AUC: a Statistically Consistent and more Discriminating Measure than Accuracy

Charles X. Ling **Jin Huang**
Department of Computer Science
The University of Western Ontario
London, Ontario, Canada N6A 5B7
{ling, jhuang}@csd.uwo.ca

Harry Zhang
Faculty of Computer Science
University of New Brunswick
Fredericton, NB, Canada E3B 5A3
hzhang@unb.ca

Abstract

Predictive accuracy has been used as the main and often only evaluation criterion for the predictive performance of classification learning algorithms. In recent years, the area under the ROC (Receiver Operating Characteristics) curve, or simply AUC, has been proposed as an alternative single-number measure for evaluating learning algorithms. In this paper, we prove that AUC is a better measure than accuracy. More specifically, we present rigourous definitions on consistency and discriminancy in comparing two evaluation measures for learning algorithms. We then present empirical evaluations and a formal proof to establish that AUC is indeed statistically consistent and more discriminating than accuracy. Our result is quite significant since we formally prove that, for the first time, AUC is a better measure than accuracy in the evaluation of learning algorithms.

1 Introduction

The predictive ability of the classification algorithm is typically measured by its predictive accuracy (or error rate, which is 1 minus the accuracy) on the testing examples. However, most classifiers (including C4.5 and Naive Bayes) can also produce probability estimations or "confidence" of the class prediction. Unfortunately, this information is completely ignored in accuracy. This is often taken for granted since the true probability is unknown for the testing examples anyway.

In many applications, however, accuracy is not enough. For example, in direct marketing, we often need to promote the top X% of customers during gradual roll-out, or we often deploy different promotion strategies to customers with different likelihood of buying some products. To accomplish these tasks, we need more than a mere classification of buyers and non-buyers. We need (at least) a ranking of customers in terms of their likelihoods of buying. If we want to achieve a more accurate ranking from a classifier, one might naturally expect that we must need the true ranking in the training examples [Cohen *et al.*, 1999]. In most scenarios, however, that is not possible. Instead, what we are given is a dataset of examples with class labels only. Thus, given only classification labels in training and testing sets, are there better methods than accuracy to evaluate classifiers that also produce rankings?

The ROC (Receiver Operating Characteristics) curve has been recently introduced to evaluate machine learning algorithms [Provost and Fawcett, 1997; Provost *et al.*, 1998]. It compares the classifiers' performance across the entire range of class distributions and error costs. However, often there is no clear dominating relation between two ROC curves in the entire range; in those situations, the area under the ROC curve, or simply AUC, provides a single-number "summary" for the performance of the learning algorithms. Bradley [Bradley, 1997] has compared popular machine learning algorithms using AUC, and found that AUC exhibits several desirable properties compared to accuracy. For example, AUC has increased sensitivity in Analysis of Variance (ANOVA) tests, is independent to the decision threshold, and is invariant to *a priori* class probability distributions [Bradley, 1997]. However, no formal arguments or criteria have been established. How can we compare two evaluation measures for learning algorithms? How can we establish that one measure is better than another? In this paper, we give formal definitions on the consistency and discriminancy for comparing two measures. We show, empirically and formally, that AUC is indeed a statistically consistent and more discriminating measure than accuracy.

One might ask why we need to care about anything more than accuracy, since by definition, classifiers only classify examples (and do not care about ranking and probability). We answer this question from two aspects. First, as we discussed earlier, even with labelled training and testing examples, most classifiers do produce probability estimations that can rank training/testing examples. Ranking is very important in most real-world applications. As we establish that AUC is a better measure than accuracy, we can choose classifiers with better AUC, thus producing better ranking. Second, and more importantly, if we build classifiers that optimize AUC (instead of accuracy), it has been shown [Ling and Zhang, 2002] that such classifiers produce not only better AUC (a natural consequence), but also better accuracy (a surprising result), compared to classifiers that optimize the accuracy. To make an analogy, when we train workers on a more complex task, they will do better on a simple task than workers who are trained only on the simple task.

Our work is quite significant for several reasons. First, we establish rigourously, for the first time, that even given only labelled examples, AUC is a better measure (defined in Section 2.2) than accuracy. Second, our result suggests that AUC should replace accuracy in comparing learning algorithms in the future. Third, our results prompt and allow us to re-evaluate well-established results in machine learning. For example, extensive experiments have been conducted and published on comparing, in terms of accuracy, decision tree classifiers to Naive Bayes classifiers. A well-established and accepted conclusion in the machine learning community is that those learning algorithms are very similar as measured by accuracy [Kononenko, 1990; Langley *et al.*, 1992; Domingos and Pazzani, 1996]. Since we have established that AUC is a better measure, are those learning algorithms still very similar as measured by AUC? We have shown [Ling *et al.*, 2003] that, surprisingly, this is not true: Naive Bayes is significantly better than decision tree in terms of AUC. These kinds of new conclusions are very useful to the machine learning community, as well as to machine learning applications (e.g., data mining). Fourth, as a new measure (such as AUC) is discovered and proved to be better than a previous measure (such as accuracy), we can re-design most learning algorithms to optimize the new measure [Ferri *et al.*, 2002; Ling and Zhang, 2002]. This would produce classifiers that not only perform well in the new measure, but also in the previous measure, compared to the classifiers that optimize the previous measure, as shown in [Ling and Zhang, 2002]. This would further improve the performance of our learning algorithms.

2 Criteria for Comparing Evaluation Measures for Classifiers

We start with some intuitions in comparing AUC and accuracy, and then we present formal definitions in comparing evaluation measures for learning algorithms.

2.1 AUC vs Accuracy

Hand and Till [Hand and Till, 2001] present a simple approach to calculating the AUC of a classifier G below.

$$\hat{A} = \frac{S_0 - n_0(n_0+1)/2}{n_0 n_1}, \tag{1}$$

where n_0 and n_1 are the numbers of positive and negative examples respectively, and $S_0 = \sum r_i$, where r_i is the rank of i_{th} positive example in the ranked list. Table 1 shows an example of how to calculate AUC from a ranked list with 5 positive examples and 5 negative examples. The AUC of the ranked list in Table 1 is $\frac{(5+7+8+9+10)-5\times 6/2}{5\times 5}$, which is 24/25. It is clear that AUC obtained by Equation 1 is a measure for the quality of ranking, as the more positive examples are ranked higher (to the right of the list), the larger the term $\sum r_i$. AUC is shown to be equivalent to the Wilcoxon statistic rank test [Bradley, 1997].

Intuitively, we can see why AUC is a better measure than accuracy from the following example. Let us consider two classifiers, Classifier 1 and Classifier 2, both producing probability estimates for a set of 10 testing examples. Assume that both classifiers classify 5 of the 10 examples as positive, and the other 5 as negative. If we rank the testing examples according to increasing probability of being $+$ (positive), we get the two ranked lists as in Table 2.

	$-$	$-$	$-$	$-$	$+$	$-$	$+$	$+$	$+$	$+$
i					1		2	3	4	5
r_i					5		7	8	9	10

Table 1: An example for calculating AUC with r_i

Classifier 1	$- - - - +$	$- + + + +$
Classifier 2	$+ - - - -$	$+ + + + -$

Table 2: An Example in which two classifiers have the same classification accuracy, but different AUC values

Clearly, both classifiers produce an error rate of 20% (one false positive and one false negative), and thus the two classifiers are equivalent in terms of error rate. However, intuition tells us that Classifier 1 is better than Classifier 2, since overall positive examples are ranked higher in Classifier 1 than 2. If we calculate AUC according to Equation 1, we obtain that the AUC of Classifier 1 is $\frac{24}{25}$ (as seen in Table 1), and the AUC of Classifier 2 is $\frac{16}{25}$. Clearly, AUC does tell us that Classifier 1 is indeed better than Classifier 2.

Unfortunately, "counter examples" do exist, as shown in Table 3 on two other classifiers: Classifier 3 and Classifier 4. It is easy to obtain that the AUC of Classifier 3 is $\frac{21}{25}$, and the AUC of Classifier 4 is $\frac{16}{25}$. However, the error rate of Classifier 3 is 40%, while the error rate of Classifier 4 is only 20% (again we assume that the threshold for accuracy is set at the middle so that 5 examples are predicted as positive and 5 as negative). Therefore, a larger AUC does not always imply a lower error rate.

Classifier 3	$- - - + +$	$- - + + +$
Classifier 4	$+ - - - -$	$+ + + + -$

Table 3: A counter example in which one classifier has higher AUC but lower classification accuracy

Another intuitive argument for why AUC is better than accuracy is that AUC is more discriminating than accuracy since it has more possible values. More specifically, given a dataset with n examples, there is a total of only $n+1$ different classification accuracies ($0/n$, $1/n$, ..., n/n). On the other hand, assuming there are n_0 positive examples and n_1 negative examples ($n_0 + n_1 = n$), there are $n_0 n_1 + 1$ different AUC values ($0/n_0 n_1$, $1/n_0 n_1$, ..., $n_0 n_1/n_0 n_1$), generally more than $n+1$. However, counter examples also exist in this regard. Table 4 illustrates that classifiers with the same AUC can have different accuracies. Here, we see that Classifier 5 and Classifier 6 have the same AUC($\frac{3}{5}$) but different accuracies (60% and 40% respectively). In general, a measure with more values is not necessarily more discriminating. For example, the weight of a person (having infinitely many possible values) has nothing to do with the number of siblings (having only a small number of integer values) he or she has.

Classifier 5	$- - + + -$	$+ + - - +$
Classifier 6	$- - + + +$	$- - + - +$

Table 4: A counter example in which two classifiers have same AUC but different classification accuracies

How do we compare different evaluation measures for learning algorithms? Some general criteria must be established.

2.2 (Strict) Consistency and Discriminancy of Two Measures

Intuitively speaking, when we discuss two different measures f and g on evaluating two learning algorithms A and B, we want at least that f and g be *consistent* with each other. That is, when f stipulates that algorithm A is (strictly) better than B, then g will not say B is better than A. Further, if f is more *discriminating* than g, we would expect to see cases where f can tell the difference between algorithms A and B but g cannot.

This intuitive meaning of consistency and discriminancy can be made precise as the following definitions.

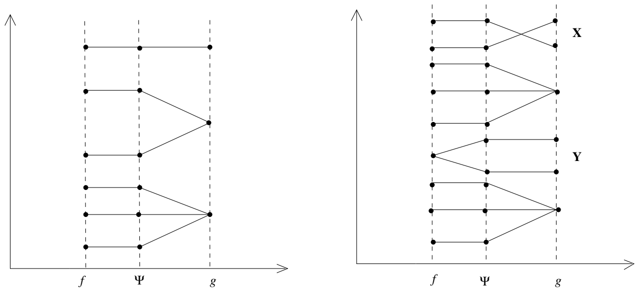
Definition 1 (Consistency) *For two measures f, g on domain Ψ, f, g are (strictly) consistent if there exist no a, $b \in \Psi$, such that $f(a) > f(b)$ and $g(a) < g(b)$.*

Definition 2 (Discriminancy) *For two measures f, g on domain Ψ, f is (strictly) more discriminating than g if there exist a, $b \in \Psi$ such that $f(a) > f(b)$ and $g(a) = g(b)$, and there exist no a, $b \in \Psi$ such that $g(a) > g(b)$ and $f(a) = f(b)$.*

As an example, let us think about numerical marks and letter marks that evaluate university students. A numerical mark gives 100, 99, 98, ..., 1, or 0 to students, while a letter mark gives A, B, C, D, or F to students. Obviously, we regard A > B > C > D > F. Clearly, numerical marks are consistent with letter marks (and vice versa). In addition, numerical marks are more discriminating than letter marks, since two students who receive 91 and 93 respectively receive different numerical marks but the same letter mark, but it is not possible to have students with different letter marks (such as A and B) but with the same numerical marks. This ideal example of a measure f (numerical marks) being strictly consistent and more discriminating than another g (letter marks) can be shown in the figure 1(a).

2.3 Statistical Consistency and Discriminancy of Two Measures

As we have already seen in Section 2.1, counter examples on consistency and discriminancy do exist for AUC and accuracy. Therefore, it is impossible to prove the consistency and discriminancy on AUC and accuracy based on Definitions 1 and 2. Figure 1(b) illustrates a situation where one measure f is not completely consistent with g, and is not strictly more discriminating than g. In this case, we must consider the probability of being consistent and degree of being more discriminating. What we will define and prove is the *probabilistic version* of the two definitions on strict consistency and discriminancy. That is, we extend the previous definitions to *degree of* consistency and *degree* of discriminancy, as follows:



(a) (b)

Figure 1: Illustrations of two measures f and g. A link between two points indicates that the function values are the same on domain Ψ. In (a), f is strictly consistent and more discriminating than g. In (b), f is not strictly consistent or more discriminating than g. Counter examples on consistency (denoted by X in the figure) and discriminancy (denoted by Y) exist here

Definition 3 (Degree of consistency) *For two measures f and g on domain Ψ, let $R = \{(a,b)|a,b \in \Psi, f(a) > f(b), g(a) > g(b)\}$, $S = \{(a,b)|a,b \in \Psi, f(a) > f(b), g(a) < g(b)\}$. The degree of consistency* [1] *of f and g is $\mathbf{C}$ ($0 \le \mathbf{C} \le 1$), where $\mathbf{C} = \frac{|R|}{|R|+|S|}$.*

Definition 4 (Degree of discriminancy) *For two measures f and g on domain Ψ, let $P = \{(a,b)|a,b \in \Psi, f(a) > f(b), g(a) = g(b)\}$, $Q = \{(a,b)|a,b \in \Psi, g(a) > g(b), f(a) = f(b)\}$. The degree of discriminancy for f over g is $\mathbf{D} = \frac{|P|}{|Q|}$.*

There are clear and important implications of these definitions of measures f and g in evaluating two machine learning algorithms, say A and B. If f and g are consistent to degree $\mathbf{C}$, then when f stipulates that A is better than B, there is a probability $\mathbf{C}$ that g will agree (stipulating A is better than B). If f is $\mathbf{D}$ times more discriminating than g, then it is $\mathbf{D}$ times more likely that f can tell the difference between A and B but g cannot than that g can tell the difference between A and B but f cannot. Clearly, we require that $\mathbf{C} > 0.5$ and $\mathbf{D} > 1$ if we want to conclude a measure f is "better" than a measure g. This leads to the following definition:

Definition 5 *The measure f is statistically consistent and more discriminating than g if and only if $\mathbf{C} > 0.5$ and $\mathbf{D} > 1$. In this case, we say, intuitively, that f is a better measure than g.*

The statistical consistency and discriminancy is a special case of the strict consistency and more discriminancy. For the example of numerical and letter marks in the student evaluation discussed in Section 2.2, we can obtain that $\mathbf{C} = 1.0$ and $\mathbf{D} = \infty$, as the former is strictly consistent and more discriminating than the latter.

[1] It is easy to prove that this definition is symmetric; that is, the degree of consistency of f and g is same as the degree of consistency of g and f.

To prove AUC is statistically consistent and more discriminating than accuracy, we substitute f by AUC and g by accuracy in the definition above. To simplify our notation, we will use *AUC* to represent AUC values, and *acc* for accuracy. The domain Ψ is the ranked lists of testing examples (with n_0 positive and n_1 negative examples). Since we require $\mathbf{C} > 0.5$ and $\mathbf{D} > 1$ we will essentially need to prove:

Theorem 1 *Given a domain Ψ, let $R = \{(a,b)|AUC(a) > AUC(b), acc(a) > acc(b), a, b \in \Psi\}$, $S = \{(a,b)|AUC(a) < AUC(b), acc(a) > acc(b), a, b \in \Psi\}$. Then $\frac{|R|}{|R|+|S|} > 0.5$ or $|R| > |S|$.*

Theorem 2 *Given a domain Ψ, let $P = \{(a,b)|AUC(a) > AUC(b), acc(a) = acc(b), a, b \in \Psi\}$, $Q = \{(a,b)|acc(a) > acc(b), AUC(a) = AUC(b), a, b \in \Psi\}$. Then $|P| > |Q|$.*

3 Empirical Verification on AUC and Accuracy

Before we present a formal proof that AUC is statistically consistent and more discriminating than accuracy, we first present an empirical verification. To simplify our notation, we will use AUC to represent AUC values, and acc for accuracy.

We conduct experiments with (artificial) testing sets to verify the statistical consistency and discriminancy between AUC and accuracy. The datasets are balanced with equal numbers of positive and negative examples. We test datasets with 4, 6, 8, 10, 12, 14, and 16 testing examples. For each number of examples, we enumerate all possible ranked lists of positive and negative examples. For the dataset with $2n$ examples, there are $\binom{2n}{n}$ such ranked lists.

We exhaustively compare all pairs of ranked lists to see how they satisfy the consistency and discriminating propositions probabilistically. To obtain degree of consistency, we count the number of pairs which satisfy "$AUC(a) > AUC(b)$ and $acc(a) > acc(b)$", and the number of pairs which satisfy "$AUC(a) > AUC(b)$ and $acc(a) < acc(b)$". We then calculate the percentage of those cases; that is, the degree of consistency. To obtain degree of discriminancy, we count the number of pairs which satisfy "$AUC(a) > AUC(b)$ and $acc(a) = acc(b)$", and the number of pairs which satisfy "$AUC(a) = AUC(b)$ and $acc(a) > acc(b)$".

Tables 5 and 6 show the experiment results for the balanced dataset. For consistency, we can see (Table 5) that for various numbers of balanced testing examples, given $AUC(a) > AUC(b)$, the number (and percentage) of cases that satisfy $acc(a) > acc(b)$ is much greater than those that satisfy $acc(a) < acc(b)$. When n increases, the degree of consistency ($\mathbf{C}$) seems to approach 0.94, much larger than the required 0.5. For discriminancy, we can see clearly from Table 6 that the number of cases that satisfy $AUC(a) > AUC(b)$ and $acc(a) = acc(b)$ is much more (from 15.5 to 18.9 times more) than the number of cases that satisfy $acc(a) > acc(b)$ and $AUC(a) = AUC(b)$. When n increases, the degree of discriminancy ($\mathbf{D}$) seems to approach 19, much larger than the required threshold 1.

These empirical experiments verify empirically that AUC is indeed a statistically consistent and more discriminating measure than accuracy for the balanced datasets, as a measure for learning algorithms.

#	$AUC(a) > AUC(b)$ & $acc(a) > acc(b)$	$AUC(a) > AUC(b)$ & $acc(a) < acc(b)$	**C**
4	9	0	1.0
6	113	1	0.991
8	1459	34	0.977
10	19742	766	0.963
12	273600	13997	0.951
14	3864673	237303	0.942
16	55370122	3868959	0.935

Table 5: Experimental results for verifying statistical consistency between AUC and accuracy for the balanced dataset

#	$AUC(a) > AUC(b)$ & $acc(a) = acc(b)$	$acc(a) > acc(b)$ & $AUC(a) = AUC(b)$	**D**
4	5	0	∞
6	62	4	15.5
8	762	52	14.7
10	9416	618	15.2
12	120374	7369	16.3
14	1578566	89828	17.6
16	21161143	1121120	18.9

Table 6: Experimental results for verifying AUC is statistically more discriminating than accuracy for the balanced dataset

4 Formal Proof

In our following discussion, we assume that the domain $\mathbf{\Psi}$ is the set of all the ranked lists with n_0 positive and n_1 negative examples. In this paper, we only study the cases where a ranked list contains an equal number of positive and negative examples, and we assume that the cutoff for classification is at the exact middle of the ranked list (each classifier classifies exactly half examples into the positive class and the other half into the negative class). That is, $n_0 = n_1$, and we use n instead of n_0 and n_1.

Lemma 1 gives the maximum and minimum AUC values for a ranked list with a fixed accuracy rate.

Lemma 1 *For a given accuracy rate α, let $AUC_{max}(\alpha) = max\{AUC(r)|acc(r) = \alpha, r \in \Psi\}$, $AUC_{min}(\alpha) = min\{AUC(r)|acc(r) = \alpha, r \in \Psi\}$. Then $AUC_{min}(\alpha) = \alpha^2$, $AUC_{max}(\alpha) = 1 - (1-\alpha)^2$.*

Proof: Assume that there are n_p positive examples correctly classified. Then there are $n - n_p$ positive examples in negative section. According to Equation 1, AUC value is determined by S_0, which is the sum of indexes of all positive examples. So when all $n - n_p$ positive examples are put on the highest positions in negative section (their indexes are $n, n-1, \cdots, n_p+1$), and n_p positive examples are put on the highest positions in positive section (their indexes are $2n, 2n-1, \cdots, 2n-n_p+1$), AUC reaches the maximum value. So

$$AUC_{max}(\alpha) = \frac{\sum_{n_p+1}^{n} i + \sum_{2n-n_p+1}^{2n} i - n(n+1)/2}{n^2}.$$

Similarly when all positive examples are put on the lowest positions in both positive and negative sections, AUC reaches the minimum value. Thus

$$AUC_{min}(\alpha) = \frac{\sum_{1}^{n-n_p} i + \sum_{n+1}^{n+n_p} i - n(n+1)/2}{n^2}.$$

Since $\alpha = \frac{n_p}{n}$, from above two formula, we can obtain $AUC_{min}(\alpha) = \alpha^2$, $AUC_{max}(\alpha) = 1 - (1-\alpha)^2$. □

Lemma 2 *For given accuracy rates $\alpha > \beta$ and a domain Ψ, $AUC_{max}(\alpha) > AUC_{max}(\beta)$, $AUC_{min}(\alpha) > AUC_{min}(\beta)$.*

Proof: It is straightforward to prove it from Lemma 1. □

Lemma 3 *Let $\Gamma(\alpha) = \{AUC(r)|acc(r) = \alpha, r \in \Psi\}$, $R_1 = \{r|acc(r) = \alpha, AUC(r) = \beta, r \in \Psi\}$, $R_2 = \{r|acc(r) = \alpha, AUC(r) = AUC_{max}(\alpha) + AUC_{min}(\alpha) - \beta, r \in \Psi\}$. Then*
(a)$|\Gamma(\alpha)| = 2nn_p - 2{n_p}^2 + 1$.
(b)$|R_1| = |R_2|$.

Proof: (a) It is straightforward to prove it from Lemma 1.
(b) For any ranked list $r \in R_1$ we can convert it to another unique ranked list $r' \in R_2$ by exchanging the positions of examples. In the negative section, for any positive example whose position is r_i, we exchange it with the example in the position of $n+1-r_i$. In the positive section, for any positive example in position r_i, we exchange it with the example in the position of $n-(r_i-n)+1+n$. It's easy to obtain that $AUC(r') = AUC_{max} + AUC_{min} - \beta$. So $r' \in R_2$. Thus, we have that $|R_1| \le |R_2|$. Similarly, we can prove $|R_2| \le |R_1|$. Therefore $|R_1| = |R_2|$. □

Theorem 1 *Given a domain Ψ, let $R = \{(a,b)|AUC(a) > AUC(b), acc(a) > acc(b), a, b \in \Psi\}$, $S = \{(a,b)|AUC(a) < AUC(b), acc(a) > acc(b), a, b \in \Psi\}$. Then $\frac{|R|}{|R|+|S|} > 0.5$ or $|R| > |S|$.*

Proof: To prove $|R| > |S|$, first we prove for given accuracy rates $\alpha > \beta$, there are more ranked list pairs (a,b) to satisfy $AUC(a) > AUC(b)$ than that to satisfy $AUC(a) < AUC(b)$, where $acc(a) = \alpha, acc(b) = \beta$. Let $R_{\alpha\beta} = \{(a,b)|AUC(a) > AUC(b), acc(a) = \alpha, acc(b) = \beta, a, b \in \Psi\}$ and $S_{\alpha\beta} = \{(a,b)|AUC(a) < AUC(b), acc(a) = \alpha, acc(b) = \beta, a, b \in \Psi\}$. Then we should prove $|R_{\alpha\beta}| > |S_{\alpha\beta}|$.

1) If $AUC_{min}(\alpha) > AUC_{max}(\beta)$, then it is obvious $|R_{\alpha\beta}| > 0, |S_{\alpha\beta}| = 0$. So $|R_{\alpha\beta}| > |S_{\alpha\beta}|$.

2) Otherwise, $|R_{\alpha\beta}|$ can be computed by adding two parts together. $|R_{\alpha\beta}| = |P_1| + |P_2|$, where $P_1 = \{(a,b)|acc(a) = \alpha, acc(b) = \beta, AUC(a) > AUC_{max}(\beta), a, b \in \Psi\}$, and $P_2 = \{(a,b)|acc(a) = \alpha, acc(b) = \beta, AUC_{max}(\beta) \ge AUC(a) > AUC(b), a, b \in \Psi\}$.

For $S_{\alpha\beta}$, we have $S_{\alpha\beta} = \{(a,b)|acc(a) = \alpha, acc(b) = \beta, AUC_{max}(\beta) \ge AUC(b) > AUC(a) > AUC_{min}(\alpha), a, b \in \Psi\}$.

It can be proved that $|R_{\alpha\beta}| > |S_{\alpha\beta}|$. We omit the details due to the space limitation.

Since $R = \bigcup_{\alpha,\beta} R_{\alpha\beta}$, $S = \bigcup_{\alpha,\beta} S_{\alpha\beta}$, and $R_{\alpha_1\beta_1} \cap R_{\alpha_2\beta_2} = \phi$, $S_{\alpha_1\beta_1} \cap S_{\alpha_2\beta_2} = \phi$, for $\alpha_1 \ne \alpha_2$ or $\beta_1 \ne \beta_2$, where $\alpha > \beta$. So $|R| = \sum_{\alpha,\beta} |R_{\alpha\beta}|$, $|S| = \sum_{\alpha,\beta} |S_{\alpha\beta}|$. For any given accuracies $\alpha > \beta$, we have proved $|R_{\alpha\beta}| > |S_{\alpha\beta}|$, therefore, $|R| > |S|$. □

Now let us explore the discriminancy of AUC and accuracy. Lemma 4 to Lemma 5 are proposed as a basis to prove Theorem 2. In the following lemmas, $|r|$ stands for the number of examples in ranked list r, and $|r|_+$ stands for the number of positive examples in r.

Lemma 4 *Assume that R_1 and R_2 are two sets of ranked lists on a given domain Ψ, and that $R_1 = \{r||r| = n, |r|_+ = m, 0 < m < \frac{n}{2}, AUC(r) = C, r \in \Psi\}$, $R_2 = \{r||r| = n, |r|_+ = n-m, 0 < m < \frac{n}{2}, AUC(r) = C, r \in \Psi\}$. Then $|R_1| = |R_2|$.*

Proof: For any ranked list $r \in R_1$, we switch the examples on position i and $n-i$, to obtain another unique ranked list s. Next we replace all positive examples in s by negative examples and all negative examples in s by positive examples to obtain ranked list s'. It is obvious that $AUC(s') = AUC(r)$. Since $|s'|_+ = n-m$, $s' \in R_2$. Thus, $|R_1| \le |R_2|$. Similarly we can prove $|R_2| \le |R_1|$. Therefore $|R_1| = |R_2|$. □

Lemma 5 *Given a domain Ψ, let α and β be accuracy rates $\alpha < \beta \le \frac{1}{2}$. Consider a series of ranked list sets $R_i = \{r|acc(r) = \alpha, AUC(r) = \frac{i}{n^2}, r \in \Psi\}$, where $i = 0, \cdots, n^2$; $S_j = \{s|acc(s) = \beta, AUC(s) = \frac{j}{n^2}, r \in \Psi\}$, where $j = 0, \cdots, n^2$; Let $a_i = |R_i|$, $b_i = |S_i|$, $c_i = a_i + b_i$. Then*

$$\sum_{i \ne j} c_i \bullet c_j \ge \sum_{i=0}^{n^2} a_i \bullet \sum_{i=0}^{n^2} b_i$$

.

Proof: We omit the proof due to the limitation of space. □

Theorem 2 *Given a domain Ψ, let $P = \{(a,b)|AUC(a) > AUC(b), acc(a) = acc(b), a, b \in \Psi\}$, $Q = \{(a,b)|acc(a) > acc(b), AUC(a) = AUC(b), a, b \in \Psi\}$. Then $|P| > |Q|$.*

Proof: Since Ψ is the set of all the ranked lists with equal number of positive and negative examples, Ψ can be partitioned into ranked list sets P_{ij} by different accuracy rates and AUC values. $P_{ij} = \{r|acc(r) = \frac{j}{n}, AUC(r) = \frac{i}{n^2}, 0 \le i \le n^2, 0 \le j \le n\}$. It is obvious $\Psi = \bigcup P_{ij}$. Let $t_{ij} = |P_{ij}|$. Then by Lemma 5 for any $i \ne j$ we have

$$\sum_{k \ne l}^{n^2} (t_{ki} + t_{kj})(t_{li} + t_{lj}) \ge (\sum_{k=0}^{n^2} t_{ki})(\sum_{l=0}^{n^2} t_{lj}).$$

After we sum all i and j, we have

$$\sum_{i \ne j}^{n} \sum_{k \ne l}^{n^2} (t_{ki} + t_{kj})(t_{li} + t_{lj}) \ge \sum_{i \ne j}^{n} (\sum_{k=0}^{n^2} t_{ki})(\sum_{l=0}^{n^2} t_{lj}). \quad (2)$$

Also for all $i \ne j$ and $k \ne l$, it's easy to obtain

$$\sum_{i \ne j}^{n} \sum_{k \ne l}^{n^2} (t_{ki} + t_{kj})(t_{li} + t_{lj}) = \sum_{i \ne k} t_{ij} t_{kl}. \quad (3)$$

Therefore, from inequality 2 and equation 3, we will get

$$\sum_{i \neq k} t_{ij} t_{kl} > \sum_{i \neq j}^{n} (\sum_{k=0}^{n^2} t_{ki})(\sum_{k=0}^{n^2} t_{kj}). \tag{4}$$

On the other hand,

$$|D| = \sum_{i=0}^{n^2} \sum_{j=0}^{n} t_{ij} = \sum_{j=0}^{n} \sum_{i=0}^{n^2} t_{ij}.$$

So

$$(\sum_{j=0}^{n} \sum_{i=0}^{n^2} t_{ij})^2 = (\sum_{i=0}^{n^2} \sum_{j=0}^{n} t_{ij})^2.$$

The above equation can also be written as

$$\begin{aligned} \sum_{i=0}^{n^2} (\sum_{j=0}^{n} t_{ij})^2 \quad &+ \quad \sum_{i \neq k}^{n^2} (\sum_{j=0}^{n} t_{ij} \sum_{j=0}^{n} t_{kj}) = \\ \sum_{j=0}^{n} (\sum_{i=0}^{n^2} t_{ij})^2 \quad &+ \quad \sum_{j \neq k}^{n} (\sum_{i=0}^{n^2} t_{ij} \sum_{i=0}^{n^2} t_{ik}). \end{aligned} \tag{5}$$

According to formula 4, we have

$$\sum_{i \neq k}^{n^2} (\sum_{j=0}^{n} t_{ij} \sum_{j=0}^{n} t_{kj}) = \sum_{i \neq k} t_{ij} t_{kl} > \sum_{j \neq k}^{n} (\sum_{i=0}^{n^2} t_{ij} \sum_{i=0}^{n^2} t_{ik}). \tag{6}$$

Then by equation 5 and inequality 6, we have

$$\sum_{i=0}^{n^2} (\sum_{j=0}^{n} t_{ij})^2 < \sum_{j=0}^{n} (\sum_{i=0}^{n^2} t_{ij})^2. \tag{7}$$

Let

$$H_i = \sum_{j=0}^{n} t_{ij},$$

and

$$V_j = \sum_{i=0}^{n^2} t_{ij}.$$

Then

$$|R| = \sum_{i=0}^{n} \binom{V_i}{2} = \sum_{i=0}^{n} \frac{V_i(V_i - 1)}{2} = \frac{1}{2} \sum_{i=0}^{n} V_i^2 - \frac{1}{2}|D|$$

$$|S| = \sum_{i=0}^{n^2} \binom{H_i}{2} = \sum_{i=0}^{n^2} \frac{H_i(H_i - 1)}{2} = \frac{1}{2} \sum_{i=0}^{n^2} H_i^2 - \frac{1}{2}|D|$$

Therefore, according to inequality 7, we have $|R| > |S|$. $\square$

5 Conclusions

In this paper, we give formal definitions of discriminancy and consistency in comparing evaluation measures for learning algorithms. We prove that AUC is statistically consistent and more discriminating than accuracy. Our result is quite significant since we have established rigourously, for the first time, that AUC is a better measure than accuracy in evaluating and comparing classification learning algorithms. This conclusion has many important implications in evaluating, comparing, and designing learning algorithms. In our future research, we will extend the results in this paper to cases with unbalanced class distribution and multiple classes.

References

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

[Cohen *et al.*, 1999] W. W. Cohen, R. E. Schapire, and Y. Singer. Learning to order things. *Journal of Artificial Intelligence Research*, 10:243–270, 1999.

[Domingos and Pazzani, 1996] P. Domingos and M. Pazzani. Beyond independence: conditions for the optimality of the simple Bayesian classifier. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 105 – 112, 1996.

[Ferri *et al.*, 2002] C. Ferri, P. A. Flach, and J. Hernandez-Orallo. Learning decision trees using the area under the ROC curve. In *Proceedings of the Nineteenth International Conference on Machine Learning (ICML 2002)*, pages 139–146, 2002.

[Hand and Till, 2001] D. J. Hand and R. J. Till. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45:171–186, 2001.

[Kononenko, 1990] I. Kononenko. Comparison of inductive and naive Bayesian learning approaches to automatic knowledge acquisition. In B. Wielinga, editor, *Current Trends in Knowledge Acquisition*. IOS Press, 1990.

[Langley *et al.*, 1992] P. Langley, W. Iba, and K. Thomas. An analysis of Bayesian classifiers. In *Proceedings of the Tenth National Conference of Artificial Intelligence*, pages 223–228. AAAI Press, 1992.

[Ling and Zhang, 2002] C. X. Ling and H. Zhang. Toward Bayesian classifiers with accurate probabilities. In *Proceedings of the Sixth Pacific-Asia Conference on KDD*. Springer, 2002.

[Ling *et al.*, 2003] C. X. Ling, J. Huang, and H. Zhang. AUC:A better measure than accuracy in comparing learning algorithms. In *Proceedings of 16th Canadian Conference on Artificial Intelligence*. Springer, 2003. To appear.

[Provost and Fawcett, 1997] F. Provost and T. Fawcett. Analysis and visualization of classifier performance: comparison under imprecise class and cost distribution. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, pages 43–48. AAAI Press, 1997.

[Provost *et al.*, 1998] F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 445–453. Morgan Kaufmann, 1998.