Constructing New and Better Evaluation Measures for Machine Learning

Jin Huang
School of Information Technology and Engineering
University of Ottawa
Ottawa, Ontario, Canada K1N 6N5
jhuang33@site.uottawa.ca

Charles X. Ling
Department of Computer Science
The University of Western Ontario
London, Ontario, Canada N6A 5B7
cling@csd.uwo.ca

Abstract

Evaluation measures play an important role in machine learning because they are used not only to compare different learning algorithms, but also often as goals to optimize in constructing learning models. Both formal and empirical work has been published in comparing evaluation measures. In this paper, we propose a general approach to construct new measures based on the existing ones, and we prove that the new measures are consistent with, and finer than, the existing ones. We also show that the new measure is more correlated to RMS (Root Mean Square error) with artificial datasets. Finally, we demonstrate experimentally that the greedy-search based algorithm (such as artificial neural networks) trained with the new and finer measure usually can achieve better prediction performance. This provides a general approach to improve the predictive performance of existing learning algorithms based on greedy search.

1 Introduction

Evaluation measures are widely used in machine learning and data mining to compare different learning algorithms. They are also often used as objective functions to construct learning models. For example, accuracy is a common evaluation measure in machine learning. It has been widely used to compare predictive performance of many learning algorithms including decision trees, neural networks, and naive Bayes (e.g., [Kononenko, 1990]), and it is also the goal of optimization in constructing decision trees [Quinlan, 1993]. In recent years, the ROC (Receiver Operating Characteristics) originated in signal detection [Green and Swets, 1966] has been introduced into machine learning [Provost and Fawcett, 1997; Provost *et al.*, 1998]. The area under the ROC curve, or simply *AUC*, has been proposed in comparing learning algorithms [Provost and Domingos, 2003] and constructing or optimizing learning models [Ferri *et al.*, 2002; Rakotomamonjy, 2004].

[Bradley, 1997] experimentally compares popular machine learning algorithms using both accuracy and *AUC*, and finds that *AUC* exhibits several desirable properties compared to accuracy. For example, *AUC* has increased sensitivity in Analysis of Variance (ANOVA) tests, is independent of the decision threshold, and is invariant to *a priori* class probability distributions. Recently, [Caruana and Niculescu-Mizil, 2004] empirically compare nine commonly used machine learning measures in terms of the correlation relations. They find that RMS (Root Mean Square error) is most correlated with the other measures on average, and thus it is a robust measure in uncertain situations. They also construct a new measure called SAR by averaging the measures of accuracy, *AUC*, and RMS, and claim that this new measure works better than RMS.

In [Ling *et al.*, 2003] two formal criteria are established to compare evaluation measures. They are called *(statistical) consistency*, and *(statistical) discriminancy*. They formally prove that *AUC* is consistent with, and more discriminant (or *finer*) than accuracy, for the binary balanced datasets (which have the same number of positive and negative examples).

In this paper we propose a general method to construct new measures based on existing ones, and prove that they are consistent with and finer than the existing ones. [1] Further, we will show experimentally that the newly constructed measure is more correlated to RMS than the existing ones. With a consistent and finer measure, a greedy hill-climbing learning algorithm (such as neural networks) would search better (i.e., not likely to get stuck in a flat plateau) and find better optimal solutions. We conduct experiments to show that neural networks optimized by the new measure predict better than the ones optimized by the existing ones. This illustrates the usefulness of designing new and finer measures for machine learning, as it provides a general method to improve learning algorithms using greedy search.

2 Review of Formal Criteria for Comparing Measures

In [Ling *et al.*, 2003] the *degree of consistency* and *degree of discriminancy* of two measures are proposed and defined. The degree of consistency between two measures f and g, denoted as $\mathbf{C_{f,g}}$, is simply the fraction (probability) that two measures are consistent over some distribution of the instance

[1] We normalize all measures in our discussion to be in [0, 1], with 0 to be the worst, and 1 to be the best. For this reason, the accuracy is used instead of the error rate. Also RMS (Root Mean Square error) in this paper is actually $(1 - RMS)$.

space. Two measures are consistent when comparing two objects a and b, if f stipulates that a is better than b, g also stipulates that a is better than b.

Using the example of AUC (as f) and accuracy (as g) on two ranked lists (as a and b), most of the time AUC and accuracy do agree on each other (i.e., they are consistent). However, there are exceptions when AUC and accuracy contradict. Table 1 lists two ranked lists of 10 testing examples, [2] presumably as the result of the prediction by two learning algorithms. The AUC of the ranked list a is $\frac{21}{25}$, [3] and the AUC of the ranked list b is $\frac{16}{25}$. Thus the ranked list a is better than the ranked list b according to AUC. But assuming that both learning algorithms classify half (the right most 5) of the 10 examples as positive, and the other 5 as negative. The accuracy of a is 60%, and the accuracy of b is 80%. Therefore, b is better than a according to accuracy. Clearly, AUC and accuracy are inconsistent here. Again the probability that two measures f and g are consistent is defined as degree of consistency $\mathbf{C_{f,g}}$. [Ling *et al.*, 2003] define that two measures f and g are *consistent* iff the degree of consistency $\mathbf{C_{f,g}} > 0.5$. That is, f and g are consistent if they agree with each other on over half of the cases.

Table 1: A counter example in which AUC and accuracy are inconsistent.

a	−	−	−	+	+	−	−	+	+	+
b	+	−	−	−	−	+	+	+	+	−

The *degree of discriminancy* of f over g, denoted as $\mathbf{D_{f/g}}$, is defined as the ratio of cases where f can tell the difference but g cannot, over the cases where g can tell the difference but f cannot. Using AUC (as f) and accuracy (as g) as example again. There are many cases in which AUC can tell the difference between two ranked lists but accuracy cannot. This is partially due to the fact that AUC has many more values than accuracy. But counter examples also exist in which accuracy can tell the difference but AUC cannot. Table 2 shows such a counter example. We can obtain that both ranked lists have the same AUC ($\frac{3}{5}$) but different accuracies (60% and 40% respectively). [Ling *et al.*, 2003] define that a measure f is *more discriminant* (or *finer*) than g iff $D_{f/g} > 1$. That is, f is finer than g if there are more cases where f can tell the difference but g cannot, than g can tell the difference but f cannot.

In the next section we will propose a general approach to construct new measures that are provably consistent with and finer than the existing ones.

[2] The domain of AUC and accuracy is the ranked lists of labeled examples, ordered according to the increasing probability of being positive. Almost all classification learning algorithms, such as decision trees, naive Bayes, support vector machines, and neural networks produce probability estimations on the classification which can be used to rank testing examples.

[3] The AUC can be calculated by the formula [Hand and Till, 2001] $AUC = \frac{\sum_{i=1}^{n_0}(r_i - i)}{n_0 n_1}$, where n_0 and n_1 are the number of positive and negative examples (both 5 here) respectively, and r_i is the position of the ith positive example.

Table 2: A counter example in which two ranked lists have the same AUC but different accuracies

a	−	−	+	+	−	+	+	−	−	+
b	−	−	+	+	+	−	−	+	−	+

3 Constructing New and Better Measures

First of all, we show formally that the finer relation is transitive (while the consistent relation is not; a counter example can be easily given). We use $f \succ g$ to denote that f is finer than g. The following theorem proves that the finer relation is transitive.

Theorem 1 *For measures f, g, and h, if $f \succ g$ and $g \succ h$, then $f \succ h$.*

Proof: Let Ψ be the set of the objects to be evaluated, $\Gamma = \Psi \times \Psi$, $A, B, X_1, X_2, Y_1, Y_2 \subset \Gamma$. In the following definitions we use "$f =$" to represent "$f(a) = f(b)$", "$g \neq$" to represent "$g(a) \neq g(b)$" etc. We define $A = \{(a,b)|a,b, \in \Psi, f \neq, g =\}$, $B = \{(a,b)|f =, g \neq\}$, $X_1 = \{(a,b)|f \neq, g \neq, h =\}$, $X_2 = \{(a,b)|f =, g \neq, h =\}$, $Y_1 = \{(a,b)|f \neq, g =, h \neq\}$, $Y_2 = \{(a,b)|f =, g =, h \neq\}$. Then clearly $D_{f/g} = \frac{|A|}{|B|} > 1$, $D_{g/h} = \frac{|X_1|+|X_2|}{|Y_1|+|Y_2|} > 1$, and $D_{f/h} = \frac{|A|-|Y_1|+|X_1|}{|B|-|X_2|+|Y_2|}$. Since we have $|A| > |B|$, and $|X_1| + |X_2| > |Y_1| + |Y_2|$, thus $|A| - |Y_1| + |X_1| > |B| - |X_2| + |Y_2|$, $D_{f/h} > 1$. □

We propose a general approach to construct a "two-level measure", denoted as f:g, based on two existing measures f and g. Intuitively, f:g is a new measure where f is used first as a "dominant" measure in comparison. If f ties in the comparison, then g would be used as a tie breaker. We can formally define the two-level measure f:g as follows.

Definition 1 *A two-level measure ϕ formed by f and g, denoted by f:g, is defined as:*

- $\phi(a) > \phi(b)$ *iff* $f(a) > f(b)$, *or* $f(a) = f(b)$ *and* $g(a) > g(b)$;
- $\phi(a) = \phi(b)$ *iff* $f(a) = f(b)$ *and* $g(a) = g(b)$.

When using AUC and accuracy as two existing measures, if AUC values of the two ranked lists are different, then the new two-level measure AUC:acc agrees with AUC, no matter what the value of accuracy is. But if AUC values are the same, then the two-level measure agrees with accuracy. Our new measure AUC:acc is different from the new measure SAR proposed in [Caruana and Niculescu-Mizil, 2004] as ours is not a simple linear combination of existing measures, and is still a measure for ranked lists with class labels.

The following theorem proves that the two-level measure defined is consistent with and finer than the existing measures.

Theorem 2 *Let $\phi = f$:g be the two-level measure formed by f and g, $f \succ g$, and $\mathbf{D}_{f/g} \neq \infty$. Then $\mathbf{C}_{\phi,f} = 1$, and $\mathbf{D}_{\phi/f} = \infty$. In addition, $\mathbf{C}_{\phi,g} \geq \mathbf{C}_{f,g}$, and $\mathbf{D}_{\phi/g} = \infty$. That is, ϕ is a finer measure than both f and g; i.e., $\phi \succ f \succ g$.*

Proof: Let $A = \{(a,b)|f$:$g(a) > f$:$g(b), f(a) < f(b)\}$. By Definition 1, $A = \Phi$. Therefore $\mathbf{C}_{\phi,f} = 1$. Let

$B = \{(a,b)|f(a) = f(b), g(a) > g(b)\}$, $C = \{(a,b)|f(a) > f(b), g(a) > g(b)\}$, $D = \{(a,b)|f(a) > f(b), g(a) < g(b)\}$, Then $\mathbf{C}_{\phi,g} = \frac{|B|+|C|}{|B|+|C|+|D|}$, $\mathbf{C}_{f,g} = \frac{|C|}{|C|+|D|}$. Thus $\mathbf{C}_{\phi,g} \geq \mathbf{C}_{f,g}$. For discriminancy there does not exist $a,b \in \Psi$ such that "f:$g(a) = f$:$g(b)$ and $f(a) > f(b)$". Since $\mathbf{D}_{f/g} > 1$, $\mathbf{D}_{f/g} \neq \infty$, there exists $a,b \in \Psi$ such that "$f(a) = f(b)$ and $g(a) > g(b)$" which is equivalent to "f:$g(a) \neq f$:$g(b)$ and $f(a) = f(b)$". Therefore $\mathbf{D}_{\phi/f} = \infty$, similarly we have $\mathbf{D}_{\phi/g} = \infty$. □

To confirm Theorem 2 when it applies to the two-level measure *AUC*:*acc*, we conduct experiment to compute the degree of consistency and discriminancy between the *AUC*:*acc* and *AUC* (and *acc*). This also gives us an intuition for the degrees of the consistency and discriminancy between *AUC*:*acc*, *AUC* and *acc*.

To conduct the experiment, we exhaustively enumerate all possible pairs of ranked lists with 6, 8, 10, 12, 14, and 16 examples of artificial datasets with an equal number of positive and negative examples.[4] The two criteria are computed, and the results are shown in Tables 3. Clearly, we can see from the table that $\mathbf{C}_{\phi,\mathbf{AUC}} = 1$, and $\mathbf{D}_{\phi/\mathbf{AUC}} = \infty$. Similarly, we can see that $\mathbf{C}_{\phi,\mathbf{acc}} > \mathbf{C}_{\mathbf{AUC},\mathbf{acc}}$, and $\mathbf{D}_{\phi/\mathbf{acc}} = \infty$. These confirm Theorem 2.

Table 3: Compare the two-level measure ϕ=*AUC*:*acc* with *AUC* and *acc*.

#	$C_{AUC,acc}$	$C_{\phi,AUC}$	$D_{\phi/AUC}$	$C_{\phi,acc}$	$D_{\phi/acc}$
6	0.991	1	∞	0.992	∞
8	0.977	1	∞	0.978	∞
10	0.963	1	∞	0.964	∞
12	0.951	1	∞	0.953	∞
14	0.942	1	∞	0.943	∞
16	0.935	1	∞	0.936	∞

One might think that we could construct a more discriminant "three-level" measure (such as $(f:g):f$) from the newly formed two-level measure f:g and an original measure f or g, and this process could repeat to get finer and finer measures. However, this will not work. Recall that in Theorem 2 one of the conditions to construct a finer two-level measure $\phi = f$:g is that $\mathbf{D}_{f/g} \neq \infty$. However, Theorem 2 proves that $\mathbf{D}_{\phi/f} = \mathbf{D}_{\phi/g} = \infty$, making it impossible for ϕ to be combined with f or g for further constructing new measures. Therefore, we can only use this method of constructing a two-level measure *once* from two existing measures.

This general method of constructing new, consistent, and finer measures is useful in evaluating learning algorithms. For example, when comparing two learning algorithms, if *AUC* is the same on a testing set, then we compare the accuracy to see which one is better. This gives rise to a finer evaluation measure in comparing learning algorithms than using *AUC* or accuracy alone. Another advantage of discovering a finer measure is that many learning algorithms build a model by optimizing some measure using hill climbing greedy search. A consistent and finer measure will guide greedy search better as it is less likely to stop prematurely in a flat plateau. We will discuss this later in the paper.

In the next section, we will experimentally compare the new measure *AUC*:*acc* with RMS, and show that it is more correlated with RMS than *AUC* and accuracy.

4 Comparing the New Measure to RMS

As indicated by [Caruana and Niculescu-Mizil, 2004], given true probabilities of examples, RMS (Root Mean Square error) [Kenney and Keeping, 1962] is shown to be the most reliable measure when the best measure is unknown. In this section, we use artificial data generated with known true probabilities to show empirically that the newly constructed measure *AUC*:*acc* is slightly more correlated with RMS than *AUC*, and significantly more correlated with RMS than accuracy.

We first randomly generate pairs of "true" ranked lists and perturbed ranked lists. The "true" ranked list always consists of n binary examples, with the i-th example having the probability of $p_i = \frac{i}{n}$ of belonging to the positive class. We then generate a perturbed ranked list by randomly fluctuating the probability of each example within a range bounded by ε. That is, if the true probability is p, the perturbed probability is randomly distributed in $[max(0, p_i - \varepsilon), min(1, p_i + \varepsilon)]$. Table 4 shows an example of the "true" and perturbed ranked lists with 10 examples. Examples with probabilities greater than 0.5 are regarded as positive, otherwise as negative. From this table the values of *RMS*, *AUC*, *acc*, and *AUC*:*acc* compared to the "true" ranked list can be easily computed as 0.293, 0.68, 0.6, 0.686 and 0.657 respectively.

Table 4: An example of "true" (T) and perturbed (P) ranked lists.

T	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
	−	−	−	−	−	+	+	+	+	+
P	0.0	0.15	0.6	0.5	0.95	0.2	0.65	0.7	1.0	0.4
	−	−	+	−	+	−	+	+	+	−

After we generate 200 tuples of *RMS*, *AUC*, *acc*, and *AUC*:*acc*, we calculate the Pearson correlation coefficients [Edwards, 1976] of *AUC*, *acc*, and *AUC*:*acc* compared to *RMS*[5]. We also vary the value of ε, and repeat this process five times. The averaged correlation coefficients are listed in Table 5. We perform a two-tailed paired t-test with 95% confidence interval to see whether the differences in these correlation coefficients are statistically significant. The values in bold means that they are significantly indifferent with the largest values in each row, but they are significantly larger than the values not in bold.

Several interesting conclusions can be drawn from the table. First of all, the correlation coefficients of *AUC* and *AUC*:*acc* are all significantly greater than that of *acc* with all ε values. It indicates that these measures are more correlated with RMS than accuracy. Secondly, we can see that when ε

[4] Artificial datasets are used for the uniform distribution of the instance space.

[5] Again, *AUC*, *acc*, and *AUC*:*acc* are all compared to $(1 - RMS)$ as all measures are normalized to be in [0, 1], with 0 as the worst and 1 as the best predictive performance.

Table 5: Comparing correlation coefficients of *acc*, *AUC*, *AUC*:*acc* with *RMS*.

	acc	AUC	AUC:acc
$\varepsilon = 0.3$	0.2454±0.076	**0.3222±0.072**	**0.3177±0.072**
$\varepsilon = 0.5$	0.4678±0.071	**0.535±0.063**	**0.5352±0.064**
$\varepsilon = 0.7$	0.5932±0.01	**0.6596±0.015**	**0.6616±0.014**
$\varepsilon = 0.8$	0.6546±0.051	0.6996±0.041	**0.7036±0.041**
$\varepsilon = 0.9$	0.6656±0.024	0.7137±0.025	**0.7168±0.025**

is small (0.3 and 0.5), there is no significant difference in correlation coefficients of *AUC* and *AUC*:*acc*. But with the increasing value of ε, our newly constructed measure *AUC*:*acc* becomes significantly more correlated with RMS than *AUC*. Thirdly, when ε is small (0.3), the values of all correlation coefficients are relatively small (0.2454), but when ε is large (0.9), the values are larger (0.7188). This can be expected as when the perturbation (ε) is small, there can often be no change in ranking ($AUC = 1$) and accuracy ($acc = 1$), while RMS is not 0. Thus the values of *AUC* and *acc* do not correlate well with RMS. When the perturbation (ε) is large, the rank list (*AUC*) and accuracy are both affected.

In sum, from the experiments conducted in this section, we can conclude that *AUC*:*acc* is slightly more correlated with RMS than *AUC*, which is significantly more correlated than accuracy.

5 Building Better Models with Finer Measures

In Section 3, we showed that the two-level measure *AUC*:*acc* is finer than *AUC* (which is in turn finer than accuracy). That is, $AUC{:}acc \succ AUC \succ acc$. As we have discussed earlier, a significant advantage of discovering consistent and finer measures is that they can be used in building learning models (such as classifiers) by optimizing the finer measures directly with greedy hill-climbing search. Intuitively, greedy search will less likely to stop prematurely with a long flat plateau if a finer measure is used for optimization. In this section, we will show that by maximizing *AUC*:*acc* or *AUC*, we will get better prediction than by maximizing the accuracy. We will conduct our experiments using artificial neural networks (ANNs). This is because ANNs are a typical hill-climbing greedy search, and are much more sensitive to small changes in the optimization process by producing different weights. On the other hand, decision trees, for example, may not be sensitive enough to changes in the attribute selection criterion.

Essentially we want to train three ANNs with the same training data optimized using *AUC*:*acc*, *AUC*, and *acc* respectively. For simplicity, we call the three ANN models $\text{ANN}_{AUC:acc}$, ANN_{AUC}, and ANN_{acc} respectively. Then we test these three ANN models on the testing sets. The predictive performance of the three different learning models on the test sets are measured by *AUC*:*acc*, *AUC*, and *acc*. We do this many times (using a 5-fold cross-validation) to obtain the average on testing *AUC*:*acc*, testing *AUC*, and testing accuracy. What we expect to see is that the model optimized by *AUC*:*acc* predicts better than the model optimized by *AUC*, measured by all of the three measures (*AUC*:*acc*, *AUC*, and *acc*). Similarly, the model optimized by *AUC* would be better than the model optimized by accuracy.

To optimize ANN with a measure f (f is either *AUC*:*acc*, *AUC*, or *acc* here), we implement the following simple optimization process: We still use the standard back-propagation algorithm that minimizes the sum of the squared differences (same as the RMS error) as it is the most robust and "strict" measure, but we monitor the change in f instead to decide when to stop training. [6] More specifically, we save the current weights in the neural network, and look ahead and train the network for N more epochs, and obtain the new f value. If the difference between the two f values is larger than a pre-selected threshold ε, it indicates that the neural network is still improving according to f, so we save the new weights (after training N epochs) as the current best weights, and the process repeats. If the difference between the two f values is less than ε, it indicates that the neural network is not improving according to f (a long flat plateau), so the training stops, and the saved weights are used as the final weights for the neural network optimized by f.

We choose $\varepsilon = 0.01$ and $N = 100$. We choose 20 real-world datasets from the UCI Machine Learning Repository [Blake and Merz, 1998]. The properties of these datasets are shown in Table 6. Each dataset is split into training and test sets using 5-fold cross-validation.

Table 6: Descriptions of the datasets used in our experiments.

Dataset	Instances	Attributes	Class
anneal	898	39	6
autos	205	26	7
breast-c.	286	10	2
cars	700	7	2
colic	368	23	2
credit-a	690	16	2
diabetes	768	9	2
eco	336	7	8
Glass	214	10	7
heart-c	303	14	5
hepatitis	155	20	2
ionosph.	351	35	2
p.-tumor	339	18	21
pima	392	8	2
segment	2310	20	7
sonar	208	61	2
soybean	683	36	19
splice	3190	62	3
vehicle	846	19	4
vote	435	17	2

The predictive performance on the testing sets from the three models $\text{ANN}_{AUC:acc}$, ANN_{AUC}, and ANN_{acc} is shown in Table 7.

[6] There are few previous works, such as [Herschtal and Raskutti, 2004], that directly optimize *AUC* in learning.

Table 7: Predictive results from the three ANNs optimized by *AUC*:*acc*, *AUC*, and accuracy.

Dataset	Model	acc	*AUC*	*AUC*:*acc*
anneal	$ANN_{AUC:acc}$	0.9631	0.9462	0.9558
	ANN_{AUC}	0.9434	0.9308	0.9402
	ANN_{acc}	0.9376	0.9055	0.9148
autos	$ANN_{AUC:acc}$	0.7880	0.9217	0.9296
	ANN_{AUC}	0.7843	0.8943	0.9021
	ANN_{acc}	0.7735	0.8809	0.8886
breast	$ANN_{AUC:acc}$	0.8432	0.6531	0.6615
	ANN_{AUC}	0.8446	0.6553	0.6637
	ANN_{acc}	0.8447	0.6527	0.6611
cars	$ANN_{AUC:acc}$	0.8686	0.9231	0.9318
	ANN_{AUC}	0.8643	0.9214	0.9301
	ANN_{acc}	0.7829	0.8782	0.8860
colic	$ANN_{AUC:acc}$	0.8239	0.8543	0.8625
	ANN_{AUC}	0.7821	0.8187	0.8265
	ANN_{acc}	0.8057	0.8087	0.8168
credit-a	$ANN_{AUC:acc}$	0.7058	0.6518	0.6589
	ANN_{AUC}	0.6936	0.6245	0.6314
	ANN_{acc}	0.7058	0.6520	0.6591
diabetes	$ANN_{AUC:acc}$	0.7509	0.8071	0.8146
	ANN_{AUC}	0.7608	0.8084	0.8160
	ANN_{acc}	0.7650	0.7936	0.8013
eco	$ANN_{AUC:acc}$	0.9488	0.9390	0.9485
	ANN_{AUC}	0.8497	0.9436	0.9521
	ANN_{acc}	0.9548	0.9458	0.9553
Glass	$ANN_{AUC:acc}$	0.5865	0.7908	0.7967
	ANN_{AUC}	0.5603	0.7752	0.7808
	ANN_{acc}	0.5298	0.7317	0.7369
heart-c	$ANN_{AUC:acc}$	0.7778	0.8201	0.8279
	ANN_{AUC}	0.7854	0.8163	0.8242
	ANN_{acc}	0.7778	0.8098	0.8176
hepatitis	$ANN_{AUC:acc}$	0.8305	0.8050	0.8133
	ANN_{AUC}	0.8305	0.8050	0.8133
	ANN_{acc}	0.8305	0.7503	0.7586
ionosph.	$ANN_{AUC:acc}$	0.9072	0.9538	0.9629
	ANN_{AUC}	0.9153	0.9622	0.9713
	ANN_{acc}	0.9047	0.9477	0.9567
primary-tumor	$ANN_{AUC:acc}$	0.4576	0.7751	0.7797
	ANN_{AUC}	0.4637	0.7803	0.7849
	ANN_{acc}	0.4505	0.7492	0.7537
pima	$ANN_{AUC:acc}$	0.7122	0.7311	0.7382
	ANN_{AUC}	0.7009	0.7038	0.7108
	ANN_{acc}	0.6233	0.6621	0.6683
segment	$ANN_{AUC:acc}$	0.9246	0.9927	1.00
	ANN_{AUC}	0.9063	0.9910	1.00
	ANN_{acc}	0.8806	0.9839	0.9927
sonar	$ANN_{AUC:acc}$	0.7419	0.8532	0.8606
	ANN_{AUC}	0.7203	0.8537	0.8609
	ANN_{acc}	0.6958	0.8710	0.8779
soybean	$ANN_{AUC:acc}$	0.9280	0.9923	1.00
	ANN_{AUC}	0.8872	0.9710	0.9799
	ANN_{acc}	0.8761	0.9229	0.9317
splice	$ANN_{AUC:acc}$	0.9546	0.9887	0.9982
	ANN_{AUC}	0.9533	0.9612	0.9707
	ANN_{acc}	0.9341	0.9253	0.9346
vehicle	$ANN_{AUC:acc}$	0.6804	0.8735	0.8803
	ANN_{AUC}	0.7019	0.8806	0.8876
	ANN_{acc}	0.6673	0.8299	0.8366
vote	$ANN_{AUC:acc}$	0.7627	0.6802	0.6878
	ANN_{AUC}	0.7586	0.6588	0.6664
	ANN_{acc}	0.7456	0.6324	0.6399
Average	$ANN_{AUC:acc}$	0.7978	0.8476	0.8556
	ANN_{AUC}	0.7853	0.8378	0.8457
	ANN_{acc}	0.7743	0.8167	0.8244

Table 8: Summary of experimental results in terms of AUC:acc, AUC, acc. Each cell indicates the number of win-draw-loss.

AUC:acc	$\mathbf{ANN}_{AUC}$	$\mathbf{ANN}_{acc}$
$\mathbf{ANN}_{AUC:acc}$	8-11-1	12-8-0
$\mathbf{ANN}_{AUC}$		10-9-1
AUC	$\mathbf{ANN}_{AUC}$	$\mathbf{ANN}_{acc}$
$\mathbf{ANN}_{AUC:acc}$	8-11-1	12-8-0
$\mathbf{ANN}_{AUC}$		10-9-1
acc	$\mathbf{ANN}_{AUC}$	$\mathbf{ANN}_{acc}$
$\mathbf{ANN}_{AUC:acc}$	8-12-0	11-8-1
$\mathbf{ANN}_{AUC}$		11-8-1

Note that in Table 7 the predictive results of different models on each dataset can only be compared vertically because it is not meaningful to compare results horizontally as values of accuracy, *AUC*, and *AUC*:*acc* are not comparable.

We perform a paired t-test with the 95% confidence level on each of the 20 datasets comparing the models of $ANN_{AUC:acc}$, ANN_{AUC} and ANN_{acc}, measured by *AUC*:*acc*, *AUC* and accuracy, respectively. We count in how many datasets that one model is statistically significantly better or worse than another model. The summary of these comparisons is listed in Table 8. The data in each cell indicates the win-draw-loss number of datasets that the model in the corresponding row over the model in the corresponding column. Several interesting conclusions can be drawn from the results in Table 7 and 8. Clearly, the result shows that the $ANN_{AUC:acc}$ model performs significantly better than ANN_{AUC} and ANN_{acc}, and ANN_{AUC} performs significantly better than ANN_{acc} in terms of the three different measures. When evaluated with *AUC*:*acc* (or *AUC*), $ANN_{AUC:acc}$ is significantly better than ANN_{AUC} (8 wins, 11 draws, 1 loss), and ANN_{AUC} is significantly better than ANN_{acc} (10wins, 9 draws, 1 loss). When evaluated with accuracy, $ANN_{AUC:acc}$ is significantly better than ANN_{AUC} (8 wins, 12 draws, 0 loss), and ANN_{AUC} is significantly better than ANN_{acc} (11wins, 8 draws, 1 loss). Therefore models optimized by *AUC*:*acc* are significantly better than models optimized by *AUC* and *acc*. This shows the advantage of using consistent and finer measures in model building by greedy search – optimizing consistent and finer measures lead to models with better predictions.

6 Discussions

The experimental results in the previous section show that the ANN model optimized by *AUC*:*acc* performs better than the ANN model optimized by accuracy even if they are both evaluated with accuracy. This is somewhat against a common intuition in machine learning that a model should be optimized by a measure that it will be measured on. However, some recent works have reported similar findings. For example, [Rosset, 2004] has compared the model selection performance of *AUC* and accuracy in highly uncertain situations. He shows that *AUC* is more likely to choose the correct model than accuracy, even if the model selection criterion is the model's future accuracy. We have conducted an extensive empirical

study that verifies [Rosset, 2004]'s conclusions. Further, we compare the model selection performance of nine evaluation measures, and show that, in general, a finer measure f is more likely to select better models than g even if the model is evaluated by g (submitted manuscript). This prompts us to believe that a finer measure may have an intrinsic advantage in model selection and model optimization.

We believe that finer measures also have advantages in hill climbing search algorithms, leading towards a better training model. The basic idea of hill climbing is to always head towards a state which is better than the current one. A heuristic measure is used to choose the best state from several candidate future states. As future states can be viewed as different learning models, the heuristics measure which chooses the best future state can be viewed as choosing the best future model. Since a finer measure is shown to be more likely to choose the better model, it is more likely to choose the better future state. Therefore, we can conclude that a finer heuristics measure is more likely to lead to a better learning model in hill climbing algorithms. The experimental results in the previous section can be well explained by this conclusion.

7 Conclusions and Future Work

Evaluation metrics are essential in machine learning and other experimental science and engineering areas. In this paper, we first review the formal criteria established in [Ling *et al.*, 2003] to compare the predictive performance of any two single-number measures. We then propose a general approach to construct new measures based on existing ones, and prove that the new measures are consistent with and finer than the existing ones. We compare experimentally the new measure *AUC*:*acc* with a best measure RMS, and show that it is more correlated with it than *AUC* and *acc*. Finally, we show that learning models optimized with hill-climbing by the new and finer measure predict better than models optimized by the existing ones.

In our future work, we plan to study how to construct new measures based on the popular evaluation measures in information retrieval and natural language processing. We will also re-design other popular learning algorithms by optimizing these new constructed measures.

References

[Blake and Merz, 1998] C.L. Blake and C.J. Merz. *UCI Repository of machine learning databases [http://www.ics.uci.edu/∼mlearn/MLRepository.html]*. Irvine, CA: University of California, Department of Information and Computer Science, 1998.

[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.

[Caruana and Niculescu-Mizil, 2004] Rich Caruana and Alexandru Niculescu-Mizil. Data mining in metric space: An empirical analysis of supervised learning performance criteria. In *Proceedings of the 10th ACM SIGKDD conference*, 2004.

[Edwards, 1976] Allen Louis Edwards. *An introduction to linear regression and correlation*. W. H. Freeman, 1976.

[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.

[Green and Swets, 1966] D.M. Green and J.A. Swets. *Signal Detection Theory and Psychophysics*. Wiley, New York, 1966.

[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.

[Herschtal and Raskutti, 2004] Alan Herschtal and Bhavani Raskutti. Optimising area under the ROC curve using gradient descent. In *Proceedings of the 21st International Conference on Machine Learning*, 2004.

[Kenney and Keeping, 1962] J. F. Kenney and E. S Keeping. *Mathematics of Statistics*. Princeton, NJ, 1962.

[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.

[Ling *et al.*, 2003] C. X. Ling, J. Huang, and H. Zhang. AUC: a statistically consistent and more discriminating measure than accuracy. In *Proceedings of 18th International Conference on Artificial Intelligence (IJCAI-2003)*, pages 519–526, 2003.

[Provost and Domingos, 2003] F. Provost and P. Domingos. Tree induction for probability-based ranking. *Machine Learning*, 52:3:199–215, 2003.

[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.

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

[Rakotomamonjy, 2004] Alain Rakotomamonjy. Optimizing AUC with SVMs. In *Proceedings of European Conference on Artificial Intelligence Workshop on ROC Curve and AI*, 2004.

[Rosset, 2004] Saharon Rosset. Model selection via the AUC. In *Proceedings of the 21st International Conference on Machine Learning*, 2004.