AUC: a Better Measure than Accuracy in Comparing Learning Algorithms

Authors:
Charles X. Ling, Department of Computer Science, University of Western Ontario, Canada
&
Jin Huang, Department of Computer Science, University of Western Ontario, Canada
&
Harry Zhang, Faculty of Computer Science, University of New Brunswick, Canada

Presented by:
William Elazmeh, Ottawa-Carleton Institute for Computer Science, Canada

Introduction

- The focus is visualization of classifier's performance
- Traditionally, performance = predictive accuracy
- Accuracy ignores probability estimations of classification in favor of class labels
- ROC curves show the trade off between false positive and true positive rates
- AUC of ROC is a better measure than accuracy
- AUC as a criteria for comparing learning algorithms
- AUC replaces accuracy when comparing classifiers
- Experimental results show AUC indicates a difference in performance between decision trees and Naive Bayes (significantly better)

Matrices

Confusion Matrix

	+	-
Y	T+	F+
N	F-	T-

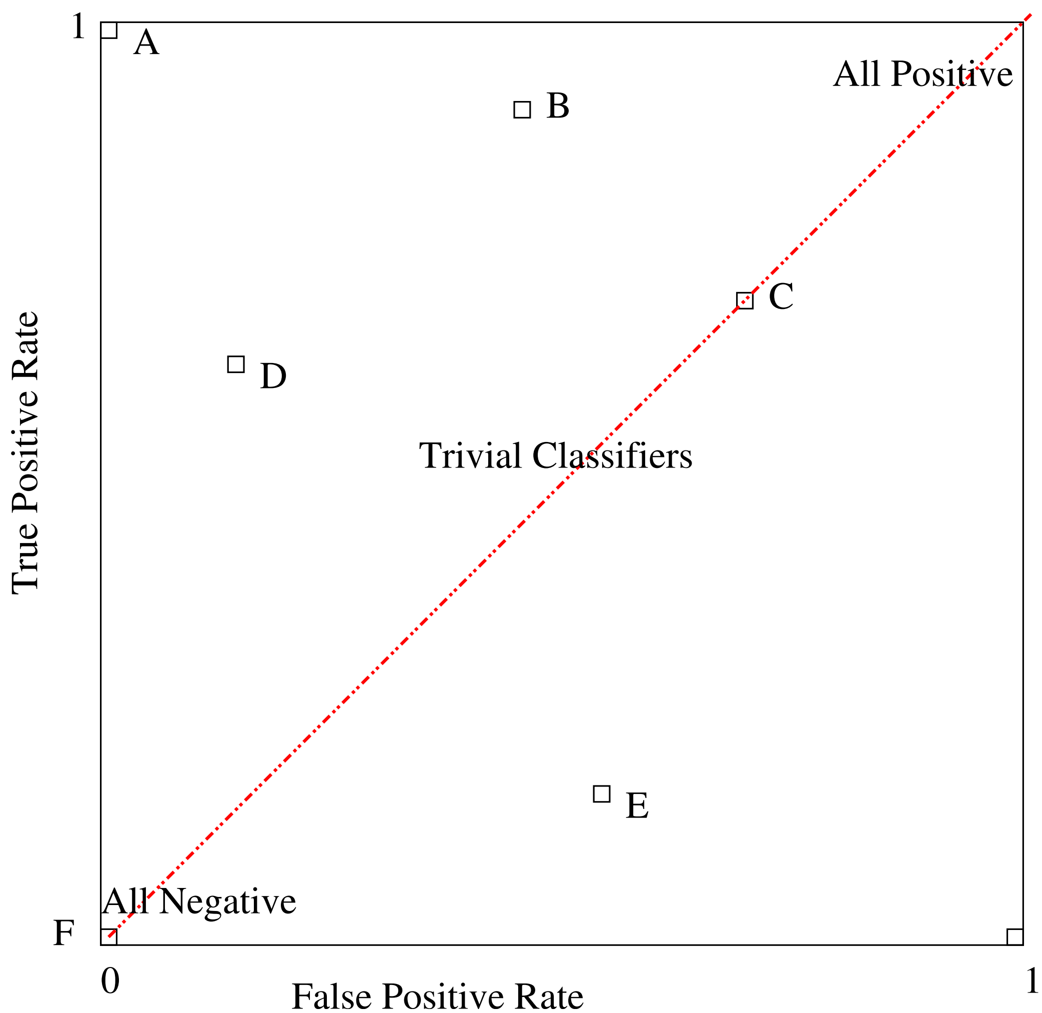
$$\text{F+ Rate} = \frac{F+}{-} \qquad \text{T+ Rate (Recall)} = \frac{T+}{+}$$

$$\text{Precision} = \frac{T+}{Y} \qquad \text{Accuracy} = \frac{(T+)+(T-)}{(+)+(-)}$$

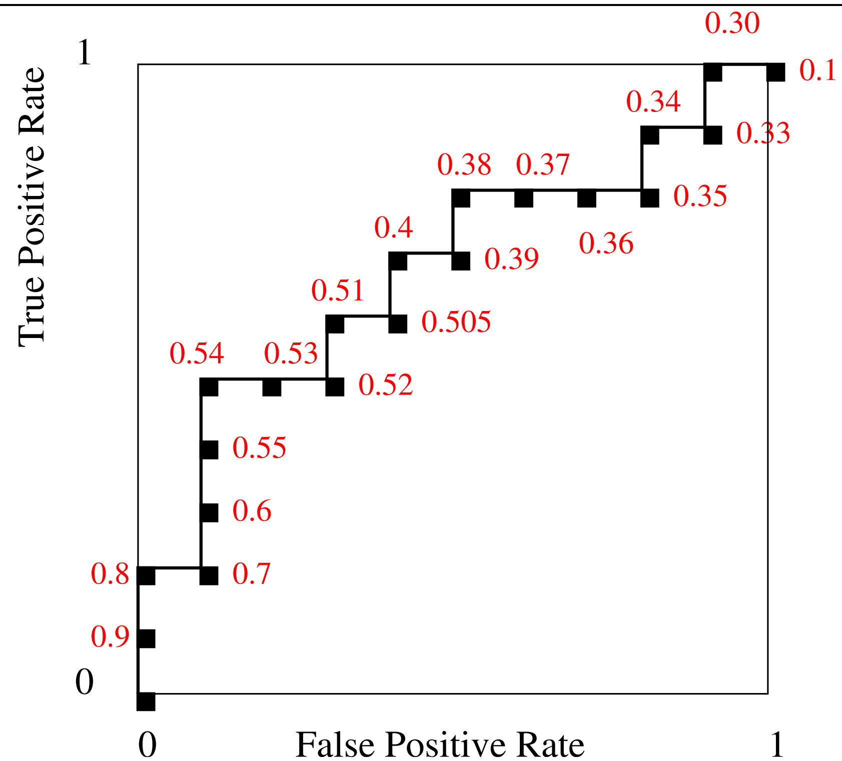
$$\text{F-Score} = \qquad \text{Precision} \times \text{Recall}$$

$$\text{Error Rate} = \qquad \text{1 - Accuracy}$$

ROC Space

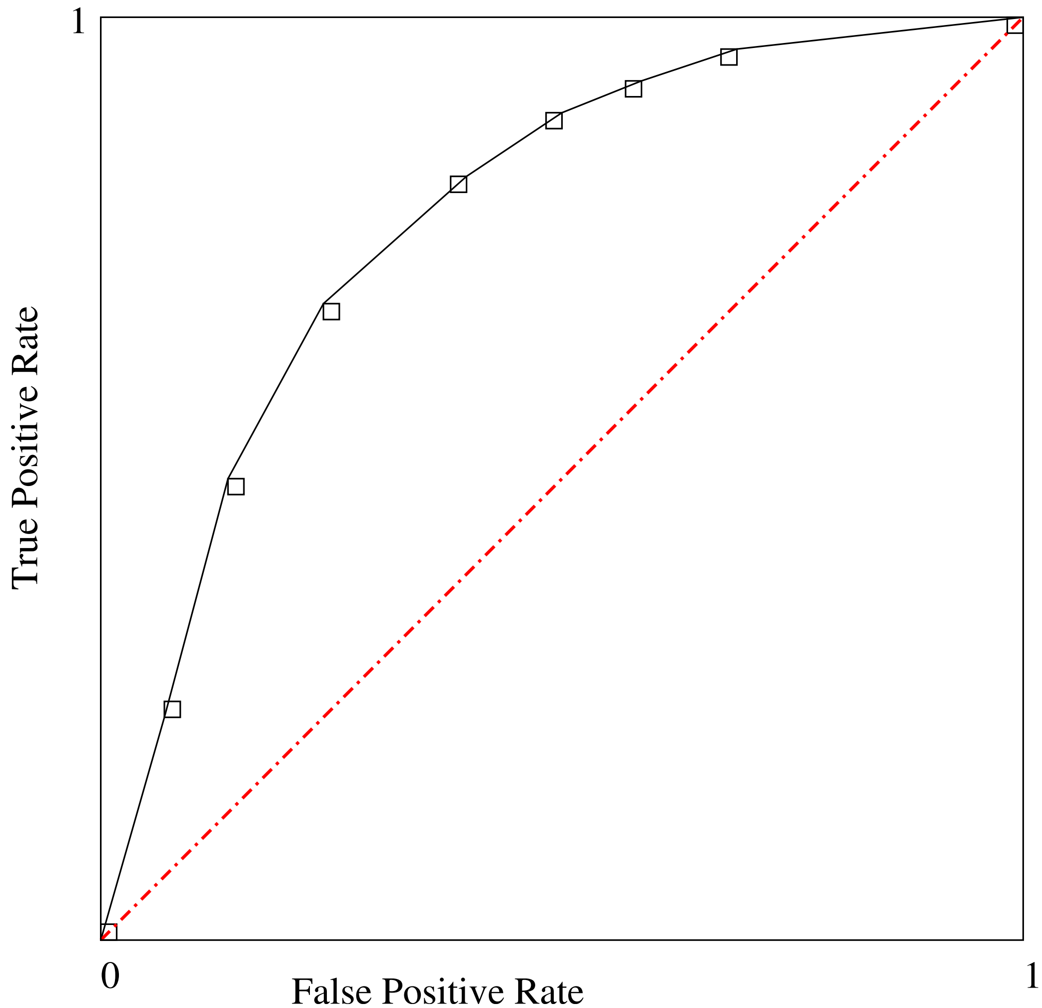

1
A
B
All Positive
C
D
Trivial Classifiers
E
All Negative
F
0
1
True Positive Rate
False Positive Rate

ROC Curves

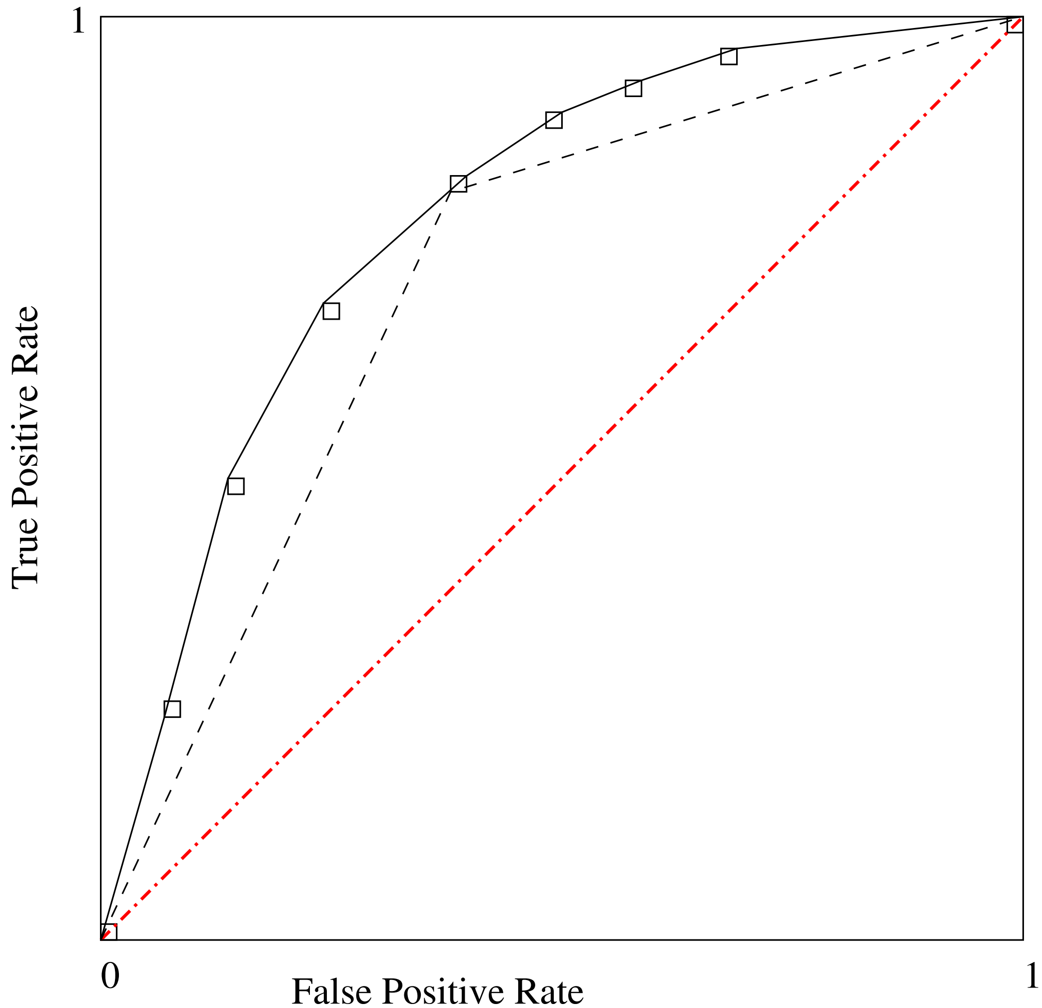


#	Class	Score	#	Class	Score
1	+	0.9	11	+	0.4
2	+	0.8	12	-	0.39
3	-	0.7	13	+	0.38
4	+	0.6	14	-	0.37
5	+	0.55	15	-	0.36
6	+	0.54	16	-	0.35
7	-	0.53	17	+	0.34
8	-	0.52	18	-	0.33
9	+	0.51	19	+	0.30
10	-	0.505	20	-	0.1

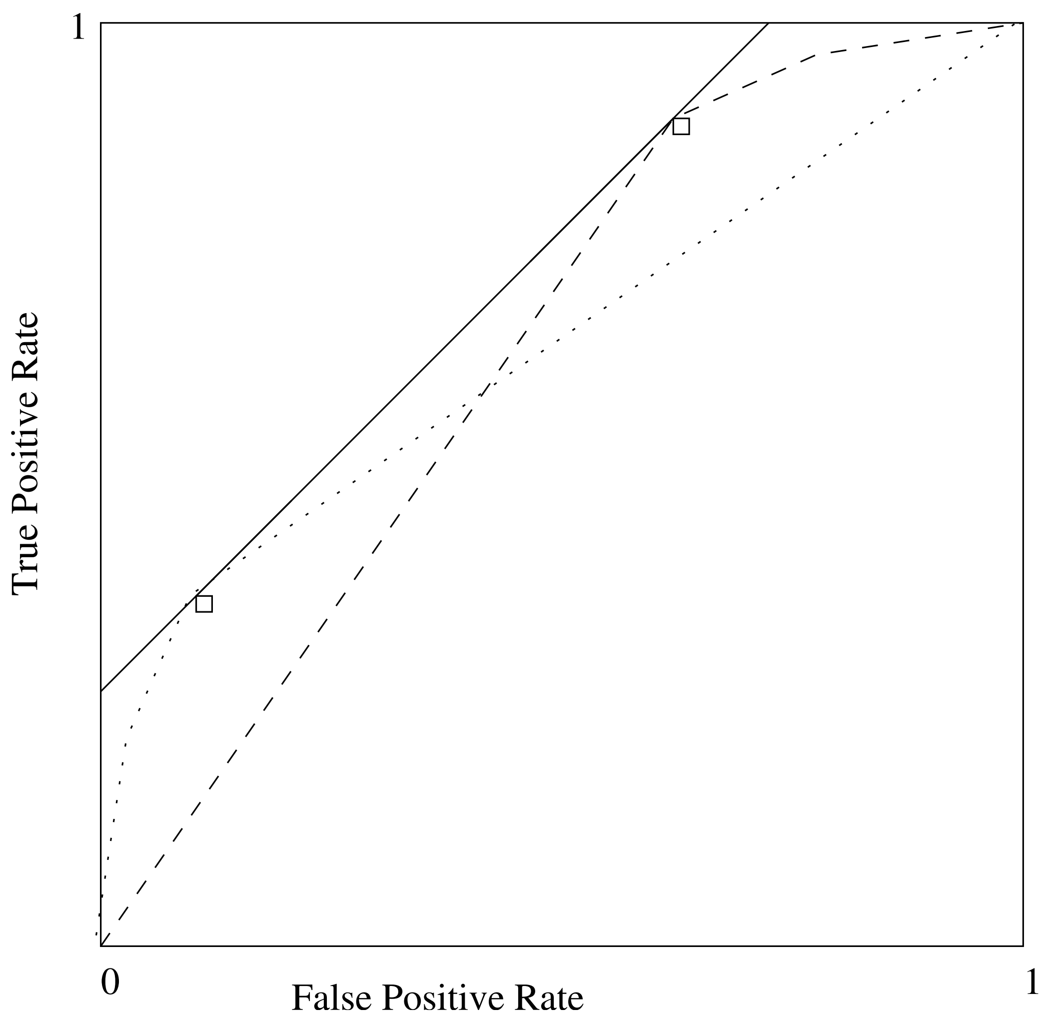
ROC Curves



Comparing Classifier Performance
ROC



Choosing Between Classifiers ROC



Area Under the Curve AUC

$$AUC = \frac{\Sigma Rank(+) - |+| \times (|+|+1)/2}{|+|+|-|}$$

where:
$\sum Rank(+)$ is the sum the ranks of all positively classified examples
$|+|$ is the number of positive examples in the dataset
$|-|$ is the number of negative examples in the dataset

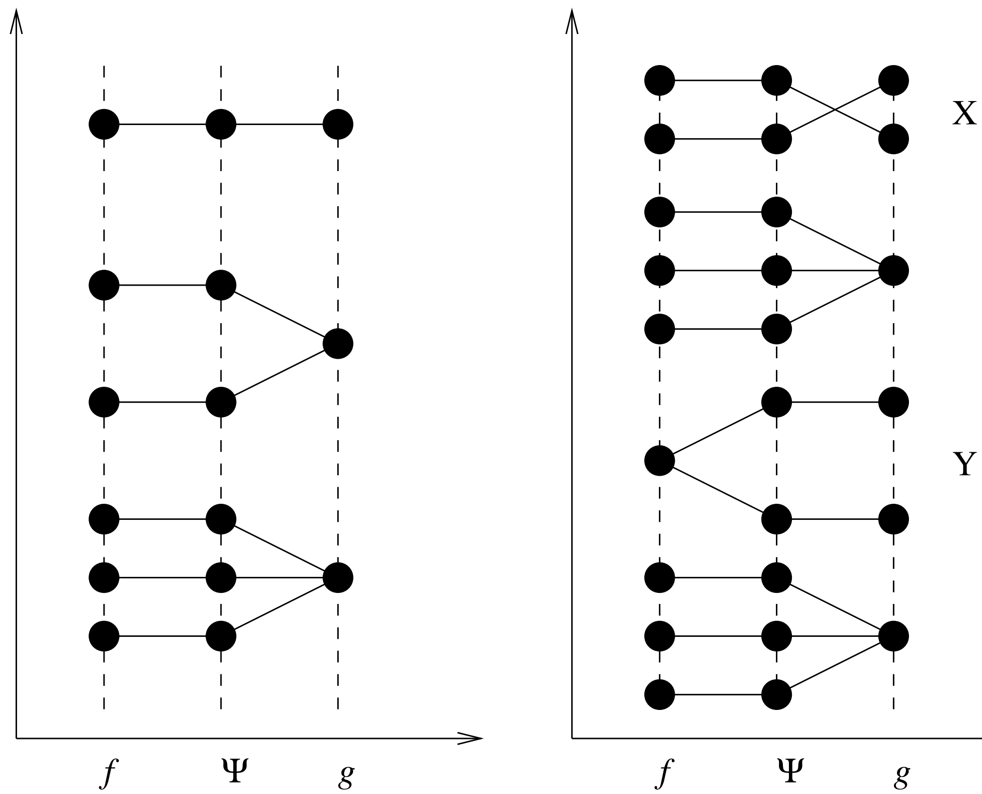
Class Label	Rank	C_1	C_2	C_3
+	10	+	-	+
+	9	+	+	+
+	8	+	+	+
+	7	+	+	-
+	6	-	+	-
-	5	+	-	+
-	4	-	-	+
-	3	-	-	-
-	2	-	-	-
-	1	-	+	-

Classifier	AUC	Error Rate
C_1	$\frac{(5+7+8+9+10)-(5\times6)/2}{5\times5} = \frac{24}{25}$	20%
C_2	$\frac{(1+6+7+8+9)-(5\times6)/2}{5\times5} = \frac{16}{25}$	20%
C_3	$\frac{(4+5+8+9+10)-(5\times6)/2}{5\times5} = \frac{21}{25}$	40%

Comparing Evaluation Measures for Learning Algorithm

- Let Ψ represent the domain and f and g are the two evaluation measures used to compare the learning algorithms A and B
- Consistency: f and g are strictly consistent if there does not exist $a, b \in \Psi | f(a) > f(b)$ and $g(a) < g(b)$
- Discriminancy: f is strictly more discriminating than g if $\exists a, b \in \Psi | f(a) > f(b)$ and $g(a) = g(b)$, and there does not exist $a, b \in \Psi | g(a) > g(b)$ and $f(a) = f(b)$

Consistency and Discriminancy



X is Consistency counter example
Y is Discriminancy counter example

Statistical Consistency and Discriminancy of Two Measures

- Let Ψ represent the domain and f and g are the two evaluation measures used to compare the learning algorithms A and B
- Degree of Consistency: let $R = \{(a,b)|a,b \in \Psi, f(a) > f(b), g(a) > g(b)\}$, $S = \{(a,b)|a,b \in \Psi, f(a) > f(b), g(a) < g(b)\}$. The degree of consistency of f and g is $C (0 \leq C \leq 1)$, where $C = \frac{|R|}{|R|+|S|}$.
- Degree of Discriminancy: 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 and g is $D = \frac{|P|}{|Q|}$.
- The measure f is statistically consistent and more discriminating than g if and only if $C > 0.5$ and $D > 1$. Intuitively, f is better than g.

For AUC and Accuracy Formally

- In domain Ψ let $R = \{(a,b)|a,b \in \Psi, AUC(a) > AUC(b), acc(a) > acc(b)\}$, $S = \{(a,b)|a,b \in \Psi, AUC(a) < AUC(b), acc(a) > acc(b)\}$. Then, $\frac{|R|}{|R|+|S|} > 0.5$ or $|R| > |S|$.
- In domain Ψ let $P = \{(a,b)|a,b \in \Psi, AUC(a) > AUC(b), acc(a) = acc(b)\}$, $Q = \{(a,b)|a,b \in \Psi, acc(a) > acc(b), AUC(a) = AUC(b)\}$. Then $|P| > |Q|$.
- Experimental results to verify the above formal results for balanced or unbalanced datasets
- Experimental results to show that the Naive Bayes classifier is significantly better than decision trees

AUC and Accuracy Experimental Results (balanced)

Statistical Consistency

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

Statistical Discriminancy

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

AUC and Accuracy Experimental Results (unbalanced)

Statistical Consistency

#	$AUC(a) > AUC(b)$ & $acc(a) > acc(b)$	$AUC(a) > AUC(b)$ & $acc(a) < acc(b)$	C
4	3	0	1.0
8	187	10	0.949
12	12716	1225	0.912
16	926884	114074	0.890

Statistical Discriminancy

#	$AUC(a) > AUC(b)$ & $acc(a) = acc(b)$	$acc(a) > acc(b)$ & $AUC(a) = AUC(b)$	D
4	3	0	NA
8	159	10	15.9
12	8986	489	18.4
16	559751	25969	21.6

Conclusions

- AUC is a better measure than accuracy based on formal definitions of discriminancy and consistency
- The above conclusion allows to the re-evaluation of conclusions made using accuracy in machine learning such as, the Naive Bayes classifier predicts significantly better than decision trees. This is contrary to the well-established conclusion of both being equivalent based on the accuracy measure.
- The paper recommends using AUC as a "single number" measure to over accuracy when evaluating and comparing classifiers