Reciever Operator Characteristic Curves and a Performance Measure for Probabalistic Classifiers with Probabilistic Reference Truth

Michael Gredlics
Computing and Digital Media
DePaul University
mgredlics@gmail.com

Daniela Raicu
Computing and Digital Media
DePaul University
draicu@cdm.depaul.edu

Jacob Furst
Computing and Digital Media
Depaul University
jfurst@cdm.depaul.edu

***Abstract*—Evaluation of a classifier's performance has been improved through the use of Receiver Operator Characteristic Curves and summary statistics such as AUC, H-Measure, and area under the cost curve. However, much of the work in this area has been done when the true class or reference truth is known and deterministic. In some instances, such as medical imaging evaluation, experts may disagree on a diagnosis. This work will propose a method for evaluating a classifier with a reference truth that is probabilistic between 0 (negative class) and 1 (positive class) rather than deterministic. In addition, a summary statistic similar to AUC is proposed called AUC$_{PRT}$**

Keywords-machine learning; probabilistic reference truth; probabilistic classifier; ROC curve; AUC; AUC$_{PRT}$

INTRODUCTION

The use of a receiver operator characteristic curve (ROC) has been generally accepted as a graphical method to visualize the performance of a classifier over its operating range without a threshold being defined. The ROC curve can be used for skewed class distributions or unknown cost analysis to understand the strengths and weaknesses of a particular classifier. In addition, the operating ranges where a particular classifier is suboptimal or dominant can be determined in order to form a hybrid classifier that outperforms each of its individual~~both~~ classifiers.

Various one degree of freedom summary statistics have been developed in order to compare the performance of different classifiers. These include the Area Under the Curve (AUC), t~~T~~he Gini Coefficient, the area under the cost curve, and the H-measure.

The basic ROC curve and resulting summary statistics, however, is generated based on the true class being deterministic and known. However, there are many instances when the ground truth is probabilistic and not known. These include situations where there is disagreement of whether an event should be labeled as a 0 (negative) or a 1 (positive) result. Circumstances such as medical readings of X-rays or CT s~~S~~cans by different radiologists or crowd sourcing applications are examples when expert disagreement may exist on a two class problem resulting in a probability for the reference truth between 0 and 1.

BACKGROUND

We start the discussion by considering the existing methods for generating an ROC curve for a two class problem. Each instance has a deterministic true class defined as either a 0 or a 1. This represents the negative and positive class respectively. A classification model is developed to predict (using the attributes of the problem) the true class. The classifier models of interest for this application are the ones that create a probabilistic result between 0 and 1 that represent the likelihood of being a part of the positive or negative class.

A threshold is then determined where if a classifier's value is above the threshold, the instance is predicted to be a 1 (positive) case. Otherwise, the instance is predicted to be a negative case. A 2 x 2 confusion matrix can then be developed to determine the performance of the classifier.

True Class **Predicted**	1 (positive)	0 (negative)
1 (positive)	True Positive (TP)	False Positive (FP)
0 (negative)	False Negative (FN)	True Negative (TN)

Table I. Confusion matrix, where the true class is binary and known and predicted values from the classifier

Many performance statistics can then be computed from the confusion matrix. These include:

$$Accuracy = \frac{TP+TN}{TP+FN+FP+FN} \qquad (1)$$

$$True\ Positive\ Rate = Sensitivity = \frac{TP}{TP+FN} \qquad (2)$$

$$False\ Positive\ Rate = 1 - Specificity = \frac{FP}{TN+FP} \qquad (3)$$

Each of these metrics, however, requires that a threshold is defined prior to evaluating the performance of the classification model. An ROC curve in ROC space can help to visualize the performance of a classifier across all of the thresholds. ROC space is defined as a unit square with the False Positive Rate (FPR) along the horizontal axis and the True Positive Rate (TPR) along the vertical axis.

A method described in (Fawcett, 2006) allows the creation of the ROC curve and is described below. The classifier scores are first sorted from highest to lowest. The threshold is then adjusted from +∞ to -∞ incrementally. The ROC curve starts at the point (FPR=0, TPR=0) which~~and~~ represents the case where all instances are classified as a negative case.

The threshold is then adjusted downward by an increment such that only one instance is above the threshold and the rest are below. If the true class of the instance with the highest classifier score is 1 (positive), a vertical line from (0,0) is drawn of length 1/(TP+FN). This represents the case where the false positive rate remains at 0, however the true positive rate has improved by the rate of 1/(TP+FN). However, if the true class of the instance with the highest classifier score is a 0 (negative), a horizontal line of length 1/(FP+TN).

The threshold is then adjusted to include two instances above the threshold and the process continues. The algorithm is continued until the curve reaches the point (FPR=1, TPR=1). This represents the point where all of the instances are classified as a 1 or positive class. An example of generating a ROC curve is shown in Figure 1~~below~~.

Instance	Classifier Scores (Sorted)	True Class
1	0.8	1
2	0.75	0
3	0.7	1
4	0.6	1
5	0.4	1
6	0.33	0
7	0.3	1
8	0.18	0
9	0.17	1
10	0.16	0
11	0.11	0
12	0.1	0

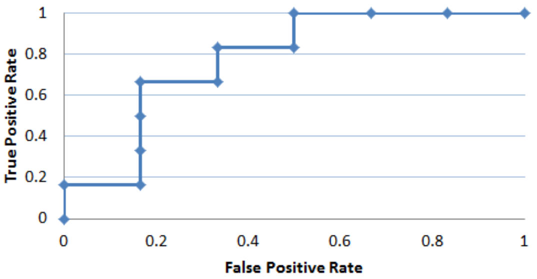

Figure 1. Example of an ROC curve created from the given dataset (blue line). A positive instance in the true class is represented by a ‘1’ and the negative class is represented by a ‘0’. The classifier scores are probabilistic with higher numbers representing a higher likelihood of a positive class. The red line represents a random classifier while the green ROC curve represents a ‘perfect’ classifier.

An ROC curve that climbs vertically from (0,0) to (0,1) and then horizontally to (1,1) represents a classifier where the probabilistic classifier scores of all of the positive true classes are higher than the classifier scores for all of the negative true classes. This ROC curve would indicate a perfect classifier. Similarly, an ROC curve that travels diagonally across the ROC space from (0,0) to (1,1) represents a random classifier.

(Fawcett, 2006) describes a method where two or more classifier scores are equal. In this case, the expected ROC segment for the tied classifier scores is simply a straight diagonal line bisecting the optimistic ROC (where the TPR is increased before the FPR) and the pessimistic ROC (where the FPR is increased before the TPR).

In all of the cases shown, the true class is deterministic and known. However, in cases where the true class is unknown and thus must be estimated by experts and/or novices, the ROC curve cannot be generated by this method. However, the benefits of creating an ROC curve to visually represent the classifier can still be beneficial for researchers. The methodology proposed in this work is a way to generate an ROC curve in the case in which we have a probabilistic reference truth.

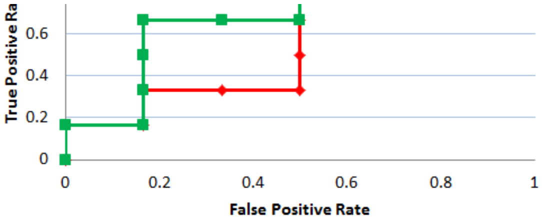

Figure 2. Example of an ROC curve where four of the instances have the same classifier score. The red curve represents the pessimistic ROC curve. The green curve represents the optimistic ROC curve. The blue line represents the expected curve and is the one that is used.

METHODOLOGY

Probabilistic Reference Truth

The remainder of the discussion and the contribution of this paper is to propose a methodology for generating an ROC curve when the reference truth is probabalistic rather than deterministic for the binary case. Situations where this exists is when the experts assigning the truth have a bias or are inconsistent with each other. Each expert is only allowed to evaluate an instance as a 0 (negative class) or a 1 (positive class). Due to differences between the experts’~~’~~s opinions, the resulting reference truth is probabalistic.

Given a dataset, f(i) represents the probabi~~a~~listic reference truth of instance i to be be of class 1. The value of f(i) will be in [0,1] and will be calculated as the average score given by the experts. Let m be the number of instances. Given a classifier, p(i) represents the score given by the classifier for instance i with a higher score indicating a higher likelihood to be of class 1. The value of p(i) will also be in [0,1].

Generating the ROC curve

First, we must calculate the number of positive classes and the number of negative classes from the probabi~~a~~listic reference truth. Let j = the number of instances and

$$Pos_{Classes}=\sum_{i=1}^{m} f(i) \qquad (4)$$

$$\neg_{Classes} ¿ m-\sum_{i=1}^{m} f(i) \qquad (5)$$

In a similar way to (Fawcett, 2006), we will sort the m instances by the classifier scores (p(i)) from 1 to 0. We will have an initial threshold of ∞ which will create a point in ROC space at (FPR=0,TPR=0) which represents the case where we classify all instances with the negative classification.

Next, we reduce the threshold such that only one case (or more if there is a tie) is above the threshold while the rest of the instances are below the threshold. If the reference truth for the instance above the threshold is exactly 1, we will draw a vertical line from (0,0) and of length 1/(Pos_Classes). If the reference truth is exactly 0, we will draw a horizontal line from (0,0) and of length 1/(Neg_Classes). If the reference truth is between [0,1], we will draw a line from (0,0) to (FPR=(1-f(i))/Neg_Classes,TPR=(f(i))/Pos_Classes)~~.~~.

We continue to lower the threshold and draw the resulting ROC curve. The final point on the ROC curve occurs at (FPR=1,TPR=1) and is a result of classifying every instance as the positive class.

Example ROC curve for a Probabalistic Reference Truth

Figure 3 shows an example of a sample dataset along with the the calculation of the first line segment. This line segment represents the change in the True Positive Rate (TPR) and False Positive Rate (FPR) when we move the threshold from +∞ to between 0.8 and 0.75. In the case of a deterministic reference class, the line segment would be horizontal or vertical. However, in the case of a probabalistic reference class, the line segment is angled and dependent on the probability of the reference class being the positive classifier.

Figure 4 shows the entire ROC curve for the example data set. This curve represents the threshold being changed from +∞ to -∞. Similar to ROC curves for deterministic reference truth, the ROC curves for probabalistic reference truth travel from (0,0) to (1,1). The convex hull is shown as a dotted line. The convex hull represents the non-dominated points on the ROC curve for a particular skew ratio range.

Instance	Classifier Scores (Sorted) p(i)	Probabalistic Reference Class f(i)
1	0.8	0.83
2	0.75	0.9
3	0.7	0.72
4	0.6	0.25
5	0.4	0.34
6	0.33	0.5
7	0.3	0.4
8	0.18	0.96
9	0.17	0.29
10	0.16	0.6
11	0.11	0.2
12	0.1	0.15

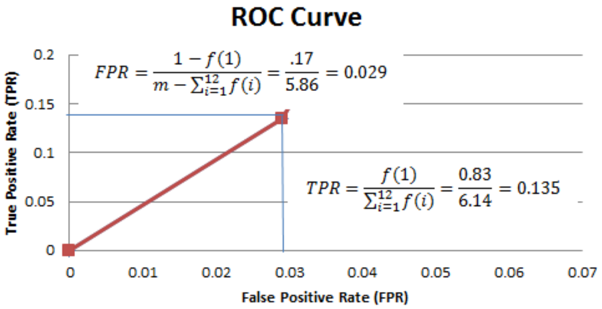

Figure 3. Example of the first line segment on an ROC curve for the sample dataset shown. The sample dataset has a probabilistic true class as well as a probabilistic classifier. The plot shows the calculation for the FPR and TPR when the threshold is between 0.75 and 0.80.

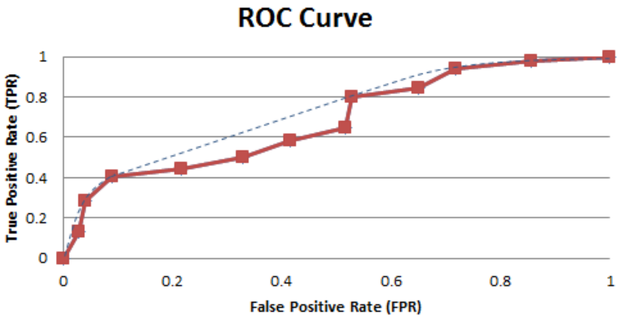

Figure 4. Example of the full ROC curve for the sample dataset shown. The dotted line represents the convex hull.

Discussion

Interpretation of ROC Curve for Probabalistic Reference Truth

In the case of a deterministic reference truth, a perfect classifier is considered to be a curve that climbs vertically from (FPR=0,TPR=0) to (FPR=0,TPR=1) and then travels horizontally to (FPR=1,TPR=1). In the case where the reference truth is probabalistic and contains values between (0,1), a perfect classifier will never reach this point. This is due to the uncertainty in the reference truth.

Figure 5 shows the perfect classifier for the dataset shown in Figure 3. Notice that the uncertainty in the reference truth affects the curve to be away from including the (0,1) point. The uncertainty in the reference truth is graphically shown to be the Area Above the Perfect Classification Curve (AAPCC). If there is little uncertainty in the reference truth (i.e. the reference truth for all instances is close to either 0 or 1), the perfect classifier curve will approach the (0,1) point. However, if the dataset contains many instances with high uncertainty in the reference truth (i.e. the reference truth for most instances is close to 0.5), the perfect classifier will approach the diagonal between (0,0) and (1,1).

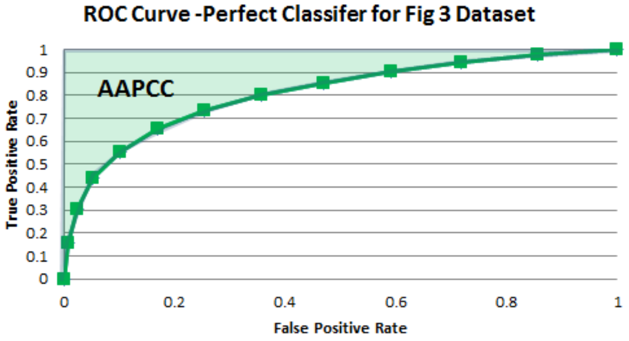

Figure 5. Perfect classifier for dataset with probabilistic reference truth found in Figure 3. Notice that the perfect classifier does not reach the (0,1) point due to the uncertainty in the ground truth. The area above the perfect classifier curve (AAPCC) represents the uncertainty in the experts determining the true class.

In this regard, we can say that:

- AAPCC is equal to 0 for datasets with deterministic reference truth.
- AAPCC is close to 0 for datasets with little uncertainty in the probabalistic reference truth
- AAPCC is close to 0.5 for dataset with high uncertainty in the probabalistic reference truth.
- AAPCC is equal to 0.5 for datasets where the f(i) for all instances is equal to 0.5.

Similarly, the worst classifier for a probabialistic reference truth dataset does not go through the point (FPR=1,TPR=0). Instead, it is a mirror image of the perfect classification curve across the diagonal in ROC space. Figure 6 shows the area under the worst classifier curve (AUWCC). The AAPCC is equal to the AUWCC.

The random classifer for a probabialistic reference truth is the same as for the deterministic case. A random classifier on a dataset with many more instances will approach the diagonol from (0,0) to (1,1).

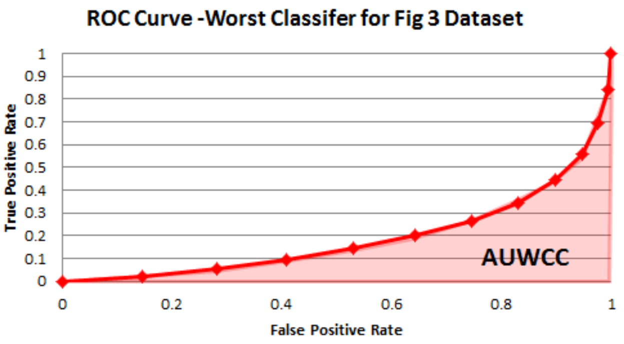

Figure 6. Worst classifier for dataset with probabilistic reference truth found in Figure 3. Notice that the perfect classifier does not reach the (1,0) point due to the uncertainty in the ground truth. The area below the worst classifier curve (AUWCC) represents the uncertainty in the experts determining the true class.

Area Under the Curve metric for probabalistic reference truth from a binary decision

The Area Under the ROC Curve (AUC) is an often used performance metric for a classifier. The uncertainty in the reference truth along with the error in the classifer, however, mean that the AUC_{prt} (Area Under the ROC curve for probabalisitic reference truth) must be defined diferently than the traditional AUC. The AUC_{prt} will be defined as the area between the classifier ROC curve and the worst performing classifier (AUWCC).

$$AUC_{PRT}=\frac{AUC-AUWCC}{1-2(AUWCC)}$$

(6)

The AUC_{PRT} will have a value of 1 when the classifier performs perfectly for a given dataset and a value of 0 when the classifier is the worse possible classifier for a given dataset.

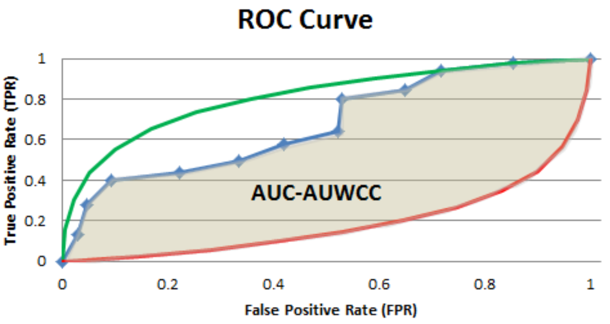

Figure 7. Graphical depiction of the numerator of the AUC_{PRT} equation. The denominator is the area of the region between the AUWCC and the AAPCC curve.

Advantages

One advantage of using ROC graphs to graph a classifier for the case of probabalistic reference truth is that we can visualize the performance without concerns for the ~~a~~ skewed distribution in the f(i). This allows a researcher to understand the performance of a classifier even if there are a limited number of positive (or highly probable) positive classes. This is due to the fact that the true positive rate and the false positive rate are evaluated separately.

Another advantage of using ROC graphs and the method described for a probabi~~a~~listic reference truth is that a researcher can find the classifier that gives the best performance for a particular operating range (Provost, Fawcett and Kohavi,1998; Fawcett 2006) using the convex hull. In addition, a researcher can determine if a classifier is completely dominated by another classifier and thus should not be used.

Finally, researchers can define iso-performance lines for a particular skew ratio and find the threshold that will match their goals

Disadvantages of the summary statistic AUC_{PRT}

As the AUC_{PRT} is an extension of the AUC, it suffers from the same weaknesses as a measure of classification performance. Many of these weaknesses have been stated in previous work in (Hand, 2009). Researchers have suggested alternate summary statistics including the H-Measure (Hand, 2009) and the area under the cost curve (Flach, Hernandez-Orallo and Ferri, 2011). These approaches to converting the multi degree of freedom ROC curve to a single degree of freedom summary statistic can be done with this method of generating an ROC curve for probabalistic reference truth datasets.

Future Work

Future work in this area will be done on validating the use of cost curves as well as calculating the area under the cost curve and the H-measure for ROC curves generated in the manner outlined in this work.

Conclusion

In this work, we have shown a method to generate an ROC curve for datasets when the reference truth is probabilistic. A probabilistic reference truth can occur in many types of situations including medical diagnosis when experts disagree to using crowd sourcing to evaluate images. The method shown involves creating and analyzing an ROC curve that is familiar to many researchers.

In addition, this paper showed a method of creating a summary statistic similar to AUC that can be used to measure the ranking performance of a classifier. The standard AUC measure is not sufficient for the case of a probabilistic reference truth as the perfect classifier on such a dataset would not reach 1. The reason is due to the uncertainty in the expert's reference truth.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation under Grant No. 1062909.

REFERENCES

C. Drummond and R. Holte, “Cost curves: an improved method for visualizing classifier performance,” Mach. Learn, vol. 65, Oct. 2006, pp.95-130.

T. Fawcett, “An introduction to ROC analysis,” Pattern Recogn., vol. 27, Jun. 2006, pp.861-874.

T. Fawcett, “ROC graphs: notes and practical considerations for data mining,” Technical report, HPL-2003-4.

C. Ferri, J. Hernández-Orallo, and R. Modroiu,”An experimental comparison of performance measures for classification,” Pattern Recogn., vol. 30, Jan. 2009, pp 27-38.

P. Flach, J. Hernández-Orallo, and C. Ferri, “A coherent interpretation of the AUC as a measure of aggregated classification performance,” Proc. of the 28th Internat. Conf on Machine Learning (ICML-’11), June 2011, pp. 657-664.

P. Flach, “The geometry of ROC space: understanding machine learning metrics through ROC isometrics,” Proc. of the 20th Internat. Conf on Machine Learning (ICML-’03), Jan. 2003, pp.194-201

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(2), Nov. 2001, pp.171-186.

D.J. Hand, “Assessing the performance of classification methods,” International Statistical Review, vol. 80(3), Aug. 22, 2012, pp. 400-414.

D.J. Hand, “Measuring classifier performance: a coherent alternative to the area under the ROC curve,” Machine Learning, vol 77(1), Oct. 2009, pp.103-123.

C.X. Ling, J. Huang, and H. Zhang, “AUC: a statistically consistent and more discriminating measure than accuracy,” Proc. of Eighteenth Internat. Conf. on Artifical Intelligence (IJCAI-2003), Aug. 2003, pp.519-526..

F. Provost and T. Fawcett, “Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions,” Proc. of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD-’97), pp. 43-48.

F. Provost, T. Fawcett, and R. Kohavi, “The case against accuracy estimation for comparing induction algorithms” Proc. of the Fifteenth Internat. Conf. on Machine Learning (ICML-’98), pp. 445-453.

F. Provost and T. Fawcett, “Robust classification for imprecise environments”. Machine Learning, 42, 203-231.

D. Zinovev, J. Furst, and D. Raicu, “Building an ensemble of probabilistic classifiers for lung nodule interpretation”. Tenth Intern. Conferen on Mach. Learn. and App. (ICMLA-2011), IEEE Press, Dec. 2011, pp.151-167, doi: 10.1109/ICMLA.2011.44.