An Improved Model Selection Heuristic for AUC

Shaomin Wu[1], Peter Flach[2], and Cèsar Ferri[3]

[1] Cranfield University, United Kingdom
Shaomin.Wu@cranfield.ac.uk
[2] University of Bristol, United Kingdom
Peter.Flach@bristol.ac.uk
[3] Universitat Politècnica de València, Spain
cferri@dsic.upv.es

Abstract. The area under the ROC curve (AUC) has been widely used to measure ranking performance for binary classification tasks. AUC only employs the classifier's scores to rank the test instances; thus, it ignores other valuable information conveyed by the scores, such as sensitivity to small differences in the score values. However, as such differences are inevitable across samples, ignoring them may lead to overfitting the validation set when selecting models with high AUC. This problem is tackled in this paper. On the basis of ranks as well as scores, we introduce a new metric called *scored AUC* (sAUC), which is the area under the *sROC curve*. The latter measures how quickly AUC deteriorates if positive scores are decreased. We study the interpretation and statistical properties of sAUC. Experimental results on UCI data sets convincingly demonstrate the effectiveness of the new metric for classifier evaluation and selection in the case of limited validation data.

1 Introduction

In the data mining and machine learning literature, there are many learning algorithms that can be applied to build candidate models for a binary classification task. Such models can be decision trees, neural networks, naive Bayes, or ensembles of these models. As the performance of the candidate models may vary over learning algorithms, effectively selecting an optimal model is vitally important. Hence, there is a need for metrics to evaluate the performance of classification models.

The predicted outcome of a classification model can be either a class decision such as positive and negative on each instance, or a score that indicates the extent to which an instance is predicted to be positive or negative. Most models can produce scores; and those that only produce class decisions can easily be converted to models that produce scores [3,11]. In this paper we assume that the scores represent likelihoods or posterior probabilities of the positive class.

The performance of a classification model can be evaluated by many metrics such as recall, accuracy and precision. A common weakness of these metrics is that they are not robust to the change of the class distribution. When the ratio of positive to negative instances changes in a test set, a model may no longer perform optimally, or even acceptably. The ROC (Receiver Operating Characteristics) curve, however, is invariant to changes in class distribution. If the class distribution changes in a test set,

J.N. Kok et al. (Eds.): ECML 2007, LNAI 4701, pp. 478–489, 2007.

but the underlying class-conditional distributions from which the data are drawn stay the same, the ROC curve will not change. It is defined as a plot of a model's true positive rate on the y-axis against its false positive rate on the x-axis, and offers an overall measure of model performance, regardless of the different thresholds used. The ROC curve has been used as a tool for model selection in the medical area since the late 1970s, and was more recently introduced to evaluate machine learning algorithms [9,10].

The area under the ROC curve, or simply AUC, aggregates the model's behaviour for all possible decision thresholds. It can be estimated under parametric [13], semi-parametric [6] and nonparametric [5] assumptions. The nonparametric estimate of the AUC is widely used in the machine learning and data mining research communities. It is the summation of the areas of the trapezoids formed by connecting the points on the ROC curve, and represents the probability that a randomly selected positive instance will score higher than a randomly selected negative instance. It is equivalent to the Wilcoxon-Mann-Whitney (WMW) U statistic test of ranks [5]. Huang and Ling [8] argue that AUC is preferable as a measure for model evaluation over accuracy.

The nonparametric estimate of the AUC is calculated on the basis of the ranks of the scores. Its advantage is that it does not depend on any distribution assumption that is commonly required in parametric statistics. Its weakness is that the scores are only used to rank instances, and otherwise ignored. The AUC, estimated simply from the ranks of the scores, can remain unchanged even when the scores change. This can lead to a loss of useful information, and may therefore produce sub-optimal results.

In this paper we argue that, in order to evaluate the performance of binary classification models, both ranks and scores should be combined. A scored AUC metric is introduced for estimating the performance of models based on their original scores. The paper is structured as follows. Section 2 reviews ways to evaluate scoring classifiers, including AUC and Brier score, and gives a simple and elegant algorithm to calculate AUC. Section 3 introduces the *scored ROC curve* and the new *scored AUC* metric, and investigates its properties. In Section 4 we present experimental results on 17 data sets from the UCI repository, which unequivocally demonstrate that validation sAUC is superior to validation AUC and validation Brier score for selecting models with high test AUC when limited validation data is available. Section 5 presents the main conclusions and suggests further work. An early version of this paper appeared as [12].

2 Evaluating Classifiers

There are a number of ways of evaluating the performance of scoring classifiers over a test set. Broadly, the choices are to evaluate its *classification* performance, its *ranking performance*, or its *probability estimation* performance. Classification performance is evaluated by a measure such as accuracy, which is the proportion of test instances that is correctly classified. Probability estimation performance is evaluated by a measure such as mean squared error, also called the *Brier score*, which can be expressed as $\sum_x (\hat{p}(x) - p(x))^2$, where $\hat{p}(x)$ is the estimated probability for instance x, and $p(x)$ is 1 if x is positive and 0 if x is negative. The calculation of both accuracy and Brier score is an $O(n)$ operation, where n is the size of the test set.

Ranking performance is evaluated by sorting the test instances on their score, which is an $O(n\log n)$ operation. It thus incorporates performance information that neither accuracy nor Brier score can access. There are a number of reasons why it is desirable to have a good ranker, rather than a good classifier or a good probability estimator. One of the main reasons is that accuracy requires a fixed score threshold, whereas it may be desirable to change the threshold in response to changing class or cost distributions. Good accuracy obtained with one threshold does not imply good accuracy with another. Furthermore, good performance in both classification and probability estimation is easily obtained if one class is much more prevalent than the other. For these reasons we prefer to evaluate ranking performance. This can be done by constructing an ROC curve.

An ROC curve is generally defined as a piecewise linear curve, plotting a model's true positive rate on the y-axis against its false positive rate on the x-axis, evaluated under all possible thresholds on the score. For a test set with t test instances, the ROC curve will have (up to) t linear segments and $t+1$ points. We are interested in the area under this curve, which is well-known to be equivalent to the Wilcoxon-Mann-Whitney sum of ranks test, and estimates the probability that a randomly chosen positive is ranked before a randomly chosen negative. AUC can be calculated directly from the sorted test instances, without the need for drawing an ROC curve or calculating ranks, as we now show.

Denote the total number of positive instances and negative instances by m and n, respectively. Let $\{y_1,\dots,y_m\}$ be the scores predicted by a model for the m positives, and $\{x_1,\dots,x_n\}$ be the scores predicted by a model for the n negatives. Assume both y_i and x_j are within the interval $[0,1]$ for all $i=1,2,...,m$ and $j=1,2,...,n$; high scores are interpreted as evidence for the positive class. By a slight abuse of language, we will sometimes use positive (negative) score to mean 'score of a positive (negative) instance'.

AUC simply counts the number of pairs of positives and negatives such that the former has higher score than the latter, and can therefore be defined as follows:

$$\hat{\theta} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\psi_{ij} \tag{1}$$

where ψ_{ij} is 1 if $y_i - x_j > 0$, and 0 otherwise. Let Z_a be the sequence produced by merging $\{y_1,\dots,y_m\}$ and $\{x_1,\dots,x_n\}$ and sorting the merged set in ascending order (so a good ranker would put the positives after the negatives in Z_a), and let r_i be the rank of y_i in Z_a. Then AUC can be expressed in terms of ranks as follows:

$$\hat{\theta} = \frac{1}{mn}\left(\sum_{i=1}^{m} r_i - \frac{m(m+1)}{2}\right) = \frac{1}{mn}\sum_{i=1}^{m}(r_i - i) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{t=1}^{r_i - i} 1 \tag{2}$$

Here, $r_i - i$ is the number of negatives before the ith positive in Z_a, and thus AUC is the (normalised) sum of the number of negatives before each of the m positives in Z_a.

Dually, let Z_d be the sequence produced by sorting $\{y_1,\dots,y_m\}\cup\{x_1,\dots,x_n\}$ in descending order (so a good ranker would put the positives before the negatives in Z_d). We then obtain

$$\hat{\theta} = \frac{1}{mn}\sum_{j=1}^{n}(s_j - j) = \frac{1}{mn}\sum_{j=1}^{n}\sum_{t=1}^{s_j - j} 1 \tag{3}$$

Table 1. Column-wise algorithm for calculating AUC

```
Inputs: $m$ positive and $n$ negative test instances, sorted by decreasing score;
Outputs: $\hat{\theta}$: AUC value of the model;
Algorithm:
    1: Initialise: $AUC \leftarrow 0$, $c \leftarrow 0$
    2: for each consecutive instance in the ranking do
    3:     if the instance is positive then
    4:         $c \leftarrow c + 1$
    5:     else
    6:         $AUC \leftarrow AUC + c$
    7:     end if
    8: end for
    9: $\hat{\theta} \leftarrow \frac{AUC}{mn}$
```

where s_j is the rank of x_j in Z_d, and $s_j - j$ is the number of positives before the jth negative in Z_d. From this perspective, AUC represents the normalised sum of the number of positives before each of the n negatives in Z_d.

From Eq. (3) we obtain the algorithm shown in Table 1 to calculate the value of the AUC. The algorithm is different from other algorithms to calculate AUC (e.g., [4]) because it doesn't explicitly manipulate ranks. The algorithm works by calculating AUC column-wise in ROC space, where c represents the (un-normalised) height of the current column. For simplicity, we assume there are no ties (this can be easily incorporated by reducing the increment of AUC in line 6). A dual, row-wise algorithm using the ascending ordering Z_a can be derived from Eq. (2). Alternatively, we can calculate the *Area Over the Curve (AOC)* row-wise using the descending ordering, and set $\hat{\theta} \leftarrow \frac{mn-AOC}{mn}$ at the end.

3 sROC Curves and Scored AUC

Our main interest in this paper is to select models that perform well as rankers. To that end, we could simply evaluate AUC on a validation set and select those models with highest AUC. However, this method may suffer from overfitting the validation set whenever small difference in the score values lead to considerable differences in AUC.

Example 1. Two models, M_1 and M_2, are evaluated on a small test set containing 3 positives and 3 negatives. We obtain the following scores:

$$M_1 : 1.0{+}\ 0.7{+}\ 0.6{+}\ 0.5{-}\ 0.4{-}\ 0.0{-}$$
$$M_2 : 1.0{+}\ 0.9{+}\ 0.6{-}\ 0.5{+}\ 0.2{-}\ 0.0{-}$$

Here, for example, $0.7+$ means that a positive instance receives a score of 0.7, and $0.6-$ means that a negative instance receives a score of 0.6. In terms of AUC, M_1 achieves the perfect ranking, while M_2 has $AUC = 8/9$. In terms of Brier score, both models perform equally, as the sum of squared errors is 0.66 in both cases, and the mean squared error is 0.11. However, one could argue that M_2 is preferable as its ranking is much less sensitive

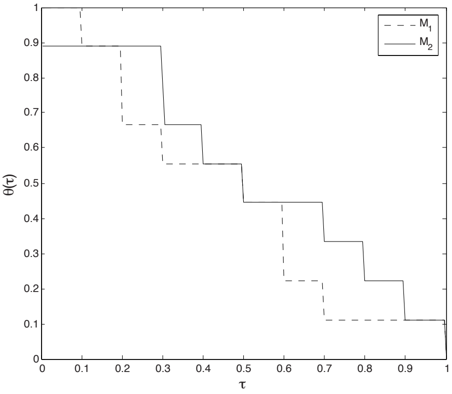

Fig. 1. sROC curves for example models M_1 and M_2 from Example 1

to drift of the scores. For instance, if we subtract 0.25 from the positive scores, the AUC of M_1 decreases to $6/9$, but the AUC of M_2 stays the same.

In order to study this more closely, we introduce the following parameterised version of AUC.

Definition 1. *Given a* margin τ *with* $0 \leq \tau \leq 1$, *the* margin-based AUC *is defined as*

$$\hat{\theta}(\tau) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\psi_{ij}(\tau) \tag{4}$$

where $\psi_{ij}(\tau)$ *is 1 if* $y_i - x_j > \tau$, *and 0 otherwise.*

Clearly, $\hat{\theta}(0) = \hat{\theta}$, and $\hat{\theta}(1) = 0$. More generally, $\hat{\theta}(\tau)$ is a non-increasing step function in τ. It has (up to) *mn* horizontal segments. For a given τ, $\hat{\theta}(\tau)$ can be interpreted as the AUC resulting from decreasing all positive scores with τ (or increasing all negative scores with τ). Figure 1 plots $\hat{\theta}(\tau)$ of models M_1 and M_2 from Example 1. It is clear that the margin-based AUC of M_1 deteriorates more rapidly with τ than that of M_2, even though its initial AUC is higher. We call such a plot of $\hat{\theta}(\tau)$ against τ an *sROC curve*.

Consider the area under the sROC curve, denoted by $\hat{\theta}_s$. This is a measure of how rapidly the AUC deteriorates with increasing margin. It can be calculated without explicitly constructing the sROC curve, as follows.

Theorem 1. $\hat{\theta}_s = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(y_i - x_j)\psi_{ij}$.

Proof

$$\begin{aligned}\hat{\theta}_s &= \int_0^1 \hat{\theta}(\tau)d\tau = \int_0^1 \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\psi_{ij}(\tau)d\tau \\ &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\int_0^1 \psi_{ij}(\tau)d\tau = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(y_i - x_j)\psi_{ij}\end{aligned} \tag{5}$$

Thus, just as $\hat{\theta}$ is the area under the ROC curve, $\hat{\theta}_s$ is the area under the sROC curve; we call it *scored AUC (sAUC)*. An equivalent definition of sAUC was introduced in [12], and independently by Huang and Ling, who refer to it as *soft AUC* [7]. Its interpretation as the area under the sROC curve is, to the best of our knowledge, novel. The sROC curve depicts the stability of AUC with increasing margins, and sAUC aggregates this over all margins into a single statistic.

Whereas in Eq. (1) the term ψ_{ij} is an indicator that only reflects the ordinal comparison between the scores, $(y_i - x_j)\psi_{ij}$ in Eq. (5) measures how much y_i is larger than x_j. Notice that, by including the ordinal term, we combine information from both ranks and scores. Indeed, if we omit ψ_{ij} from Eq. (5) the expression reduces to $\frac{1}{m}\sum_{i=1}^{m} y_i - \frac{1}{n}\sum_{i=1}^{n} x_i = M^+ - M^-$; i.e., the difference between the mean positive and negative scores. This measure (a quantity that takes scores into account but ignores the ranking) is investigated in [2].

We continue to analyse the properties of sAUC. Let $R^+ = \frac{1}{m}\sum_{i=1}^{m}\frac{r_i - i}{n} y_i$ be the weighted average of the positive scores, weighted by the proportion of negatives that are correctly ranked relative to each positive. Similarly, let $R^- = \frac{1}{n}\sum_{j=1}^{n}\frac{s_j - j}{m} x_j$ be the weighted average of the negative scores, weighted by the proportion of positives that are correctly ranked relative to each negative (i.e., the height of the column under the ROC curve). We then have the following useful reformulation of sAUC.

Theorem 2. $\hat{\theta}_s = R^+ - R^-$.

Proof

$$\hat{\theta}_s = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(y_i - x_j)\psi_{ij} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} y_i\psi_{ij} - \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} x_j\psi_{ij} =$$

$$= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{r_i - i} y_i - \frac{1}{mn}\sum_{j=1}^{n}\sum_{i=1}^{s_j - j} x_j = \frac{1}{m}\sum_{i=1}^{m}\frac{r_i - i}{n} y_i - \frac{1}{n}\sum_{j=1}^{n}\frac{s_j - j}{m} x_j = R^+ - R^-$$

This immediately leads to the algorithm for calculating $\hat{\theta}_s$ in Table 2. The algorithm calculates R^+ column-wise as in the AUC algorithm (Table 1), and the complement of R^- row-wise (so that the descending ordering can be used in both cases).

Example 2. Continuing Example 1, we have

M_1**:** $R^+ = 0.77$, $R^- = 0.3$ and $\hat{\theta}_s = 0.47$;
M_2**:** $R^+ = 0.74$, $R^- = 0.2$ and $\hat{\theta}_s = 0.54$.

We thus have that M_2 is somewhat better in terms of sAUC than M_1 because, even though its AUC is lower, it is robust over a larger range of margins.

The following theorems relate $\hat{\theta}_s$, $\hat{\theta}$ and $M^+ - M^-$.

Theorem 3. *(1)* $R^+ \leq M^+$ *and* $R^- \leq M^-$.
(2) $M^+ - M^- \leq \hat{\theta}_s \leq \hat{\theta}$.

Table 2. Algorithm for calculating sAUC

```
Inputs: $m$ positive and $n$ negative test instances, sorted by decreasing score;
Outputs: $\hat{\theta}_s$: scored AUC;
Algorithm:
    Initialise: $AOC \leftarrow 0, AUC \leftarrow 0, r \leftarrow 0, c \leftarrow 0$
    for each consecutive instance with score $s$ do
        if the instance is positive then
            $c \leftarrow c+s$
            $AOC \leftarrow AOC+r$
        else
            $r \leftarrow r+s$
            $AUC \leftarrow AUC+c$
        end if
    end for
    $R^- \leftarrow \frac{mr-AOC}{mn}$
    $R^+ \leftarrow \frac{AUC}{mn}$
    $\hat{\theta}_s \leftarrow R^+ - R^-$
```

Proof. (1)

$$R^+ = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{r_i-i} y_i \leq \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} y_i = M^+$$

$$R^- = \frac{1}{mn}\sum_{j=1}^{n}\sum_{i=1}^{s_j-j} x_j \leq \frac{1}{mn}\sum_{j=1}^{n}\sum_{i=1}^{m} x_j = M^-$$

(2)

$$\begin{aligned} M^+ - M^- &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(y_i - x_j) \leq \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(y_i - x_j)\psi_{ij} \\ &= \hat{\theta}_s \leq \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\psi_{ij} = \hat{\theta} \end{aligned}$$

The last step follows because $y_i \leq 1$ and $0 \leq x_j \leq 1$, hence $y_i - x_j \leq 1$, for any i and j.

Theorem 4. *(1) For separated scores (i.e., $y_i > x_j$ for any i and j), $M^+ - M^- = \hat{\theta}_s \leq \hat{\theta} = 1$.*
(2) For perfect scores (i.e., $y_i = 1$ and $x_j = 0$ for any i and j), $M^+ - M^- = \hat{\theta}_s = \hat{\theta} = 1$.

Proof. (1) For separated scores we have $\psi_{ij} = 1$ for any i and j, hence $M^+ - M^- = \hat{\theta}_s$ and $\hat{\theta} = 1$.
(2) For perfect scores we additionally have $y_i - x_j = 1$ for any i and j, hence $\hat{\theta}_s = 1$.

Finally, we investigate the statistical properties of sAUC. We note that $\hat{\theta}_s$ is an unbiased estimate of $\theta_s = \int_0^1 P(y > x + \tau)d\tau$, which is proved in the following theorem.

Theorem 5. *$\hat{\theta}_s$ is an unbiased estimate of θ_s.*

Proof. From Eq. (5), we have

$$E(\hat{\theta}_s) = \mathrm{E}\left[\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\int_0^1 \psi_{ij}(\tau)d\tau\right] = \int_0^1 \left(\mathrm{E}\left[\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\psi_{ij}(\tau)\right]\right)d\tau$$
$$= \int_0^1 P(y > x+\tau)d\tau$$

The variance of the estimate $\hat{\theta}_s$ can be obtained using the method of DeLong et al. [1] (we omit the proof due to lack of space).

Theorem 6. *The variance of* $\hat{\theta}_s$ *is estimated by*

$$\mathrm{var}(\hat{\theta}_s) = \frac{n-1}{mn(m-1)}\sum_{i=1}^{m}\left(\frac{1}{n}\sum_{j=1}^{n}(y_i - x_i)\psi_{ij} - \hat{\theta}_s\right)^2 + \frac{m-1}{mn(n-1)}\sum_{j=1}^{n}\left(\frac{1}{m}\sum_{i=1}^{m}(y_i - x_i)\psi_{ij} - \hat{\theta}_s\right)^2 .$$

4 Experimental Evaluation

Our experiments to evaluate the usefulness of sAUC for model selection are described in this section. Our main conclusion is that sAUC outperforms AUC and BS (Brier score) for selecting models, particularly when validation data is limited. We attribute this to sAUC having a lower variance than AUC and BS. Consequently, validation set values generalise better to test set values.

In the first experiment, we generated two artificial data sets (A and B) of 100 examples, each labelled with a 'true' probability p which is uniformly sampled from $[0,1]$. Then, we label the instances ($+$ if $p \geq 0.5$, $-$ otherwise). Finally, we swap the classes of 10 examples of data set A, and of 11 examples of data set B. We then construct 'models' M_A and M_B by giving them access to the 'true' probabilities p, and record which one is better (either M_A on data set A or M_B on data set B). For example, by thresholding p at 0.5, M_A has accuracy 90% on data set A, and M_B has accuracy 89% on data set B. We then add noise to obtain 'estimated' probabilities in the following way: $p' = p + k * U(-0.5, 0.5)$, where k is a noise parameter, and $U(-0.5, 0.5)$ obtains a pseudo-random number between -0.5 and 0.5 using a uniform distribution (if the corrupted values are > 1 or < 0, we set them to 1 and 0 respectively).

After adding noise, we again determine which model is better according to the four measures. In Figure 2, we show the proportion of cases where noise has provoked a change in the selection of the better model, using different values of the noise parameter k (averaged over 10,000 runs for each value of k). As expected, the percentage of changes increases with respect to noise for all four measures, but sAUC presents the most robust behaviour among all these four measures. This simple experiment shows that AUC, BS and accuracy are more vulnerable to the existence of noise in the predicted probabilities, and therefore, in this situation, the model selected by sAUC is more reliable than the models selected by the other three measures.

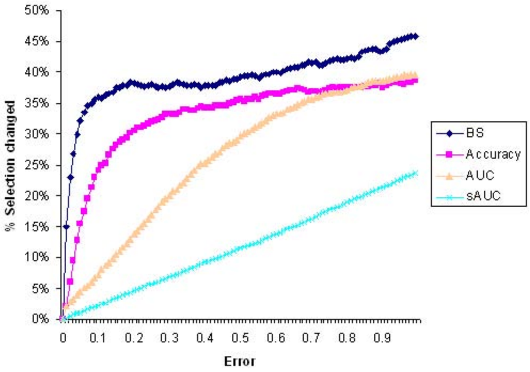

Fig. 2. The effect of noise in the probability estimates on four model selection measures

We continue reporting our experiments with real data. We use the three metrics (AUC, sAUC and BS) to select models on the validation set, and compare them using the AUC values on the test set. 17 two-class data sets are selected from the UCI repository for this purpose. Table 3 lists their numbers of attributes, numbers of instances, and relative size of the majority class.

Table 3. UCI data sets used in the experiments (larger data sets used in separate experiment in **bold face**)

# Data set	#Attrs	#Exs	%Maj.Class
1 Monk1	6	556	50.00
2 Monk2	6	601	65.72
3 Monk3	6	554	55.41
4 **Kr-vs-kp**	36	3,196	52.22
5 Tic-tac-toe	9	958	64.20
6 Credit-a	15	690	55.51
7 German	20	1,000	69.40
8 **Spam**	57	4,601	60.59
9 House-vote	16	435	54.25

# Data set	#Attrs	#Exs	%Maj.Class
10 Breast Cancer	9	286	70.28
11 Breast-w	9	699	65.52
12 Colic	22	368	63.04
13 Heart-statlog	13	270	59.50
14 **Sick**	29	3,772	93.87
15 **Caravan**	85	5,822	94.02
16 **Hypothyroid**	25	3,163	95.22
17 **Mushroom**	22	8,124	51.80

The configuration of the experiments is as follows. We distinguish between small data sets (with up to 1,000 examples) and larger data sets. For the 11 small data sets, we randomly split the whole data set into two equal-sized parts. One half is used as training set; the second half is again split into 20% validation set and 80% test set. In order to obtain models with sufficiently different performance, we train 10 different classifiers with the same learning technique (J48 unpruned with Laplace correction, Naive Bayes, and Logistic Regression, all from Weka) over the same training data, by randomly removing three attributes before training. We select the best model according to three measures: AUC, sAUC and BS using the validation set. The performance of each selected model is assessed by AUC on the test set. Results are averaged over 2000 repetitions of this

Table 4. Experimental results (AUC) on small data sets. Figures in **bold face** indicate a win of sAUC over AUC/BS. The last line indicates the total number of wins, which is never smaller than the critical value (9 out of 11).

	J48			Naive Bayes			Logistic Regression		
#	sAUC	AUC	BS	sAUC	AUC	BS	sAUC	AUC	BS
1	86.34	**83.76**	**85.81**	70.80	**67.98**	**69.96**	70.07	**67.28**	**69.23**
2	51.79	**51.32**	**51.05**	51.19	51.81	51.78	51.19	51.76	51.80
3	95.92	**93.20**	**95.47**	95.47	**92.21**	**94.96**	95.98	**92.65**	**95.58**
5	79.48	**77.72**	**78.16**	72.13	**70.88**	**71.05**	74.62	**72.11**	**72.68**
6	90.16	**89.25**	**89.56**	89.70	**89.06**	**89.61**	91.12	**90.62**	**90.55**
7	68.95	**68.75**	**68.85**	77.69	**77.24**	**77.25**	77.60	**77.29**	**77.20**
9	98.11	**97.81**	**97.98**	96.90	**96.74**	**96.81**	98.36	**98.24**	**98.28**
10	61.75	62.10	62.09	69.62	**69.09**	**68.98**	65.19	**64.94**	65.33
11	97.68	**97.64**	**97.67**	98.01	**97.94**	**98.00**	99.24	**99.18**	**99.22**
12	87.13	**85.65**	**86.13**	83.85	**83.60**	**83.82**	84.18	**83.74**	**83.76**
13	83.42	83.56	83.45	88.69	**88.68**	**88.49**	89.24	**89.12**	**89.13**
wins		9	9		10	10		10	9

experiment to reduce the effect of the random selection of attributes. These results are reported in Table 4. We performed a sign test over these results to compare the overall performance. The critical value for a two-tailed sign test over 11 data sets at $\alpha = 0.05$ is 9 wins. We conclude that sAUC significantly outperforms AUC/BS in all experiments. Given that the sign test is relatively weak, we consider this to be strong experimental evidence that sAUC is a good model selector for AUC in cases where we have limited validation data.

For the 6 larger data sets we employed a slightly different experimental configuration. In this case we employ 50% of the data for training the models, 25% for validation, and 25% for test. Here we only run 100 iterations. Our intuition is that when we have enough validation data, sAUC demonstrates less of an advantage for selecting models with higher test AUC because the variance of validation AUC is drastically reduced. The results included in Table 5 confirm this intuition, as the critical number of wins or losses (6 at $\alpha = 0.10$) is never achieved, and thus no significant differences in performance are observed.

Table 5. Experimental results (AUC) on larger data sets. Figures in **bold face** indicate a win of sAUC over AUC/BS. According to the sign test, the numbers of wins and losses are not significant.

	J48			Naive Bayes			Logistic Regression		
#	sAUC	AUC	BS	sAUC	AUC	BS	sAUC	AUC	BS
4	99.92	**99.91**	**99.91**	95.88	96.45	96.45	99.59	**99.55**	**99.57**
8	96.69	96.78	**96.67**	95.88	96.50	96.45	96.95	**96.93**	**96.91**
14	98.70	**98.67**	**98.65**	91.85	92.00	**91.62**	93.68	93.78	**93.59**
15	69.55	69.67	69.90	70.47	70.59	70.75	94.83	96.55	94.90
16	96.73	97.28	**96.59**	98.00	**97.99**	**97.90**	96.91	97.01	96.98
17	100	100	100	99.80	99.88	**99.79**	100	100	100
wins		2	3		1	3		2	3

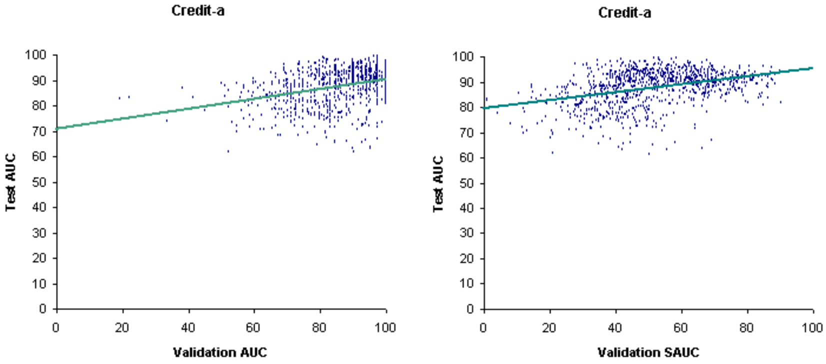

Fig. 3. Scatter plots of test AUC vs. validation AUC (left) and test AUC vs. validation sAUC (right) on the Credit-a data set.

Finally, Figure 3 shows two scatter plots of the models obtained for the Credit-a data set, the first one plotting test AUC against validation AUC, and the second one plotting test AUC against validation sAUC. Both plots include a straight line obtained by linear regression. Since validation sAUC is an underestimate of validation AUC (Theorem 3), it is not surprising that validation sAUC is also an underestimate of test AUC. Validation AUC appears to be an underestimate of test AUC on this data set, but this may be caused by the outliers on the left. But what really matters in these plots is the proportion of variance in test AUC not accounted for by the linear regression (which is $1-g^2$, where g is the linear correlation coefficient). We can see that this is larger for validation AUC, particularly because of the vertical lines observed in Figure 3 (left). These lines indicate how validation AUC fails to distinguish between models with different test AUC. This phenomenon particularly occurs for a number of models with perfect ranking on the validation set. Since sAUC takes the scores into account, and since these models do not have perfect scores on the validation set, the same phenomenon is not observed in Figure 3 (right).

5 Conclusions

The ROC curve is useful for visualising the performance of scoring classification models. ROC curves contain a wealth of information about the performance of one or more classifiers, which can be utilised to improve their performance and for model selection. For example, Provost and Fawcett [10] studied the application of model selection in ROC space when target misclassification costs and class distributions are uncertain.

In this paper we introduced the scored AUC (sAUC) metric to measure the performance of a model. The difference between AUC and scored AUC is that the AUC only uses the ranks obtained from scores, whereas the scored AUC uses both ranks and scores. We defined sAUC as the area under the sROC curve, which shows how quickly AUC deteriorates if the positive scores are decreased. Empirically, sAUC was found to select models with larger AUC values then AUC itself (which uses only ranks) or the Brier score (which uses only scores).

Evaluating learning algorithms can be regarded as a process of testing the diversity of two samples, that is, a sample of the scores for positive instances and that for negative instances. As the scored AUC takes advantage of both the ranks and the original values of samples, it is potentially a good statistic for testing the diversity of two samples, in a similar vein as the Wilcoxon-Man-Whitney U statistic. Preliminary experiments suggest that sAUC has indeed higher power than WMW. Furthermore, while this paper only investigates sAUC from the non-parametric perspective, it is worthwhile to study its parametric properties. We plan to investigate these further in future work.

Acknowledgments

We thank José Hernández-Orallo and Thomas Gärtner for useful discussions. We would also like to thank the anonymous reviewers for their helpful comments.

References

1. DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44, 837–845 (1988)
2. Ferri, C., Flach, P., Hernández-Orallo, J., Senad, A.: Modifying ROC curves to incorporate predicted probabilities. In: Proceedings of the Second Workshop on ROC Analysis in Machine Learning (ROCML'05) (2005)
3. Fawcett, T.: Using Rule Sets to Maximize ROC Performance. In: Proc. IEEE Int'l Conf. Data Mining, pp. 131–138 (2001)
4. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Let. 27-8, 861–874 (2006)
5. Hanley, J.A., McNeil, B.J.: The Meaning and Use of the AUC Under a Receiver Operating Characteristic (ROC) Curve. Radiology 143, 29–36 (1982)
6. Hsieh, F., Turnbull, B.W.: Nonparametric and Semiparametric Estimation of the Receiver Operating Characteristic Curve. Annals of Statistics 24, 25–40 (1996)
7. Huang, J., Ling, C.X.: Dynamic Ensemble Re-Construction for Better Ranking. In: Proc. 9th Eur. Conf. Principles and Practice of Knowledge Discovery in Databases, pp. 511–518 (2005)
8. Huang, J., Ling, C.X.: Using AUC and Accuray in Evaluating Learing Algorithms. IEEE Transactions on Knowledge and Data Engineering 17, 299–310 (2005)
9. Provost, F., Fawcett, T., Kohavi, R.: Analysis and Visualization of Classifier Performance: Comparison Under Imprecise Class and Cost Distribution. In: Proc. 3rd Int'l Conf. Knowledge Discovery and Data Mining, pp. 43–48 (1997)
10. Provost, F., Fawcett, T.: Robust Classification for Imprecise Environments. Machine Learning 42, 203–231 (2001)
11. Provost, F., Domingos, P.: Tree Induction for Probability-Based Ranking. Machine Learning 52, 199–215 (2003)
12. Wu, S.M., Flach, P.: Scored Metric for Classifier Evaluation and Selection. In: Proceedings of the Second Workshop on ROC Analysis in Machine Learning (ROCML'05) (2005)
13. Zhou, X.H., Obuchowski, N.A., McClish, D.K.: Statistical Methods in Diagnostic Medicine. John Wiley and Sons, Chichester (2002)