FBE-CMPE-06/2009-03

STATISTICAL COMPARISON OF CLASSIFIERS USING AREA UNDER THE ROC CURVE

ÖZLEM ASLAN [1]
OLCAY TANER YILDIZ [2]
ETHEM ALPAYDIN [1]

Kasım 2009

Fen Bilimleri Enstitüsü

November 2009

Institute for Graduate Studies
in Science and Engineering

Boğaziçi University, Bebek, Istanbul, Turkey

[1] Computer Eng. Dept., Boğaziçi University, Istanbul, Turkey
[2] Computer Eng. Dept., Işık University, Istanbul, Turkey

Statistical Comparison of Classifiers Using Area Under the ROC Curve

Özlem Aslan, Olcay Taner Yıldız and Ethem Alpaydın

Abstract

Statistical tests in the literature mainly use error rate for comparison. Receiver Operating Characteristics (ROC) curves and/or Area Under the ROC Curve (AUC) can also be used for comparing classifier performances under a spectrum of loss values. A ROC curve and hence an AUC value is calculated from one training/test pair and to average over randomness in folds, we propose to use k-fold cross-validation to generate a set of ROC curves and AUC values to which we can fit a distribution and test hypotheses on. Experiment results on 15 datasets using 5 different classification algorithms show that our proposed test using AUC values is to be preferred over the usual paired t test on misclassification errors because it can detect equivalences and differences which the error test cannot.

I. Introduction

Comparing the performances of classifiers is a critical problem in machine learning. In the literature, to compare the generalization error of learning algorithms, statistical tests have been proposed [1], [2]. In choosing between two learning algorithms, one can use a pairwise test to compare their generalization error and select the one that has lower error. Typically, cross-validation is used to generate a set of training, validation folds, and we compare the expected error on the validation folds after training on the training folds. Examples of such tests are parametric tests, such as k-fold paired t test, 5×2 cv t test [1], 5×2 cv F test [3], nonparametric tests, such as the sign test and Friedman's test, or range tests, such as Wilcoxon signed rank test.

AUC has been related to the *Wilcoxon* statistic. Wilcoxon statistic has been defined as an estimate of 'true' area under the ROC curve, area constructed from an infinite sample [4]. In signal detection, people have been using ROC curves to visualize the trade-off between hit rate and false alarm rate [5]. Area under the ROC curve (AUC) is also used for comparing classifiers. However, ROC and AUC use a single training and testing pair [6], [7], [8]. In this paper, we extend this idea and use k-fold cross-validation to generate k ROC curves and hence k AUC values and then we fit a distribution to the set of AUC values and test hypothesis on these distributions.

In Section II, performance metrics for comparing classifiers, ROC curves and AUC are discussed. In Section III, our AUC based statistical test is explained. In Section IV, the experimental setup and results of the experiments are given. In Section V, we give the related work on ROC curve and AUC. We conclude in Section VI.

II. Statistical Methods for Comparing Classification Algorithms

A. Performance Metrics for Classifiers

If we define class labels of the two-class classification problem as positive and negative, the confusion matrix that is shown in Table I contains the following items:

- True positive (TP): If both the class label and the predicted class are positive.
- False negative (FN): If the class label is positive and the predicted class is negative.
- False positive (FP): If the class label is negative and the predicted class is positive.
- True negative (TN): If both the class label and the predicted class are negative.

Different metrics calculated from these values are used in the literature:

$$\begin{aligned} \text{hit rate} &= \frac{TP}{TP+FN} \\ \text{false alarm rate} &= \frac{FP}{TN+FP} \\ \text{error} &= \frac{FP+FN}{TP+TN+FP+FN} \end{aligned}$$

Özlem Aslan and Ethem Alpaydın are with the Department of Computer Engineering, Boğaziçi University, TR-34342, Istanbul, Turkey. Email: ozlem.aslan1@boun.edu.tr, alpaydin@boun.edu.tr

Olcay Taner Yıldız is with the Department of Computer Engineering, Işık University, TR-34980, İstanbul, Turkey. Email: olcaytaner@isikun.edu.tr

TABLE I
CONFUSION MATRIX

	Predicted Class	
True Class	Positive	Negative
Positive	TP	FN
Negative	FP	TN

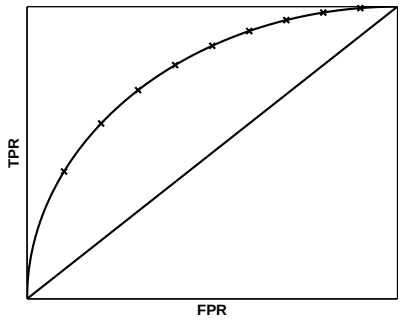


Fig. 1. Example ROC Curve: different treshold points are marked on the upper ROC curve. Lower diagonal line is the ROC curve for random prediction.

B. ROC curves

ROC curves are being used in signal processing to plot the trade-off between hit rate and false alarm rates. They allow visualization of performance for a set of conditions instead of just the misclassification error. Hit rate or true positive rate (TPR) defines the y axis and false alarm rate or false positive rate (FPR) defines the x axis. A classifier is good if it has a high hit rate and a low false alarm rate, that is, if the curve is closer to the upper left corner. The diagonal line indicates the curve for random prediction (see Figure 1)

In a two-class problem, if the posterior probability of the positive class is greater than the posterior probability of the negative class, the classifier predicts the class label of a test instance as positive, otherwise it predicts the class label as negative. This is equivalent to checking if the posterior probability of the positive class is greater than the *threshold value* of $\theta = 0.5$. For a given test set, a ROC curve is constructed by plotting the hit rates on y axis and false alarm rates in x axis for different threshold values. The ROC curve construction algorithm is given in [5].

Let us define the positive class as class C_1 and the negative class as C_2. Given the loss matrix in Table II, for a test instance x, the risk of choosing C_1 is:

$$\begin{aligned} R(C_1|x) &= \lambda_{11}P(C_1|x) + \lambda_{12}P(C_2|x) \\ &= \lambda P(C_2|x) \end{aligned}$$

and the risk of choosing C_2 is:

$$R(C_2|x) = P(C_1|x)$$

Then we choose C_1 if

$$\begin{aligned} R(C_1|x) &< R(C_2|x) \\ \lambda P(C_2|x) &< P(C_1|x) \end{aligned}$$

that is, if

$$\frac{P(C_1|x)}{P(C_2|x)} > \lambda$$

Since $P(C_1|x) + P(C_2|x) = 1$, this gives

$$\begin{aligned} \frac{P(C_1|x)}{1 - P(C_1|x)} &> \lambda \\ P(C_1|x) &> \frac{\lambda}{1+\lambda} \end{aligned}$$

TABLE II
LOSS MATRIX

	C_1	C_2
C_1	0	λ
C_2	1	0

We see that the threshold of 0.5 which we use to calculate misclassification error corresponds to $\lambda = 1$, that is, when a false positive and a false negative has equal loss. We get a variety of thresholds when we vary λ:

$$\theta = \frac{\lambda}{1+\lambda}$$
$$\begin{aligned} \lambda = 0.5 &\rightarrow \theta = 1/3 \\ \lambda = 1 &\rightarrow \theta = 1/2 \\ \lambda = 2 &\rightarrow \theta = 2/3 \\ \lambda = 10 &\rightarrow \theta = 10/11 \end{aligned}$$

That is, the threshold points on the ROC curve indicate the λ values in the risk calculation. This is the reason why using ROC curve (or AUC value, as we will see shortly) is better than using misclassification error because error gives equal emphasis and makes no distinction between false positives and false negatives and thus may not be the best measure for many applications; ROC curve (and AUC) takes a set of possible loss proportions into account and hence average over that defines a more robust measure.

If the ROC curve of the first classifier is always over the ROC curve of the second classifier, we can easily say that the first classifier is better than the second classifier. But this case does not always happen. In some cases, the ROC curve of the first classifier may be over the ROC curve of the second classifier in one part, whereas the second classifier's curve is over the ROC curve of the first one in some other part; this implies that the two classifiers are preferred under different loss conditions. ROC is a curve; one may reduce the ROC curve to a single value using the *area under ROC curve* (*AUC*). If a ROC curve is closer to the upper left corner, the area gets closer to 1. The area under the ROC curve is estimated by summing tropezoidal areas formed by successive points on the ROC curve. A classifier with a greater AUC is said to be better than a classifier with a smaller AUC. AUC calculation algorithm is given in [5].

III. PROPOSED TECHNIQUE

In general, a classifier is trained using a training set and the ROC curve is constructed and AUC is calculated only once using a test set. To average over randomness in the training and testing split, one can use more than one training and testing pair, which results in multiple ROC curves and AUC values. The main idea of this paper is to fit a distribution to these values and test hypotheses on such distributions.

We use k-fold cross-validation to generate k training sets and train k classifiers whose ROC curves and AUC values we calculate over a test set. At the end, for each classification algorithm we have k AUC values. To have a *paired* test, we use the same training and test sets for all algorithms. Afterwards, the two classification algorithms can be compared by applying the paired t test with the null hypothesis that two classifiers have the same mean AUC values and the alternative hypothesis that the two AUC means are different.

k-fold cross-validated paired t test will be called as *AUC test* in the rest of the paper. In AUC test, there are K train-test set pairs because of applying k-fold cross-validation. Each classification algorithm is trained on training set T_i and posterior probabilities are calculated by testing on test set Te_i , $i = 1, \ldots, K$. Using these posterior probabilities, the AUC of each classifier is calculated as A_i^1 and A_i^2. AUC difference is calculated in each fold as $A_i = A_i^1 - A_i^2$ for $i = 1, \ldots, K$. The distribution of difference is normal since the A_i^1 and A_i^2 distributions are approximately normal (we extend the normality assumption for errors to AUC values). Then, if the mean of this distribution equals to zero, we can say that classifiers have equal AUC's:

$$H_0 : \mu = 0 \tag{1}$$
$$H_1 : \mu \neq 0 \tag{2}$$

Then $m = \sum_{i=1}^{K} \frac{A_i}{K}$, $S^2 = \sum_{i=1}^{K} \frac{(A_i - m)^2}{K-1}$. t statistic is calculated as $\frac{\sqrt{K}.m}{S} \sim t_{K-1}$. The null hypothesis is accepted at significance level α, if the test statistic is in the interval $(-t_{\alpha/2,K-1}, t_{\alpha/2,K-1})$.

IV. EXPERIMENTS

A. Experimental Setup

1) Data sets: We use a total of 15 data sets where 11 (*aibocolor, chess, connect-4, mushroom, nursery, pageblock, report, shuttle, spambase, thyroid, wave*) are from UCI and 4 (*ada, caravan, gina and sylva*) are from IJCNN 2007 [9]. The datasets

with the number of instances greater than 3000 or approximately 3000 are selected to decrease the dependency between folds of 30-fold cross validation. Since two-class classification is applied, the datasets with more classes are converted to two classes—this is done by selecting the two classes which are most confused by looking at the confusion matrix (We first use 1-nearest neighbor over all classes to choose these two).

2) Learning algorithms: We use five algorithms:

1) *C4.5*: C4.5 decision tree algorithm [10].
2) *LP*: Linear perceptron with softmax outputs trained by gradient-descent to minimize cross-entropy
3) k-*NN*: k-nearest neighbor. For the optimization of k, values of 1, 3, 5, 7, 11, 21 are tried and the one with minimum validation error is selected.
4) *NB*: naive nayes that is a parametric discriminator assuming independent inputs.
5) *Ripper*: Rule learning algorithm with 2 optimization steps [11].

3) Division of training, validation, and test sets: Our methodology is as follows: A data set is first divided into two parts, with $1/3$ as the test set, *test*, and $2/3$ as the training set, *train-all*. The training set, *train-all*, is then resampled using 30 times cross-validation to generate *tra*$_i, i = 1, \ldots, 30$, which are used to train the classifiers and the tests are run on the test set.

B. Overall Results

We compare 5 algorithms in a pairwise manner on 15 datasets using the paired t test on misclassification errors or AUC values at the significance level of 0.05, which makes a total of 150 comparisons. The null hypothesis of both k-fold cv paired t test on errors (error test) and k-fold cv paired t test on AUC values (AUC test) are that the two populations have the same mean. There are four possible cases:

- Both error test and AUC test accept the null hypothesis. This case occured only 1 time.
- Error test accepts and AUC test rejects the null hypothesis. This case occured 10 times.
- Error test rejects and AUC test accepts the null hypothesis. This case occured 9 times.
- Both error test and AUC test reject the null hypothesis. This case occured 130 times.

We now discuss some examples of these cases: Figure 2 shows the results on *chess* dataset for *C4.5* and *Ripper* algorithms, where both the error test and our AUC test accepts the null hypothesis. It can be seen in (a) that the error distribution of the two algorithms overlap and in (b) that the AUC distribution of the two algorithms also overlap. ROC graph supports the agreement, since ROC curves of algorithms overlap (c) and 0.5 threshold points shown on the ROC curves (by circle and triangle for the two algorithms) also overlap.

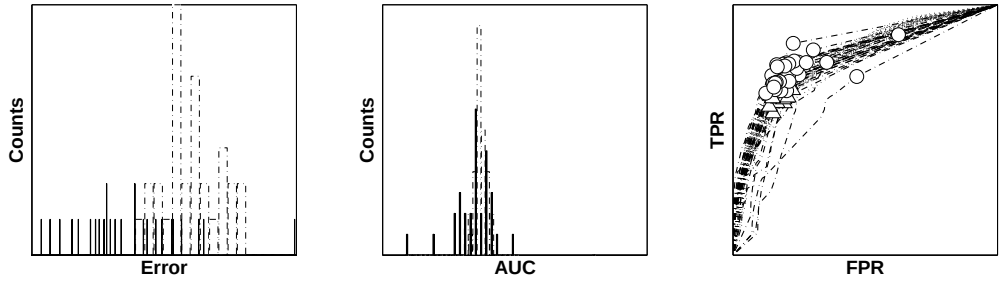
Figure 3 shows the second case where the error test accepts and our AUC test rejects the null hypothesis. In Figure 3(a), it can be seen that the error distributions of the k-*NN* (white) and *Ripper* (black) on the *report* dataset overlap and this supports the decision of the error test. In Figure 3(b), it can be seen that AUC distributions are significantly separated. In Figure 3(c), we see why; it can be seen that ROC curves of k-*NN* (white) are above the ROC curves of the *Ripper* (black). For large values of θ; this implies that the two algorithms have different performances in such cases. The marked points (decisions at the threshold of 0.5) overlap and this supports the error test decision but if we look overall, we see that the algorithms have indeed different behavior over all possible thresholds. We see that the AUC test is able to detect differences that the error test cannot and that is why, we can say that the AUC test has higher power.

Figure 4 shows the third case where error test rejects and our AUC test accepts the null hypothesis that the algorithms have equal expected performance. If we look at Figure 4(a), we see that there is a significant difference in error distributions of k-*NN* (white) and *NB* (black) on the *shuttle* dataset. Looking at Figure 4(b), it can be seen that there is not a significant difference in AUC distributions. In Figure 4(c), the ROC curves intersect. To the left of the intersection, *NB* (black) is better and to the right, k-*NN* (white) is better. Though, the error test says that they are different, if we average over all possible losses (as AUC does), we see that there is no significant difference. The AUC test does not reject such cases and can therefore be said to have lower type I error.

Figure 5 is an example of the fourth case where both the error test and our AUC test reject the null hypothesis. In Figure 5(a) and 5(b), the error and area distributions of *C4.5* (white) and *LP* (black) on *nursery* dataset are well-separated. Figure 5(c) also supports this claim, ROC curves of *LP* (black) are over the ROC curves of *C4.5* (white) and the threshold marks are also quite well-separated.

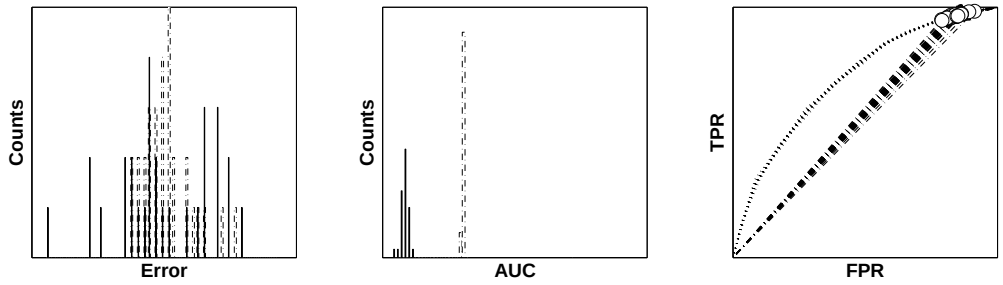
V. Related Work

Hanley and McNeil [4] stated that Wilcoxon statistic is an estimate of 'true' area under the ROC curve, the area constructed from an infinite sample. They have also given a standard error formula which takes five parameters: the probability that two randomly chosen abnormal images will both be ranked higher than a randomly chosen normal image, the probability that one randomly chosen positive example will be ranked higher than two randomly chosen negative examples, the number of positive examples, the number of negative examples and the estimated area under the ROC curve. However, for calculating the standard error of the estimated AUC, the distributions of the positive and negative examples should also be known. Using



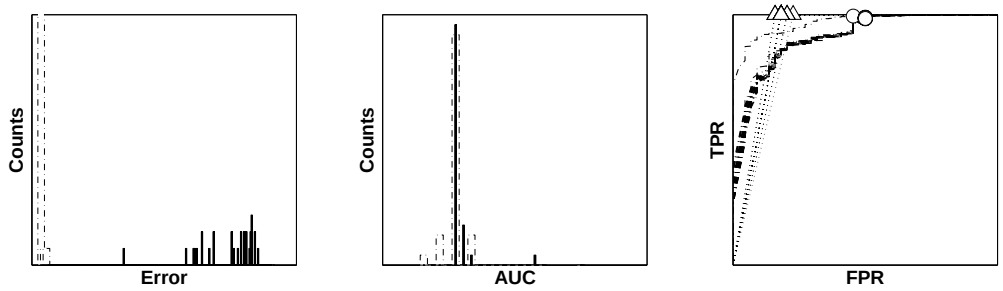
(a) ERROR DISTRIBUTION (b) AUC DISTRIBUTION (c) ROC CURVES

Fig. 2. An example for the case 1, where both error test and our AUC test accepts the null hypothesis



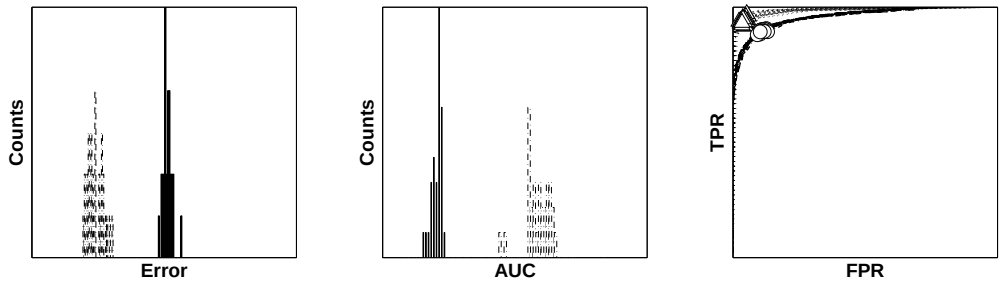
(a) ERROR DISTRIBUTION (b) AUC DISTRIBUTION (c) ROC CURVES

Fig. 3. An example for the case 2, where error test accepts and our AUC test rejects the null hypothesis



(a) ERROR DISTRIBUTION (b) AUC DISTRIBUTION (c) ROC CURVES

Fig. 4. An example for the case 3, where error test rejects and our proposed AUC test accepts the null hypothesis



(a) ERROR DISTRIBUTION (b) AUC DISTRIBUTION (c) AUC DISTRIBUTION

Fig. 5. An example for the case 4, where both error test and our proposed AUC test rejects the null hypothesis

the probabilities defined in the calculation of standard error, they have also given a formula that finds the required number of positive and negative examples for detecting the difference of two AUC's depending on the specified type I and type II error rates (It also requires specific distributions for the values of positive and negative samples).

Hanley and McNeil [12] have argued that comparing different ROC curves with a single dataset limits their usefulness. They state that there is a correlation between AUC's calculated from the same dataset, where correlation is included in the calculation of the standard error of difference in AUC's. They have noticed that a paired test can be used for comparing two

algorithms and therefore included the correlation in the statistical test for applying the behaviour of paired t test. A z test statistic is constructed using this standard error and the null hypothesis that 'true' AUC's are equal. They state that they make a correction for pairing like t test. However, we directly use the paired t test, by applying cross-validation to dataset. Therefore, their motivation supports our work. Paired t test is applied to AUC results, but it is not compared with an error test. It is only used for evaluating the results [13].

Cortes and Mohri [14] have also proposed to calculate confidence intervals for AUC. A confidence interval for AUC has been derived from the confidence interval of error. First, they define expectation and variance of AUC in terms of the expected error, the number of negative instances and the number of positive instances by using the Wilcoxon-Mann-Whitney statistic. Using these values, the confidence intervals are constructed without any assumption on the distribution for AUC. For large values of the sample size, they make a normal distribution assumption for error.

We argue that there are two weaknesses in their work. First, using error for deriving a confidence interval for AUC is not a good idea, because as we show below, in some cases, AUC intervals can be significantly different although the error intervals are not significantly different. However, their confidence interval formulations give the same AUC interval for the same error value. For comparing our results with their results, we trained and tested the classification algorithms without cross-validation and substituted the error results in their formulations since they use one error value.

In Figure 6 (a), the error distributions of the classifiers *Ripper* and *LP* on dataset *ada* are shown, they overlap indicating the equality of their means and the error test can not reject the null hypothesis that the means of these error distributions are equal. However, in Figure 6(b), it can be seen that the corresponding AUC distributions are separated despite the overlapping of error distributions and our AUC test rejects the null hypothesis that the means of these AUC distributions are equal. The dashed-dotted lines above the distributions in Figure 6(b) are the AUC confidence intervals found by the method of Cortes and Mohri [14]. Their confidence intervals do not show a good fit to the empirical AUC distributions since AUC confidence intervals can be significantly different although error confidence intervals are not. Their confidence intervals also fail when the error results are different. In Figure 7(a), the error distributions of the classifiers k-*NN* and *LP* on dataset *ada* are shown, they do not overlap and the error test can not reject the null hypothesis that the means of these AUC distributions are equal. In Figure 7(b), the AUC distributions do not overlap and our AUC test can not reject the null hypothesis that the means of these AUC distributions are equal. The confidence intervals of Cortes and Mohri do not fit to the distributions. Another point to note is that, as seen in Figures 6(b) and 7(b), their confidence intervals are too large because their approach is nonparametric. However, they are inefficient when the sufficient conditions for the distribution assumptions are met.

Another approach for finding the confidence intervals for AUC has been proposed in [15]. Agarwal et al. give a large deviation bound for the distribution independent case. In Figures 6(b) and 7(b), their confidence intervals are shown with dotted lines above the AUC distributions. The figures support their claim that confidence intervals are too large since no distribution assumption is made. They state that the AUC value follows an asymptotically normal distribution and for large N, the normal approximation can be used to obtain a tighter bound (as we do for deriving the parametric t test). They also state that one can estimate the actual variance of AUC directly from data for obtaining tighter intervals, for example, one can use resampling methods to approximate it that they can be useful in practice despite being approximate. This is similar to what we have done in our proposed test. They also criticize the AUC definition of Cortes and Mohri because of the same reason that we have stated above. They argue that AUC and error are different metrics, therefore different analyses should be done for them.

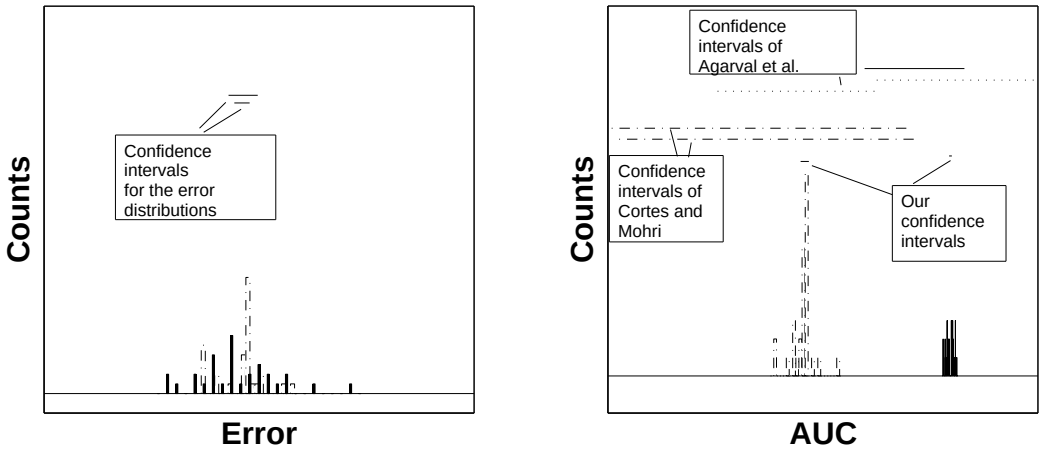
AUC values have been used to compare classifiers over multiple datasets [2]. However, in our work, we try to gain an insight to the difference in the behavior of the error and AUC tests. J. Demsar compares two classifiers with paired t test over multiple datasets. They state that this test makes normality assumption on the difference of random variables and for this, the dataset size should be approximately 30. They also use the Wilcoxon signed-ranks test since it is nonparametric compared to the paired t test. They calculate AUC values by applying 5-fold internal cross validation and take the average of them, thus they do not apply test on these values like us. They compare AUC of different C4.5 algorithms over 14 datasets. They state that commensurability of differences over datasets can be assumed and no distribution assumption is done in this nonparametric test compared to the paired t test. They compare AUC's of C4.5 algorithms with 5-fold internal cross validation over 14 datasets using the Friedman test which is a nonparametric version of ANOVA.

The ROC curves are preferred when there is class skewness and/or different misclassification costs. The effect of class distribution on error and AUC is explored in [16]. On the other hand, we explore the effect of imbalanced cost in error and AUC.

VI. Discussion

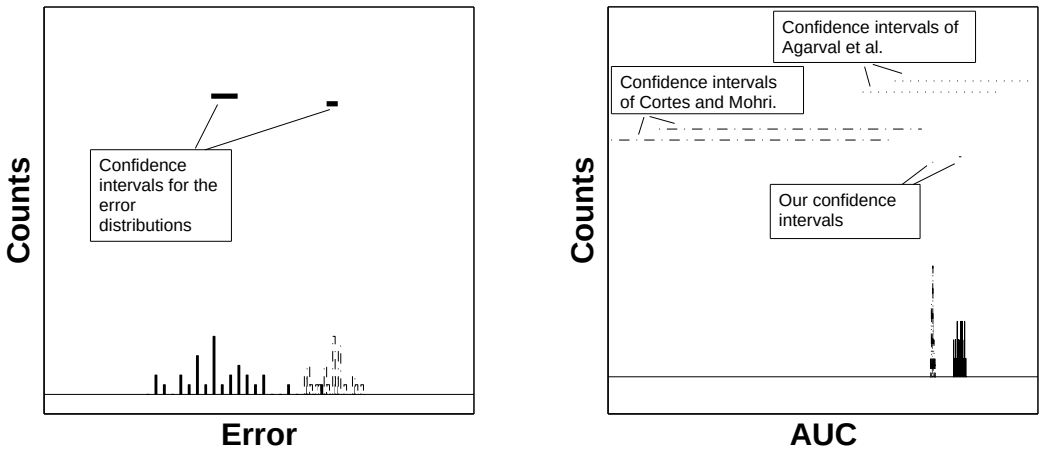
It has been known that the ROC curve or the AUC value gives more information than the misclassification error [5], but still, tests in literature all use misclassification error.

In this paper, we propose a novel statistical comparison procedure based on AUC of the ROC curves. To check for significant difference (unaffected by randomness), for each classifier, we use k-fold cross validation to construct multiple ROC curves and calculate an AUC value for each. We then use the paired t test to test hypotheses on such AUC distributions.



(a) ERROR DISTRIBUTION (b) AUC DISTRIBUTION

Fig. 6. Confidence intervals for error and AUC for the case where the error test accepts and the AUC test rejects the null hypothesis



(a) ERROR DISTRIBUTION (b) AUC DISTRIBUTION

Fig. 7. Confidence intervals for error and AUC for the case where the both error test and the AUC test reject the null hypothesis

To validate our test, we compare it with the paired t test on misclassification errors. We see that our AUC test and the one using error give consistent decisions on a high proportion of cases. When they disagree, we believe that the one using AUC values are more to be trusted because they compare under a set of possible losses and not just a single one of equal loss for false positives and false negatives.

Both the error test and our test use the central limit theorem which states that the sum of a large number of iid random variables (the Bernoulli random variables corresponding to 0/1 decisions on test instances) is approximately normal. We see in practice that the distributions for error or AUC are sometimes not normal, probably due to dependence between folds which share data and the fact that 30 is a relatively small number for central limit theorem to hold. We therefore believe that it may also be interesting to check how nonparametric tests can be used to compare AUC distributions; this is future work.

REFERENCES

[1] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning classifiers," *Neural Computation*, vol. 10, pp. 1895–1923, 1998.

[2] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[3] A. E., "Combined 5×2 cv F test for comparing supervised classification learning classifiers," *Neural Computation*, vol. 11, pp. 1975–1982, 1999.

[4] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, pp. 29–36, 1982.

[5] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.

[6] C. X. Ling, J. Huang, and H. Zhang, "AUC: a better measure than accuracy in comparing learning algorithms," in *In Proc. of IJCAI03*. Springer, 2003, pp. 329–341.

[7] J. Huang, J. Lu, and C. Ling, "Comparing naive Bayes, decision trees, and SVM with AUC and accuracy," in *Proceedings of the Third IEEE International Conference on Data Mining*, 2003, pp. 553–556.

[8] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, pp. 1145–1159, 1997.

[9] I. Guyon, A. R. S. Azar, G. Dror, and G. Cawley, "Agnostic learning vs. prior knowledge challenge & data representation discovery workshop," Florida, 2007. [Online]. Available: http://www.agnostic.inf.ethz.ch/datasets.php

[10] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Meteo, CA: Morgan Kaufmann, 1993.

[11] W. W. Cohen, "Fast effective rule induction," in *The Twelfth International Conference on Machine Learning*, 1995, pp. 115–123.

[12] J. A. Hanley and B. J. McNeil, "A method of comparing the areas under receiver operating characteristic curves derived from the same cases," *Radiology*, vol. 148, pp. 839–843, 1983.

[13] H. C. Bravo, G. Wahba, K. E. Lee, B. E. K. Klein, R. Klein, and S. K. Iyengar, "Examining the relative influence of familial, genetic, and environmental covariate information in flexible risk models," in *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, vol. 106, 2009, pp. 8128–8133.

[14] C. Cortes and M. Mohri, "Confidence intervals for the area under the ROC curve," in *Advances in Neural Information Processing Systems 17, NIPS 2004*, Vancouver, Canada, 2004.

[15] S. Agarwal, T. Graepel, R. Herbrich, and D. Roth, "Generalization bounds for the area under the ROC curve," *Journal of Machine Learning Research*, vol. 6, pp. 393–425, 2005.

[16] G. M. Weiss and F. Provost, "Learning when training data are costly: The effect of class distribution on tree induction," *Journal Of Artificial Intelligence Research*, vol. 19, pp. 315–354, 2003.