same mean value as 100 negative instances, but with a larger variance (specifically, $\mathcal{N}(0, 1)$ for the negative class and $\mathcal{N}(0, 4)$ for the positive). This classifier,

Figure 1: Empirical ROC curve showing the operating point of the KS statistic (×).



is unlikely to be performing a random labelling of the test instances, as confirmed by the KS statistic, even though the probability of correct ranking, and hence AUC, is 0.5. This demonstrates the limitation of AUC in this context and that the KS test correctly indicates that the negative and positive distributions differ. Clearly, the KS test and ROC curves are related as they both utilise the class conditional CDFs: one finds the maximum difference between them; the other plots one against the other. However, application of the KS test to the comparison of different classifiers raises two important questions: how do we handle multiple class conditional distributions from multiple classifiers? and how should the scores from the different classifiers be compared?

3. ROC Equivalence using the KS test

Suppose, we have two classifiers, Y and Z, which produce scores $s_Y = m_Y(\mathbf{x})$ and $s_Z = m_Z(\mathbf{x})$ over the intervals $\mathcal{I}_\mathcal{Y} \subseteq \Re$ and $\mathcal{I}_\mathcal{Z} \subseteq \Re$ respectively. Further, suppose these scores have continuous distributions with densities $f(s_Y)$ and $g(s_Z)$ which are zero outside the intervals $\mathcal{I}_\mathcal{Y}$ and $\mathcal{I}_\mathcal{Z}$. Extending the KS statistic to perform a paired comparison between the scores s_Y and s_Z requires that they are mapped to the same interval (Antoch et al., 2010). However, here our intention is to use the KS test to compare the class dependent CDF's produced by the two classifiers. That is, to compare $F_n(s)$ to $G_n(s)$ and $F_p(s)$ to $G_p(s)$, rather than comparing $F_n(s)$ to $F_p(s)$ as in the standard KS test.

Under the null hypothesis of equivalent ROC curves, for any operating point on ROC_Y there exists an identical operating point, with the same TPR and FPR, on ROC_Z. Therefore, any threshold $t_Y \in \mathcal{I}_\mathcal{Y}$ has an equivalent threshold $t_Z \in \mathcal{I}_{\mathcal{Z}}$, i.e.,

$$\forall t_Y \in \mathcal{I}_{\mathcal{Y}} \quad \exists \quad t_Z \in \mathcal{I}_{\mathcal{Z}} \quad \text{where } F_n(t_Y) = G_n(t_Z) \text{ \& } F_p(t_Y) = G_p(t_Z) \tag{2}$$

As the distribution functions are strictly increasing on $\mathcal{I}_{\mathcal{Y}}$ and $\mathcal{I}_{\mathcal{Z}}$, there exists an increasing transformation function $\tau(t)$ that maps $\mathcal{I}_{\mathcal{Z}} \to \mathcal{I}_{\mathcal{Y}}$ (Antoch et al., 2010) such that $F_n(t) = G_n(\tau(t))$ and $F_p(t) = G_p(\tau(t))$, i.e.,

$$\tau(t) = G_n^{-1}\left(F_n(t)\right) = G_p^{-1}\left(F_p(t)\right) \quad \forall t \in \mathcal{I}_{\mathcal{Y}}. \tag{3}$$

Applying this transformation to the mixture distributions for each classifier gives,

$$F(t) = \pi_n F_n(t) + \pi_p F_p(t) = G\left(\tau(t)\right) = \pi_n G_n\left(\tau(t)\right) + \pi_p G_p\left(\tau(t)\right) \tag{4}$$

That is, if the ROC curves are equivalent, application of the transformation $\tau(t)$ will map both classifier's scores to the same interval ($\mathcal{I}_{\mathcal{Y}}$) with identical class conditional and mixture distributions. Note, (4) assumes the case of a *paired* comparison, that is different classifiers evaluated on the same test set (as implied in the definition of the scores s_Y and s_Z). Indeed, (Berrar and Flach, 2012) have cautioned against comparing ROC curves when the classifiers were *not* trained and tested on the same (paired) data. Importantly, there is no requirement that equivalent ROC curves behave in exactly the same manner, only that they agree on the same proportion of negative and positive instances (Antoch et al., 2010).

In practice the transformation $\tau(t)$ is estimated from a set of data. That is, from the *empirical* mixture distribution

$$\hat{\tau}(t) = \hat{G}^{-1}\left(\hat{F}(t)\right) \quad \forall t \in \mathcal{I}_{\mathcal{Y}}. \tag{5}$$

This transformation can then be used to map $\mathcal{I}_{\mathcal{Z}} \to \mathcal{I}_{\mathcal{Y}}$ enabling the scores from both classifiers to be directly compared.

$$s_{ZY} = \hat{\tau}(s_Z) \tag{6}$$

The transformed scores (s_{ZY}) have the same value and rank order as s_Y, but potentially different class labels, as the scores come from different classifiers. In this way, the classifiers are given identical mixture distributions, regardless of the validity of the null hypothesis and the class conditional distributions are only identical when the ROC curves are equivalent (when $m_Y(\mathbf{x}) \equiv m_Z(\mathbf{x})$). Put another way, as the (monotonic) transformation, $\hat{\tau}(t)$, preserves rank order $s_Z \to s_{ZY}$ it does not alter classifier Z's ROC curve or AUC (Campbell, 1994); it simply maps the scores from both classifiers to the same interval.

The test for ROC equivalence then consists of two independent KS tests,

$$D_n = \max_{s_Y} |F_n(s_Y) - G_n(s_{ZY})| \tag{7}$$

$$D_p = \max_{s_Y} |F_p(s_Y) - G_p(s_{ZY})| \tag{8}$$

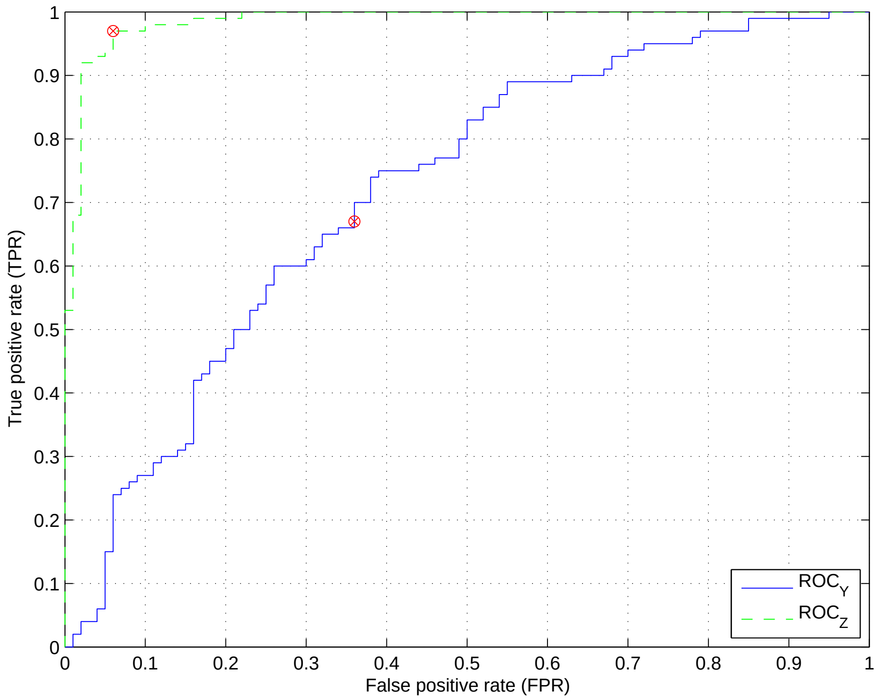
The KS statistics D_n and D_p indicate the maximum distances between the two classifier's negative and positive CDFs respectively. These can then be used to calculate the p-value of the observed D_n and D_p and hence accept or reject the null hypothesis that the distributions (and hence ROC curves) are the same (Press et al., 2007, Section 14.3.3). The advantage of having two KS tests applied independently to the negative and positive CDFs is that the critical values of D_n and D_p are based on the number of instances in each class. For example, in the case of skewed class priors, the class conditional distributions will be estimated from significantly different numbers of instances. Therefore, for a given value of D, the class with the larger number of instances will have a lower p-value. Of course, as the null hypothesis now involves two comparisons, a Bonferroni correction (or similar) should be applied to maintain the type I error rate. That is, each individual hypothesis should be tested at the $\alpha/2$ level of significance.

3.1. Examples

Figure 2 demonstrates empirical ROC curves from two classifiers Y and Z, where Z dominates Y. Clearly, comparing the performance of these classifiers at any individual operating point, using error rate or the (TPR, FPR) pair, or over a number of operating points using AUC, will indicate the superiority of classifier Z. In this example, the scores from classifier Y are $\mathcal{N}(0,1)$ for the negative class and $\mathcal{N}(1,1)$ for the positive. For classifier Z the distributions are unchanged for the negative class and $\mathcal{N}(3,1)$ for the positive. In both cases there are 100 instances in each class.

Figure 3 shows the cumulative density functions for the negative class (top) and positive class (bottom) for classifier scores s_Y, s_Z and s_{ZY}. For the negative class it shows that originally $F_n(s_Y)$ and $G_n(s_Z)$ are similar, but for the positive class $F_p(s_Y) > G_p(s_Z)$ resulting in an improved TPR and FPR at all operating points (score thresholds). The superiority of classifier Z is maintained after $\mathcal{I_Z} \rightarrow \mathcal{I_Y}$ as it can be seen that $F_n(s_Y) < G_p(s_{ZY})$ and $F_p(s_Y) > G_p(s_{ZY})$ at virtually all operating points (as of course $ROC_{ZY} \equiv ROC_Z$). In this case, both D_n and D_p occurred at the same operating point (score ≈ 0.7) and so there is one operating point where classifier Z is maximally different to Y in both TPR and FPR. We can can therefore reject the null hypothesis that ROC_Y and ROC_Z are equivalent

Figure 2: Empirical ROC curves where classifier Z dominates Y, showing the operating points related to the KS statistics D_n ($\circ$) and D_p ($\times$).



at the $p = 0.05$ level of significance.

Figure 3: Class conditional CDFs for classifiers Y (s_Y) and Z (s_Z); and for Z mapped to the same interval as Y (s_{ZY}).

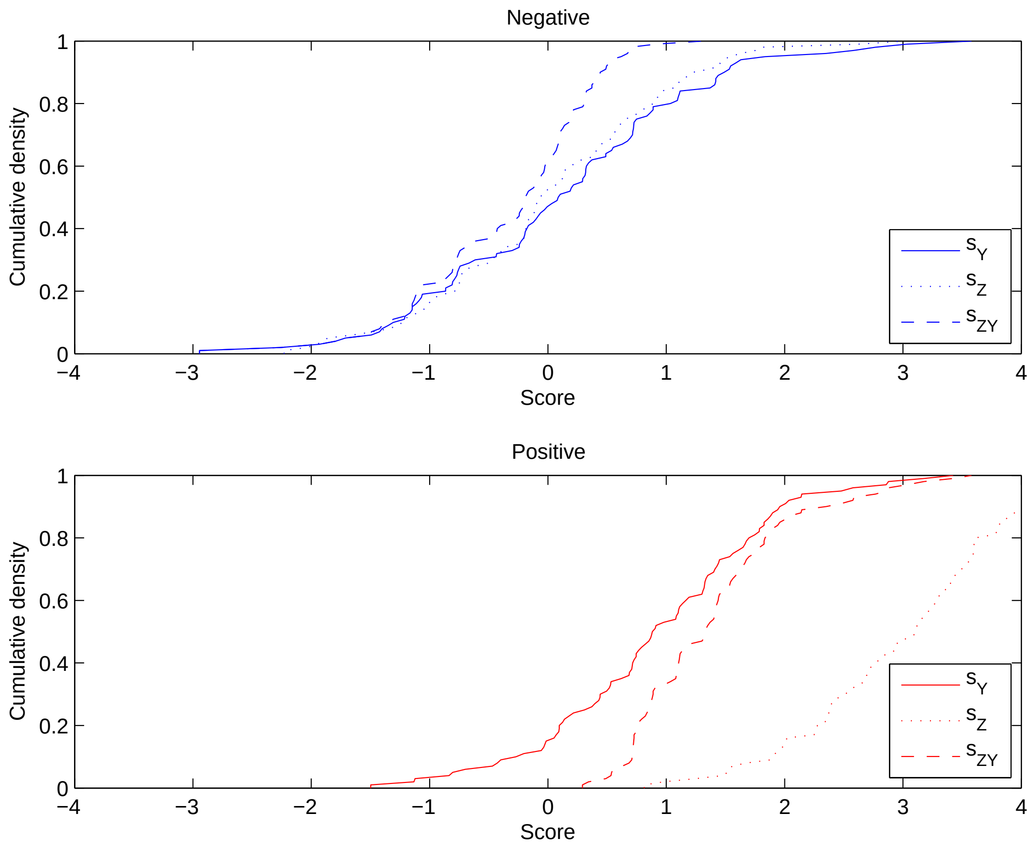


Figure 4 demonstrates empirical ROC curves from two classifiers Y and Z that not only cross, but have the same AUC (0.78). In this example, the scores from classifier Y are $\mathcal{N}(0, 1)$ for the negative class and $\mathcal{N}(1, \frac{1}{3})$ for the positive. For classifier Z the distributions are swapped and negated so that they are $\mathcal{N}(-1, \frac{1}{3})$ for the negative class and $\mathcal{N}(0, 1)$ for the positive. This results in the classifiers having the same minimum (Bayes) error rate, with TPR_Y = 1 - FPR_Z and FPR_Y = 1 - TPR_Z. In both cases there are 140 instances in each class.

Figure 4: Crossing ROC curves for classifiers Y and Z showing the operating points related to the KS statistics D_n ($\circ$) and D_p ($\times$).

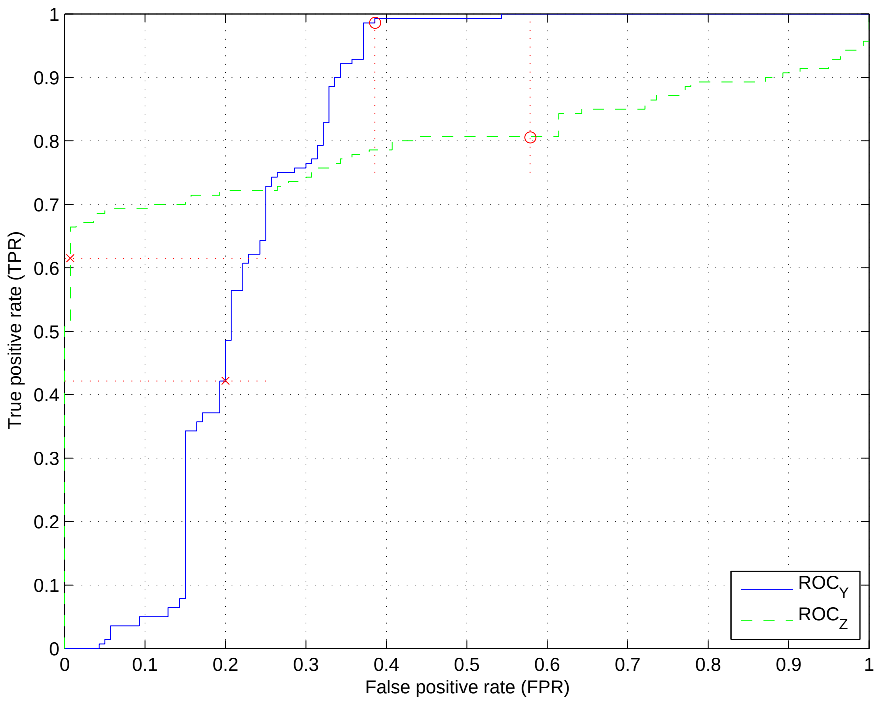


Figure 4 shows that we can can reject the null hypothesis that ROC_Y and ROC_Z are equivalent at the $p = 0.05$ level of significance. The maximum difference in TPR (D_p) occurs between the operating points (0.007, 0.615) and (0.2, 0.422). The maximum difference in FPR (D_n) between (0.386, 0.986) and (0.579, 0.805). While these difference occur at the same score for both classifiers, there is no constraint that they occur at the same TPR or FPR, as in the Neyman-Pearson method. To determine if classifier Y performs better than Z depends on whether the application domain requires that we operate at a high TPR (where Y is likely to be preferred) or low FPR (where Z is likely to be preferred).

Figure 5: Empirical ROC curves for three classifiers X, Y and Z showing the operating points related to the KS statistics D_n ($\circ$) and D_p ($\times$) where Y most differs from Z.

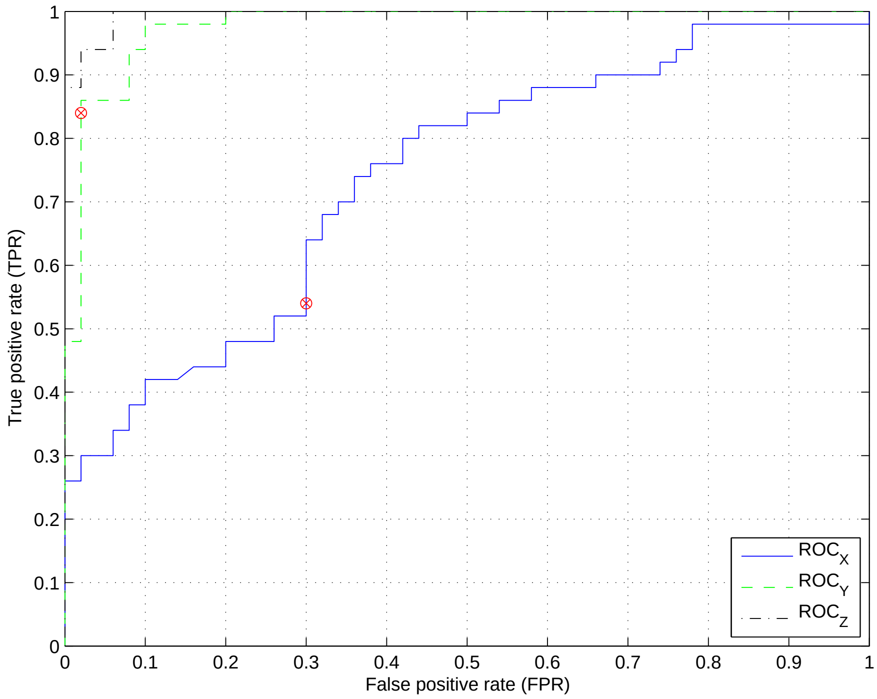


Figure 5 demonstrates empirical ROC curves from three classifiers X, Y and Z, where Y and Z are equivalent, but both dominate X. In this example, the scores from classifiers X, Y and Z are estimated by merging the posterior probabilities obtained using 10-fold cross validation (Fawcett, 2006; Bradley, 1997). The classifiers are all of the same type (quadratic discriminant functions), but are trained using different feature sub-sets. Specifically, a two-class (Versicolor, Virginica) version of Fisher's Iris dataset is used where the species is predicted: by classifier X using two features only (sepal length and width); by classifier Y using three features (previous two plus petal length) and by classifier Z using all four features (previous three plus petal width). For simplicity Figure 5 only shows the operating points where classifiers Y and Z differ the most. There are no operating points where X and Y differ significantly and so on the available data (50 instances per class) they are deemed equivalent.

4. Discussion

The examples presented in this paper demonstrate that, once the scores from different classifiers are mapped to the same interval, the KS statistic can be used to test the null hypothesis that their ROC curves are equivalent. The proposed test consists of measuring the maximum difference between both the positive and negative CDFs when mapped to the same interval. The advantage of the method is that the threshold at which this maximum difference occurs relates to a specific TPR and/or FPR and therefore to specific operating points on both ROC curves. Therefore, if the null hypothesis can be rejected the operating points that differ the most in terms of TPR and FPR can be displayed.

It is of interest here to note the difference between (5) and the method proposed by Antoch et al. (2010) which tests the null Hypothesis that the transformations applied to the negative and positive distributions are equal, i.e.,

$$\tau_n(t) = \tau_p(t) \quad \forall t \in \mathcal{I}_{\mathcal{Y}}. \tag{9}$$

This requires the development of a bespoke test statistic and, if the null hypothesis is rejected, does not indicate where on the ROC curves the classifiers differ. Also, the modification to the KS test presented here differs from that described in (Campbell, 1994) in that initially a conventional KS test is used to created confidence intervals on a single ROC curve. Then the KS test is applied to the maximum distance between two ROC curves along a line with slope $b = -\sqrt{\pi_n/\pi_p}$, using a bootstrap technique to estimate the p-value. This joint confidence interval was shown to be "too loose" by Macskassy and Provost (2004).

It has been argued that displaying ROC curves with confidence intervals is more meaningful that p-values (Berrar and Flach, 2012). However, when there are multiple ROC curves to compare, p-values are of use for automatically detecting equivalent ROC curves; thereby reducing the number (unique) ROC curves to compare in detail. Again, having a hypothesis test that can indicate on the ROC curve which operating points are significantly different can guide this detailed (and application dependent) comparison.

Hand (2009) showed that using AUC to compare classifiers is equivalent to taking an average of the losses at different thresholds, using the mixture distribution as a weighting function. He then went on to argue that the implication of this, is that AUC is "fundamentally incoherent" as it depends on the classifier's score distribution (effectively $F(t)$ and $G(t)$) and so the weight distribution used to combine different cost ratios varies from classifier to classifier. However, (4) demonstrates that by applying the transformation, $\tau(t)$, the scores from any two classifiers can always be given identical mixture distributions. In addition, when the ROC curves are equivalent, this transformation also ensures that the scores have identical class conditional distributions. Therefore, for equivalent ROC curves, after the application of the transform the weight distributions become equal and AUC is coherent. When two ROC curves are *not* equivalent, the transformation produces identical mixture distributions, but different class conditionals. In this case, an additional constraint is required, as per the Neyman-Pearson method, so that the classifiers are compared at the same sensitivity or specificity (Hand and Anagnostopoulos, 2012).

It is well known that ROC curves (and AUC) are invariant to *any* monotonic transformation, as rank order is preserved (Campbell, 1994). This is also the implication of the equivalence between AUC and the Wilcoxon-Mann-Whitney test of ranks. Therefore, provided AUC is estimated independently of the costs, it is always coherent. Specifically, as Flach et al. (2011) show, AUC is coherent when estimated using both optimal and non-optimal thresholds. While this is the implicit choice for calculating AUC (using as many thresholds as there are test instances) it is often not realistic. For example, Figure 4 shows the "incoherent" example of two very different ROC curves producing identical AUCs. While they both have the same overall probability of correct ranking, this probability does not distinguish a classifier with a high sensitivity (Y) from one with a high specificity (Z).

Future work could apply extensions of the KS test, such as the Anderson-Darling statistic, that have been shown to be more sensitive in the tails of this distributions (Press et al., 2007, Section 14.3.4). This may be important to increase the sensitivity of the proposed ROC equivalence test, as the tails of the distributions are likely to be where practically important differences between different classifiers can be found, e.g., when TPR ≥ 0.9. It may also be beneficial to in indicate on the ROC curves all values of D_n and D_p that exceed the critical value, so that an end-user can see if the ROC curves differ at an operating point of practical significance.

5. Conclusions

This paper has presented a straight-forward extension of the KS test that allows two competing ROC curves to be compared for equivalence. If the curves are found to be not equivalent the method indicates the operating points where the two ROC curves are most dissimilar in both TPR and FPR. The proposed KS test was shown to correctly handle cases where the ROC curves can be distinguished based on AUC, but also the confounding case of where two different and crossing ROC curves have the same AUC. Therefore, the test is a useful addition to the classifier evaluation toolbox.

6. Acknowledgements

I would like to thank the anonymous reviewers for their constructive comments on an earlier draft of this paper. The author is the recipient of an Australian Research Council Future Fellowship (FT110100623).

Antoch, J., Prchal, L., Sarda, P., 2010. Nonparametric comparison of ROC curves: Testing equivalence. Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis 7, 12–24.

Berrar, D., Flach, P., 2012. Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). Briefings in Bioinformatics 13 (1), 83–97.

Bradley, A. P., 1996. ROC curves and the X2 test. Pattern Recognition Letters 17 (3), 287–294.

Bradley, A. P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30 (7), 1145–1159.

Bradley, A. P., Longstaff, I., 2004. Sample size estimation using the receiver operating characteristic curve. In: Proceedings 17th International Conference on Pattern Recognition. Vol. 4. pp. 428–431.

Campbell, G., 1994. Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. Statistics in Medicine 13 (5-7), 499–508.

Drummond, C., Holte, R. C., 2006. Cost curves: An improved method of visualising classifier performance. Machine Learning 65, 95–130.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters 27 (8), 861–874.

Flach, P., Hernandez-Orallo, J., Ferri, C., June 2011. A coherent interpretation of auc as a measure of aggregated classification performance. In: Getoor, L., Scheffer, T. (Eds.), Proceedings of the 28th International Conference on Machine Learning (ICML-11). ICML '11. ACM, New York, NY, USA, pp. 657–664.

Hand, D. J., 2005. Good practice in retail credit scorecard assessment. Journal of the Operational Research Society 56, 1109–1117.

Hand, D. J., 2009. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine Learning 77, 103–123.

Hand, D. J., Anagnostopoulos, C., 2012. When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? Pattern Recognition Letters preprint.

Hilden, J., 1991. The area under the ROC curve and its competitors. Medical Decision Making 11 (2), 95–101.

Krzanowski, W. J., Hand, D. J., 2011. Testing the difference between two kolmogorovsmirnov values in the context of receiver operating characteristic curves. Journal of Applied Statistics 38 (3), 437–450.

Landgrebe, T. C., Paclik, P., Duin, R. P., Bradley, A. P., 2006. Precision-recall operating characteristic (P-ROC) curves in imprecise environments. In: Proceedings 18th International Conference on Pattern Recognition. Vol. 4. pp. 123–127.

Macskassy, S., Provost, F., 2004. Confidence bands for ROC curves: Methods and an empirical study. In: Proceedings of the First Workshop on ROC Analysis in AI.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P., 2007. Numerical Recipes: The Art of Scientific Computing, 3rd Edition. Cambridge University Press.

Swets, J. A., Dawes, R. M., Monahan, J., 2000. Better decisions through science. Scientific American, 82–87.

Venkatraman, E. S., Begg, C. B., 1996. A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. Biometrika 83 (4), 835–848.