A scored AUC Metric for Classifier Evaluation and Selection

Shaomin Wu SHAOMIN.WU@READING.AC.UK
School of Construction Management and Engineering, The University of Reading, Reading RG6 6AW, UK

Peter Flach PETER.FLACH@BRISTOL.AC.UK
Department of Computer Science, University of Bristol, Bristol BS8 1UB, UK

Abstract

The area under the ROC (Receiver Operating Characteristic) curve, or simply AUC, has been widely used to measure model performance for binary classification tasks. It can be estimated under parametric, semiparametric and nonparametric assumptions. The non-parametric estimate of the AUC, which is calculated from the ranks of predicted scores of instances, does not always sufficiently take advantage of the predicted scores. This problem is tackled in this paper. On the basis of the ranks and the original values of the predicted scores, we introduce a new metric, called a *scored AUC* or sAUC. Experimental results on 20 UCI data sets empirically demonstrate the validity of the new metric for classifier evaluation and selection.

1. Introduction

In the data mining and machine learning literature, there are many learning that algorithms 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 performance metrics for evaluating the models.

The predicted outcome of a classification model can be either a class decision such as positive and negative on each instance, or a predicted score that gives how much an instance is predicted to be positive or negative. Most models can produce predicted scores; and those that only produce class decisions can easily be converted to models that produce predicted scores [6, 13]. In this paper we assume that the scores represent class-conditional likelihoods.

Appearing in *Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning*, Bonn, Germany, 2005.

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 proportion of positive to negative instances changes in a test set, they may no longer perform optimally, or even acceptably. The ROC (Receiver Operating Characteristics) curve, however, is insensitive to the change in the class distribution. If the class distribution changes in a test set, the ROC curve will not change. The ROC curve has been used as a tool for model selection in the medical area since the late 1970s, and more recently introduced to evaluate machine learning algorithms [11, 12]. It is defined as a plot of a model's true positive rate as the y coordinate versus its false positive rate as the x coordinate, under all possible score thresholds.

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 [14], semiparametric [9] and nonparametric [7] 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) statistic test of ranks [7]. Huang and Ling [10] also show show theoretically and empirically that AUC is a better measure for model evaluation than accuracy.

The nonparametric estimate of the AUC is calculated on the basis of the ranks of the predicted 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 predicted scores are only used to rank instances, and otherwise ignored. The AUC, estimated simply from the ranks of the predicted scores, can remain unchanged even when the predicted scores change. This can lead to a loss of useful information, and may therefore produce sub-optimal results.

This paper attempts to combine both the ranks and the original values of the predicted scores to evaluate the performance of binary classification models. A scored AUC metric is introduced for estimating the performance of models based on their original predicted scores. The paper has similar aims to [3], the approaches are however different and have been developed independently.

The paper is structured as follows. Section 2 introduces a new algorithm to calculate the AUC, and a new AUC-like metric derived from the AUC. Section 3 investigates the properties of the new metric, which we call sAUC (scored AUC). In Section 4 we present experimental results on 20 data sets from the UCI repository [1], using sAUC for model selection. Section 5 presents the main conclusions and suggests further work.

2. A Scored AUC

The purpose of this section is to introduce a new algorithm for calculating the AUC, and then propose a new metric, called scored AUC (abbreviated to sAUC).

2.1. Calculating AUC

Denote the total number of positive instances and negative instances by N_+ and N_-, respectively. Let $\{x_1, \ldots, x_{N_+}\}$ (where $x_1 \geq \ldots \geq x_{N_+}$) be the scores predicted by a model for the N_+ positives, and $\{y_1, \ldots, y_{N_-}\}$ (where $y_1 \leq \ldots \leq y_{N_-}$) be the scores predicted by a model for the N_- negatives. Assume both x_i and y_j have been normalized and they are within interval (0,1), where $i = 1, 2, \ldots, N_+$ and $j = 1, 2, \ldots, N_-$. Since AUC is equivalent to the WMW statistic, estimating the probability that a randomly selected positive instance will score higher than a randomly selected negative instance, it can be expressed as follows:

$$\mathrm{AUC} = \frac{1}{N_+ N_-} \sum_{i=1}^{N_+} \sum_{j=1}^{N_-} I(x_i > y_j) \tag{1}$$

where $I(\bullet)$ is an indicator function satisfying I(true)=1 and I(false)=0. Let Z_a be the sequence produced by merging the $\{x_1, \ldots, x_{N_+}\}$ and $\{y_1, \ldots, y_{N_-}\}$, and sorting the merged set in ascending order (so a good ranker would put the positives after the negatives in Z_a). Then the expression of the AUC can be simplified as follows [8].

$$\mathrm{AUC} = \frac{1}{N_+ N_-} \left(\sum_{i=1}^{N_+} r_i - \frac{N_+(N_+ + 1)}{2} \right) \tag{2}$$

where r_i is the rank of x_i in Z_a. Further simplifying the AUC in Eq. (2), we obtain

$$\mathrm{AUC} = R_+ = \frac{1}{N_+ N_-} \sum_{i=1}^{N_+} (r_i - i) \tag{3}$$

$r_i - i$ in Eq. (3) is the number of negatives before the ith positive in Z_a, and the AUC is the sum of the number of negatives before each of the N_+ positives in Z_a, divided by the number of pairs of one positive and one negative.

Let Z_d be the sequence produced by merging the $\{x_1, \ldots, x_{N_+}\}$ and $\{y_1, \ldots, y_{N_-}\}$, and sorting in descending order (so a good ranker would put the positives before the negatives in Z_d). Analogously, the AUC can be expressed as

$$\mathrm{AUC} = R_- = \frac{1}{N_+ N_-} \sum_{j=1}^{N_-} (s_j - j) \tag{4}$$

where s_j is the rank of y_j, and $s_j - j$ in Eq. (4) is the number of positives before the jth negative in Z_d. Then, the AUC represents the normalised sum of the number of positives before each of the N_- negatives in Z_d.

Based on Eq. (3) and Eq. (4), we derived the algorithm shown in Table 1 to calculate the value of the AUC. The algorithm is different from other algorithms to calculate AUC because it doesn't calculate ranks.

Table 1. Algorithm for calculating the AUC.

Inputs.
  $Z = \{C_i, f(i)\}$
  C_i: instance i
  $f(i)$: score for predicting instance C_i to be positive
  N_-: number of negative instances in the training set
  N_+: number of positive instances in the training set
Outputs.
  R_+: AUC value of the model
  R_-: AUC value of the model
Begin
  $Z_d \leftarrow Z$ sorted decreasingly by f scores
  initialize: $R_- \leftarrow 0$, $R_+ \leftarrow 0$, $n_- \leftarrow 0$, $n_+ \leftarrow 0$
  for $C_i \in Z_d$ **do**
    if C_i is a positive instance **then**
      $R_- \leftarrow R_- + N_- - n_-$
      $n_+ \leftarrow n_+ + 1$
    else
      $n_- \leftarrow n_- + 1$
      $R_+ \leftarrow R_+ + n_+$
    end if
  end for
  $R_- \leftarrow \frac{R_-}{N_+ N_-}$
  $R_+ \leftarrow \frac{R_+}{N_+ N_-}$
End

2.2. Scored AUC

The AUC may fail to detect the performance difference between models because it is calculated simply from the ranks of predicted scores. The following example demonstrates this weakness of the AUC.

Example 1 *Given a data set containing 3 positive and 4 negative instances, and two models, M_1 and M_2, built on the data set. M_1 and M_2 have the following predicted scores respectively (the underlined scores indicate where the two models differ).*

$$M_1: +0.95 \quad -0.89 \quad \underline{+0.86} \quad \underline{+0.84} \quad -0.15 \quad -0.13 \quad -0.10$$
$$M_2: +0.95 \quad -0.89 \quad \underline{+0.20} \quad \underline{+0.16} \quad -0.15 \quad -0.13 \quad -0.10$$

Here, for example, +0.95 means that a positive instance is predicted positive with a score of 0.95, and -0.89 means that a negative instance is predicted positive with a score of 0.89. The two models result in the same ranking and therefore have the same AUC. However, model M_1 might be considered better than M_2 because of the following two reasons.

- *If we select a threshold (i.e., operating point) from the interval (0.21, 0.83) and estimate the performance of the two models using metrics precision and accuracy, M_1 outperforms M_2, and*
- *when the scores predicted for the two positives change within interval (0.15, 0.89), the AUC value remains unchanged. This may lead to a biased estimate for model performance.*

An alternative understanding of the AUC is as follows. In Eq. (1), the component $I(x_i > y_j)$ is an indicator that only reflects the ordinal comparison between the predicted scores, but it does not reflect how much x_i is larger than y_j.

Obviously, $(x_i - y_j)I(x_i > y_j)$ measures not only whether $x_i > y_j$ but also how much x_i is larger than y_j. If we replace $I(x_i > y_j)$ with $(x_i - y_j)I(x_i > y_j)$ in Eq. (1), and denote the new metric by sAUC (scored AUC), we have the following definition.

Definition 1 *The scored AUC is defined as*

$$\text{sAUC} = \frac{1}{N_+N_-}\sum_{i=1}^{N_+}\sum_{j=1}^{N_-}(x_i - y_j)I(x_i > y_j) \qquad (5)$$

The scored AUC uses both the ranks $I(x_i > y_j)$ and the original predicted scores $(x_i - y_j)$. It estimates the score difference between randomly chosen positives and negatives with positive score differences, multiplied with the probability that a positive score difference occurs. The latter term is estimated by AUC, and thus sAUC may be interpreted as a multiplicative correction on AUC.

In order to simplify sAUC in Eq. (5), we replace $r_i - i$, which is the number of negatives before the ith positive instance, with the sum of predicted scores for the negatives before the ith positive instance in Z_a in Eq. (3). Similar replacement can be made in Eq. (4). Then, we can obtain:

$$\text{sAUC} = \frac{1}{N_+N_-}\left(\sum_{j=1}^{N_-}\sum_{t=1}^{s_j-j}x_t - \sum_{i=1}^{N_+}\sum_{t=1}^{r_i-i}y_t\right)$$

where $\sum_{t=1}^{s_j-j}x_t$ is the sum of the scores predicted for positive instances before the jth negative instance in Z_a, and $\sum_{t=1}^{r_i-i}y_t$ is the sum of the scores predicted for negative instances before the ith positive instance in Z_d. Denote

$$\text{R}_{\text{s}+} = \frac{1}{N_+N_-}\sum_{j=1}^{N_-}\sum_{t=1}^{s_j-j}x_t$$

and

$$\text{R}_{\text{s}-} = \frac{1}{N_+N_-}\sum_{i=1}^{N_+}\sum_{t=1}^{r_i-i}y_t$$

Then, $\text{sAUC} = \text{R}_{\text{s}+} - \text{R}_{\text{s}-}$. The algorithm for calculating the values of sAUC, $\text{R}_{\text{s}+}$ and $\text{R}_{\text{s}-}$ is shown in Table 2.

Table 2. Algorithm for calculating sAUC.

Inputs.
$Z = \{C_i, f(i)\}$
C_i: instance i
$f(i)$: score for predicting instance C_i to be positive
N_-: number of negative instances in the training set
N_+: number of positive instances in the training set
V_{p-}: sum of $f(i)$ of negative instances
Outputs.
sAUC: scored AUC
Begin
$Z_d \leftarrow Z$ sorted decreasingly by f scores
initialize: $R_{s-} \leftarrow 0$, $R_{s+} \leftarrow 0$, $n_{s-} \leftarrow 0$, $n_{s+} \leftarrow 0$
for $C_i \in Z_d$ **do**
  if C_i is a positive instance **then**
    $R_{s-} \leftarrow R_{s-} + V_{p-} - n_{s-}$
    $n_{s+} \leftarrow n_{s+} + f(i)$
  else
    $n_{s-} \leftarrow n_{s-} + f(i)$
    $R_{s+} \leftarrow R_{s+} + n_{s+}$
  end if
end for
$R_{s-} \leftarrow \frac{R_{s-}}{N_+N_-}$
$R_{s+} \leftarrow \frac{R_{s+}}{N_+N_-}$
sAUC $\leftarrow R_{s+} - R_{s-}$
End

In building a classification model, one hopes that the scores predicted for positive instances are larger whereas those for negative instances are smaller. Hence, a good model should have a large $\text{R}_{\text{s}+}$ and a small $\text{R}_{\text{s}-}$.

Example 2 *Continuing Example 1, we have*
$\text{R}_{\text{s}+} = 0.7417$, $\text{R}_{\text{s}-} = 0.1692$ *and* sAUC $= 0.5725$, *for model M_1, and*
$\text{R}_{\text{s}+} = 0.4067$, $\text{R}_{\text{s}-} = 0.1692$ *and* sAUC $= 0.2375$, *for model M_2.*

3. Properties of sAUC

The properties of sAUC are investigated in this section.

Lemma 1 *If $x_i = 1$ and $y_j = 0$ for any i and j, then $R_{s-} = 0$, $R_{s+} = 1$ and* sAUC = AUC = 1.

Denote $M^+ = \frac{1}{N_+}\sum_{i=1}^{N_+} x_i$ and $M^- = \frac{1}{N_-}\sum_{i=1}^{N_-} y_i$. The quantity $M^+ - M^-$ estimates the score difference between randomly selected positives and negatives (i.e., not just those that have positive score differences), and is investigated in [3].

Lemma 2 *If $x_i > y_j$ for any i and j, then* $\text{sAUC} = M^+ - M^-$.

Theorem 1 *$R_{s+} \leq M^+$ and $R_{s-} \leq M^-$.*

Proof.

$$R_{s+} = \frac{1}{N_+N_-}\sum_{j=1}^{N_-}\sum_{t=1}^{s_j-j} x_t \leq \frac{1}{N_+N_-}\sum_{j=1}^{N_-}\sum_{t=1}^{N_+} x_t = \text{M}^+$$

and

$$R_{s-} = \frac{1}{N_+N_-}\sum_{i=1}^{N_+}\sum_{t=1}^{r_i-i} y_t \leq \frac{1}{N_+N_-}\sum_{i=1}^{N_+}\sum_{t=1}^{N_-} y_t = \text{M}^-$$

Theorem 2 $M^+ - M^- \leq \text{sAUC} \leq \text{AUC}$.

Proof.

$$\begin{aligned} M^+ - M^- &= \frac{1}{N_+N_-}\sum_{i=1}^{N_+}\sum_{j=1}^{N_-}(x_i - y_j) \\ &\leq \frac{1}{N_+N_-}\sum_{i=1}^{N_+}\sum_{j=1}^{N_-}(x_i - y_j)I(x_i > y_j) \\ &= \text{sAUC} \end{aligned}$$

Because $x_i \leq 1$ and $0 \leq y_i \leq 1$, it follows that $x_i - y_j \leq 1$. We then have

$$\begin{aligned} \text{sAUC} &= \frac{1}{N_+N_-}\sum_{i=1}^{N_+}\sum_{j=1}^{N_-}(x_i - y_j)I(x_i > y_j) \\ &\leq \frac{1}{N_+N_-}\sum_{i=1}^{N_+}\sum_{j=1}^{N_-} I(x_i > y_j) \\ &= \text{AUC} \end{aligned}$$

4. sAUC for model selection

Our initial experiments to evaluate the approach introduced in this paper are described in this section. We use both the sAUC and the AUC metrics to select models on the training set, and compare them using the AUC values on the test set. 20 data sets are selected from the UCI repository [1] 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 our experiments.

#	Data set	#Attrs	Size	Majority Class
1	Australia	14	690	55.51
2	Sonar	60	208	51.92
3	Glass	9	214	67.29
4	German	20	1000	69.40
5	Monk1	6	556	50.00
6	Monk2	6	601	65.72
7	Monk3	6	554	55.41
8	Hepatitis	19	155	78.71
9	House	16	435	62.07
10	Tic-tac-toe	9	958	64.20
11	Heart	13	227	55.56
12	Ionosphere	34	351	64.10
13	Breast Cancer	9	286	70.28
14	Lymphography	17	148	56.81
15	Primary Tumor	17	339	55.75
16	Solar-Flare	12	323	56.35
17	Hayes-Roth	4	133	60.91
18	Credit	15	690	55.51
19	Balance	4	625	53.92
20	Bridges	12	108	66.67

In the experiments, a data set was divided into ten folds: eight for training, one for validation, and one for testing. We first trained five models, (naive Bayes, logistic, decision tree, kstar, and voting feature interval [2]) on the training set, selected the model with maximum values of sAUC or AUC on the validation set, and finally tested the selected models on the test set. We ran the experiment ten times yielding 100 pairs of AUC values, and performed a paired t-test with level of confidence 0.05 to test the significance of the average difference in AUC. The AUC values of the models selected by the two metrics are shown in Table 4. As indicated in the table, we obtained 6 significant wins for sAUC and 2 significant losses.

We noticed that the 6 data sets where the sAUC metric won are relatively small, and hypothesised that sAUC may be particularly suitable for smaller data sets. We tested this by randomly selecting 150 instances from each data set. Applying the same approach to model selection, we obtained the results shown in Table 5. Among the 20 data sets, we now obtain 9 significant wins and no losses. Although further investigations are necessary, we believe these are promising results.

Table 4. Results of experiment with 20 UCI data sets (AUC values). The last column (labelled S?) indicates whether this is a statistically significant win or loss for sAUC, using a paired t-test with 0.05 level of confidence.

#	Data set	use AUC	use sAUC	S?
1	Australia	90.15±0.53	90.25±0.60	
2	Sonar	93.67±1.03	94.48±0.93	
3	Glass	95.23±0.90	97.16±0.61	∨
4	German	92.34±0.86	92.34±0.86	
5	Monk1	99.98±0.017	99.89±0.071	
6	Monk2	97.05±0.32	94.06±0.58	×
7	Monk3	98.63±0.28	98.84±0.23	
8	Hepatitis	90.74±1.15	91.13±1.01	
9	House	99.66±0.089	99.55±0.19	
10	Tic-tac-toe	99.68±0.034	99.71±0.024	
11	Heart	92.60±0.68	92.47±0.75	
12	Ionosphere	95.47±0.41	92.35±0.53	×
13	Breast Cancer	85.88±1.33	87.67±1.09	∨
14	Lymphography	88.22±1.04	88.89±1.00	∨
15	Primary Tumor	87.28±0.85	87.84±1.04	
16	Solar-Flare	89.62±0.72	89.50±0.67	
17	Hayes-Roth	93.00±1.03	94.25±0.85	
18	Credit	90.14±0.50	91.05±0.45	∨
19	Balance	99.88±0.043	99.98±0.0083	∨
20	Bridges	86.16±1.51	88.13±1.3	∨
	average	93.05	93.45	

Table 5. Results of experiment with 150 randomly selected instances (AUC values). The data sets marked with * have less than 150 instances, so the results are the same as in Table 4.

#	Data set	use AUC	use sAUC	S?
1	Australia	84.06 ±1.63	86.02±1.55	∨
2	Sonar	91.02±0.76	91.03±0.87	
3	Glass	93.37±1.47	96.13±0.76	∨
4	German	70.39±2.14	68.80±1.88	
5	Monk1	93.66±1.65	95.62±1.27	∨
6	Monk2	79.82±2.11	80.98±1.65	
7	Monk3	96.31±0.77	97.74±0.60	∨
8	Hepatitis	90.74±1.15	91.13±1.01	
9	House	98.84±0.33	99.75±0.11	∨
10	Tic-tac-toe	95.60±1.05	95.96±1.07	
11	Heart	93.86±1.15	95.65±0.87	∨
12	Ionosphere	93.07±1.25	92.75±1.38	
13	BreastCancer	78.39±2.32	78.86±2.52	
14	Lymphography*	88.22±1.04	88.89±1.00	∨
15	PrimaryTumor	85.33±1.22	86.62±1.22	
16	Solar-Flare	89.25±1.15	88.54±1.07	
17	Hayes-Roth*	93.00±1.03	94.25±0.85	
18	Credit	83.51±1.39	83.61±1.39	
19	Balance	97.83±0.58	99.41±0.15	∨
20	Bridges*	86.16±1.51	88.13±1.3	∨
	average	89.36	89.97	

5. Conclusions

The ROC curve is useful for visualising the performance of scoring classification models. Both the ROC curve and the AUC have drawn considerable attentions. 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 [12] studied the application of model selection in ROC space when target misclassification costs and class distributions are uncertain; the AUC values have been used by Ferri, Flach and Hernández-Orallo to find optimal labellings of decision trees [4]; and Flach and Wu [5] introduce an approach to model improvement using ROC analysis.

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 predicted scores, whereas the scored AUC uses both the ranks and the original values of the predicted scores. sAUC was evaluated on 20 UCI data sets, and found to select models with larger AUC values then AUC itself, particularly for smaller data sets.

The scored AUC metric is derived from the Wilcoxon-Mann-Whitney (WMW) statistic which is equivalent to AUC. The WMW statistic is widely used to test if two samples of data come from the same distribution, when no distribution assumption is given. Evaluating learning algorithms can be regarded as a process of testing the diversity of two samples, that is, a sample of the predicted scores for postive instances and that for negative instances. As the scored AUC takes advantage of both the ranks and the orignal values of samples, it is potentially a good statistic for testing the diversity of two samples. Our future work will focus on studying the scored AUC from this statistical point of view.

Acknowledgments

We thank José Hernández-Orallo and Cèsar Ferri from Universitat Politècnica de València for enlightening discussions on incorporating scores into ROC analysis. We also gratefully acknowledge the insightful comments made by the two anonymous reviewers. In particular, we adopted the improved experimental procedure suggested by one reviewer.

References

[1] 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, 1998.

[2] G. Demiroz and A. Guvenir, "Classification by Voting Feature Intervals," *Proc. 7th European Conf. Machine Learning*, pp. 85-92, 1997.

[3] C. Ferri, P. Flach, J. Hernández-Orallo, and A. Senad, "Modifying ROC curves to incorporate predicted probabilities," appearing in ROCML'05, 2005.

[4] C. Ferri, P. Flach, and J. Hernández-Orallo, "Decision Tree Learning Using the Area Under the ROC Curve," *Proc. 19th Int'l Conf. Machine Learning*, pp. 139-146, 2002.

[5] P. Flach and S. Wu, "Repairing Concavities in ROC Curves," *Proc. 19th Int'l Joint Conf. Artificial Intelligence*, 2005.

[6] T. Fawcett, "Using Rule Sets to Maximize ROC Performance," *Proc. IEEE Int'l Conf. Data Mining*, pp. 131-138, 2001.

[7] J.A. Hanley and B.J. McNeil, "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve," *Radiology*, vol. 143, pp. 29-36, 1982.

[8] D.J. Hand and R.J. Till, "A simple generalisation of the Area Under the ROC Curve for Multiple-Class Classification Problems," *Machine Learning*, vol. 45, no. 2, pp. 171-186, 2001.

[9] F. Hsieh and B.W. Turnbull, "Nonparametric and Semiparametric Estimation of the Receiver Operating Characteristic Curve," *Annals of Statistics*, vol. 24, pp. 25-40, 1996.

[10] J. Huang and C.X. Ling "Using AUC and Accuray in Evaluating Learing Algorithms", *IEEE Transactions on Knowledge and Data Engineering* vol. 17, no. 3, pp. 299-310, 2005.

[11] F. Provost, T. Fawcett and R. Kohavi, "Analysis and Visualization of Classifier Performance: Comparison Under Imprecise Class and Cost Distribution," *Proc. 3rd Int'l Conf. Knowledge Discovery and Data Mining*, pp. 43-48, 1997.

[12] F. Provost and T. Fawcett, "Robust Classification for Imprecise Environments," *Machine Learning*, vol. 42, pp. 203-231, 2001.

[13] F. Provost and P. Domingos, "Tree Induction for Probability-Based Ranking," *Machine Learning,* vol. 52, pp. 199-215, 2003.

[14] X.H. Zhou, N.A. Obuchowski and D.K. McClish, *Statistical Methods in Diagnostic Medicine,* John Wiley and Sons, 2002.