Using AUC and Accuracy in Evaluating Learning Algorithms

Jin Huang and Charles X. Ling

Abstract—The area under the ROC (Receiver Operating Characteristics) curve, or simply AUC, has been traditionally used in medical diagnosis since the 1970s. It has recently been proposed as an alternative single-number measure for evaluating the predictive ability of learning algorithms. However, no formal arguments were given as to why AUC should be preferred over accuracy. In this paper, we establish formal criteria for comparing two different measures for learning algorithms and we show theoretically and empirically that AUC is a better measure (defined precisely) than accuracy. We then reevaluate well-established claims in machine learning based on accuracy using AUC and obtain interesting and surprising new results. For example, it has been well-established and accepted that Naive Bayes and decision trees are very similar in predictive accuracy. We show, however, that Naive Bayes is significantly better than decision trees in AUC. The conclusions drawn in this paper may make a significant impact on machine learning and data mining applications.

Index Terms—Evaluation of learning algorithms, ROC, AUC of ROC, accuracy.

◆

1 INTRODUCTION

THE goal of classification learning algorithms is to build a classifier from a set of training examples with class labels such that the classifier can well predict the unseen testing examples. 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 decision trees [30] and Naive Bayes [9]) can also produce probability estimations[1] or "confidence" of the class prediction. Unfortunately, this information is completely ignored in accuracy. That is, the accuracy measure does not consider the probability (be it 0.51 or 0.99) of the prediction; as long as the class with the largest probability estimation is the same as the target, it is regarded as correct. This is often taken for granted since the true probability is unknown for the testing examples anyway.

In many data mining applications, however, accuracy is not enough. For example, in direct marketing, we often need to promote the top X percent (X can be 5 or 10) customers during gradual roll-out, or we often deploy different promotion strategies to customers with different likelihood of purchasing. To accomplish these tasks, we need more than a mere classification of buyers and nonbuyers. We need (at least) a ranking of customers in terms of their likelihoods of buying. Thus, a ranking is much more desirable than just a classification [20] and it can be easily obtained since most classifiers do produce probability estimations that can be used for ranking (testing) examples.

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 [6]. In most scenarios, however, that is not possible. Instead, what we are given is a data set 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 answer lies in the ROC curve.

The ROC (Receiver Operating Characteristics) curve was first used in signal detection theory to represent the tradeoff between the hit rates and false alarm rates [10], [13]. It has been extensively studied and applied in medical diagnosis since the 1970s [24], [35]. Spackman [33] was one of the first researchers who used the ROC graph to compare and evaluate machine learning algorithms. In recent years, extensive research on ROC has been done in machine learning [28], [29]. The area under the ROC curve, or simply AUC, provides a good "summary" for the performance of the ROC curves. Below we will provide a brief overview of ROC and its AUC.

Let $\{P, N\}$ be the positive and negative instance classes and let $\{\tilde{P}, \tilde{N}\}$ be the classifications produced by a classifier. Let $P(P|I)$ be the posterior probability that an instance I is positive. The true positive rate, TP, of a classifier is:

$$TP = P(\tilde{P}|P) \approx \frac{\text{positives correctly classified}}{\text{total positives}}.$$

The false positive rate, FP, of a classifier is:

$$FP = P(\tilde{P}|N) \approx \frac{\text{negatives incorrectly classified}}{\text{total negatives}}.$$

---

[1] 1. Decision trees that produce probability estimations are called PETs (Probability Estimation Trees) [26]. The leaf nodes of a decision tree may contain training examples of different classes. The probability of a testing instance belonging to a specific class is normally the ratio of training instances of that class over all examples in the leaf node that the testing instance falls in.

- *The authors are with the Department of Computer Science, The University of Western Ontario, London, Ontario, Canada N6A 5B7.*
E-mail: {jhuang, cling}@csd.uwo.ca.

Manuscript received 2 Dec. 2003; revised 3 May 2004; accepted 23 Aug. 2004; published online 19 Jan. 2005.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-0245-1203.

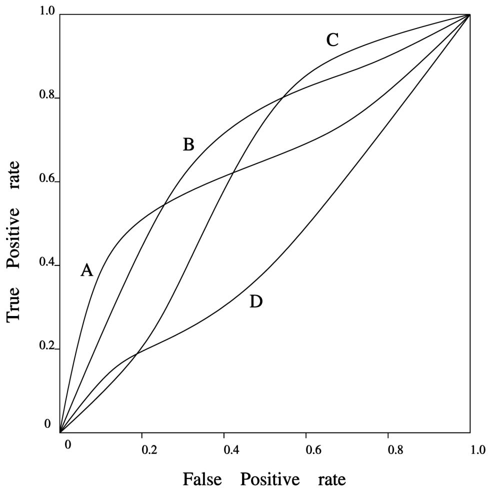


Fig. 1. An example of four ROC curves.

On an ROC graph, TP is plotted on the Y axis and FP is plotted on the X axis. In the ROC space, each classifier with a given class distribution and cost matrix is represented by a point (FP, TP) on the ROC curve. For a model that produces a continuous output, i.e., the probability estimates for Bayesian networks, TP and FP can vary as the threshold on the output varies between its extremes (0 and 1). The resulting curve is called the ROC curve.

The ROC curve compares the classifiers' performance across the entire range of class distributions and error costs. Fig. 1 shows a plot of four ROC curves, each representing one of the four classifiers, A through D. A ROC curve X is said to *dominate* another ROC curve Y if X is always above and to the left of Y. This means that the classifier of X always has a lower expected cost than that of Y, over all possible error costs and class distributions. In this example, A and B dominate D.

However, often there is no clear dominating relation between two ROC curves. For example, curves A and B are not dominating each other in the whole range. In those situations, or when the class distribution and error costs are unknown, the area under the ROC curve, or simply AUC,[2] is a good "summary" for comparing the two ROC curves. We use $AUC(X)$ to denote the area under the ROC curve X in the ROC space. In Fig. 1, since ROC curves A and B dominate ROC curve D, we can easily obtain that $AUC(A) > AUC(D)$ and $AUC(B) > AUC(D)$. Therefore, if domination relations exist among the ROC curves, AUC values can reflect these domination relations. The reverse is not true. However, AUC has a special statistical meaning: it represents the probability that a randomly chosen negative example will have a smaller estimated probability of belonging to the positive class than a randomly chosen positive example [15]. Moreover, AUC also equals to the quantity of Wilcoxon statistic [13]. See Section 2.1 for more details.

Bradley [3] 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 [3]. However, no formal arguments or criteria have been established. Recently, other researchers have even used AUC to construct learning algorithms [12], [22]. But, it is not clear if and why AUC is a better measure than accuracy. In general, 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 first establish formal criteria for comparing two arbitrary single number evaluation measures and then show, both formally and empirically, that AUC is a better measure than accuracy (see Sections 2 and 3). Our notion of a "better" measure is not based on some subjective criterion that a measure is closer to some true target than the other; instead, it is based on the objective criteria of consistency and discriminancy (see later) between the two measures themselves. Our result suggests that AUC should replace accuracy in comparing learning algorithms in the future. Our result also prompts us to reevaluate 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 when compared by accuracy [8], [18], [19]. Since we will establish that AUC is a better measure, are those learning algorithms still very similar when compared by AUC? How does the recent Support Vector Machine (SVM) [2], [7], [31] compare to traditional learning algorithms, such as Naive Bayes and decision trees, in accuracy and AUC? We perform extensive experimental comparisons to compare Naive Bayes, decision trees, and SVM to answer these questions in Section 4. These kinds of new conclusions are very useful in machine learning and data mining.

Conclusions drawn in this paper may spur new research in machine learning. As a new measure (such as AUC) is discovered and proved to be better than a previous measure (such as accuracy), we can redesign most learning algorithms to optimize the new measure [12], [22]. 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 [22]. Results in the paper suggest that, in real-world applications of machine learning and data mining, we should use learning algorithms to optimize AUC instead of accuracy. Most learning algorithms today still optimize accuracy directly (or indirectly through entropy, for example) as their goals.

2 CRITERIA FOR COMPARING EVALUATION MEASURES

We start with some intuitions in comparing AUC and accuracy and then present formal definitions in comparing evaluation measures for learning algorithms. This section is based on our previous work [21].

---

[2] It can be shown easily that the AUC of ROC is equivalent to the area under the curve of the lift curve [20]. More precisely, there is a one-to-one mapping between AUC of ROC and AUC of the lift.

TABLE 1
An Example for Calculating AUC with r_i

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

TABLE 2
An Example in Which Two Classifiers Have the Same Classification Accuracy but Different AUC Values

Classifier 1	−	−	−	−	+	|	−	+	+	+	+
Classifier 2	+	−	−	−	−	|	+	+	+	+	−

2.1 AUC versus Accuracy

As we have seen, AUC is equivalent to the probability that a randomly chosen negative example will have a smaller estimated probability of belonging to the positive class than a randomly chosen positive example. Hand and Till [14] present the following simple approach to calculating AUC of a classifier for binary classification:[3]

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

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 the $i\mathrm{th}$ positive example in the ranked list. Table 1 shows an example of how to calculate AUC from a ranked list with five positive examples and five 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 (1) is a way to measure 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 [13].

On the other hand, the accuracy can easily be defined as $\frac{TP\cdot n_0 + TN\cdot n_1}{n_0+n_1}$, where TP is the true positive rate and TN is the true negative rate ($TN = 1 - FP$). 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 five of the 10 examples as positive and the other five as negative. If we rank the testing examples according to the increasing probability of being $+$ (positive), we get the two ranked lists in Table 2.

Clearly, both classifiers produce an accuracy of 80 percent (or an error rate of 20 percent with one false positive and one false negative) and, thus, the two classifiers are equivalent in terms of accuracy. 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 (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 tells us that Classifier 1 is indeed better than Classifier 2.

TABLE 3
A Counterexample in Which One Classifier Has a Higher AUC but a Lower Classification Accuracy

Classifier 3	−	−	−	+	+	|	−	−	+	+	+
Classifier 4	+	−	−	−	−	|	+	+	+	+	−

TABLE 4
A Counterexample in Which Two Classifiers Have the Same AUC but Different Classification Accuracies

Classifier 5	−	−	+	+	−	|	+	+	−	−	+
Classifier 6	−	−	+	+	+	|	−	−	+	−	+

Unfortunately, “counterexamples” 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 accuracy of Classifier 3 is 60 percent, while the accuracy of Classifier 4 is 80 percent (again, we assume that the threshold for accuracy is set at the middle so that five examples are predicted as positive and five as negative). Therefore, a larger AUC does not always imply a higher accuracy; that is, AUC and accuracy sometimes contradict to each other. Therefore, which one should we really “trust”?

Another intuitive argument for AUC against accuracy is that AUC is more discriminating than accuracy since it has more possible values. More specifically, given a data set with n examples, there is a total of only $n+1$ different classification accuracies ($0/n$, $1/n$, $\ldots$, 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 ($0/n_0n_1$, $1/n_0n_1$, $\ldots$, n_0n_1/n_0n_1), generally more than $n+1$. However, counterexamples also exist in this regard. Table 4 illustrates two classifiers with the same AUC but different accuracies. Here, we see that both Classifier 5 and Classifier 6 have the same AUC ($\frac{3}{5}$) but different accuracies (60 percent and 40 percent, respectively). In general, a measure with more values is not necessarily more discriminating. The two measures may have nothing to do with each other.

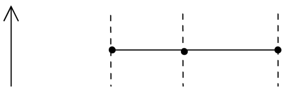
Last, there exist cases where neither AUC nor accuracy can tell the difference. Table 5 shows such an example. The AUC of both classifiers is $\frac{3}{5}$ and the accuracy of both classifiers is 60 percent. If two measures have too many such indistinguishable cases, they would not be considered

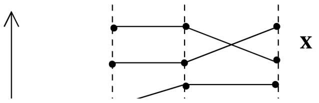
TABLE 5
An Example in Which Two Classifiers Have the Same AUC and the Same Classification Accuracy

Classifier 7	−	−	+	+	−	|	+	+	−	−	+
Classifier 8	−	−	+	+	−	|	+	−	+	+	−

---

[3] See Appendix A (available on the IEEE Computer Society’s Digital Library at http://www.computer.org/publications/dlib) for the detailed arguments about how the statistical meaning and calculation formula of AUC are obtained.





desirable. How often does this happen for AUC and accuracy?

In general, how do we compare different evaluation measures for learning algorithms? Some general criteria can be established, as we will show in the following sections.

2.2 (Strict) Consistency and Discriminancy

Intuitively speaking, when we discuss two different measures, f and g, on evaluating two learning algorithms, A and B, we want f and g to at least 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. Furthermore, 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, but not vise versa.[4]

This intuitive meaning of consistency and discriminancy can be made precise by the following definitions. We assume that Ψ is the domain of two functions, f and g, which return values as some performance measure.

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* $f(a) \neq f(b)$ *and* $g(a) = g(b)$, *and there exist no* $a, b \in \Psi$ *such that* $g(a) \neq g(b)$ *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 (A), but it is not possible to have students with different letter marks (such as A and B) but with the same numerical mark. This ideal example of a measure f (numerical mark) being strictly consistent and more discriminating than another g (letter mark) can be shown in the Fig. 2a.

2.3 Statistical Consistency and Discriminancy

As we have already seen in Section 2.1, counterexamples on consistency (Table 3) and discriminancy (Table 4) do exist for AUC and accuracy. Therefore, it is *impossible* to prove the consistency and discriminancy of AUC and accuracy based on Definition 1 and Definition 2. Fig. 2b 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 degrees 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 the *degree of* consistency and the *degree* of discriminancy, as follows:

Definition 3 (Degree of Consistency). *For two measures* f *and* g *on domain* Ψ, *let* $R = \{(a,b)|a,b \in \Psi, f(a) > f(b), g(a) > g(b)\}$ *and* $S = \{(a,b)|a,b \in \Psi, f(a) > f(b), g(a) < g(b)\}$. *The degree of consistency*[5] *of* f *and* g *is* $\mathbf{C}$ $(0 \leq \mathbf{C} \leq 1)$, *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* $\mathbf{D} = \frac{|P|}{|Q|}$.

As we have seen in Table 5 and Fig. 2b, there may exist cases where the two measures cannot tell the difference. The frequency of such cases is the Degree of Indifferency defined below.

Definition 5 (Degree of Indifferency). *For two measures* f *and* g *on domain* Ψ, *let* $V = \{(a,b)|a,b \in \Psi, a \neq b, f(a) = f(b), g(a) = g(b)\}$, $U = \{(a,b)|a,b \in \Psi, a \neq b\}$. *The degree of indifferency for* f *and* g *is* $\mathbf{E} = \frac{|V|}{|U|}$.

---

[4] As we have already seen in Section 2.1, counterexamples on strict consistency and discriminancy do exist for AUC and accuracy. See Section 2.3 for definitions on *statistical* consistency and discriminancy between two measures.

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