Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005

ROC ANALYSIS FOR PREDICTIONS MADE BY PROBABILISTIC CLASSIFIERS

ZENG-CHANG QIN

Artificial Intelligence Group, Department of Engineering Mathematics, University of Bristol, BS8 1TR, UK
E-MAIL: z.qin@bris.ac.uk

Abstract:

Receiver Operating Characteristics (ROC) analysis was originated from signal detection theory and has been introduced to machine learning community in recent years to evaluate the algorithm performance under imprecise environment. ROC graphs have become increasingly popular in machine learning, because they offer a more robust framework for evaluating classifier performance than the traditional accuracy measure. In this paper, we investigate the relation between a probabilistic classifier and its corresponding predictor in a view of ROC analysis. A method of generating ROC curves for prediction (or regression) problems is proposed and some properties of ROC curves for prediction are discussed with examples.

Keywords:

ROC Analysis; AUC; ranker; probabilistic classifier

1. Introduction

Traditionally, the main criterion for evaluating the performance of a classifier is accuracy (percentage of test examples that are correctly classified) or error (percentage of misclassified examples). However, in many situations, not every misclassification has the same consequences when misclassification costs have to be taken into account. Recent study shows that the accuracy of a classifier is also influenced by the class distribution [7]. Receiver Operating Characteristics (ROC) analysis, which was originated from signal detection theory, has been introduced to evaluate machine learning algorithms [1, 2, 6, 7] and it has become increasingly popular in machine learning research. In addition to being a generally useful performance graphing method, they have properties that make them especially useful for domains with skewed class distribution and unequal classification error costs. For example, given a classifier which has accuracy of 80%. The accuracy doesn't make sense without knowing the class distribution: If the database consists of 90% positive and 10% negative examples. We can do better simply by classifying all the data as positive that will give 90% accuracy. So, ROC analysis is not just about cost-sensitive learning, it considers the relative importance of negative vs. positive examples. This relative importance can be represented by a skew ratio by considering both costs and class distribution [3].

Many classifiers not only give discreet predicted classes but also the estimates of class membership probabilities (e.g., Naive Bayes). The former are referred to as discreet classifiers and the latter as probabilistic classifiers or rankers, because the membership probabilities can be used to rank instances from most to least likely positive. By setting a threshold, a rankers can act as a classifier. Area under the curve (AUC) of ROC is used to measure the quality of ranking for a probabilistic classifier [4, 8]. Ling *et al.* proved that AUC is statistically consistent and more discriminating than the accuracy measure [5]. So, it is fair to use AUC rather than accuracy to evaluate a learning algorithm. Currently, all the ROC analysis research are for classification problems. However, in many real-world applications, data ranging from financial analysis to weather forecasting are prediction problems. We are wondering if we can extend to the ROC analysis to predictions? This is the motivation of this research. Here in this paper, some initial investigations is presented where we only consider the predictors based on defuzzification with two fuzzy labels on probabilistic classifiers.

This paper is organized as follows: we first introduce the basics of ROC analysis for classification in section 2. In section 3, the ROC analysis is extended to prediction problems and the method for plotting ROC curves is proposed. In the section 4, the method of calculating AUC values is proposed and some special properties of ROC curves for prediction are discussed with examples.

2. ROC Analysis for Classification

Traditionally, accuracy and error are widely used measures for evaluating performance of a classifier. Using accuracy as a performance measure assumes that the error

0-7803-9091-1/05/$20.00 ©2005 IEEE

costs are equal and the class distribution is balanced. However, this is not realistic if we consider problems such as medical diagnosis or fraud detection. We begin by considering classification problems using only two classes (or binary classification problem). Usually, the instances are divided according to the following contingency table or confusion matrix:

	$(\hat{P})$ Predicted Positives	$(\hat{N})$ Predicted Negatives
(P) Positive Examples	(TP)-True Positives	(FN)-False Negatives
(N) Negative Examples	(FP)-False Positives	(TN)-True Negatives

Figure 1. Confusion matrix for a binary (i.e., positive and negative) classification problem.

If the number of positives are negatives are denoted by P and N, respectively, the predicted positives and negatives are denoted by $\hat{P}$ and $\hat{N}$, then, the classification accuracy is defined as:

$$Accuracy = \frac{TP + TN}{P + N} \tag{1}$$

ROC analysis decomposes performance into true and false positive rates defined as follows: the true positive rate (TPR) of a classifier is: $TPR = TP / P$ and the false positive rate (FPR) of a classifier is: $FPR = FP / N$. If we plot FPR on the X axis and TPR on the Y axis. A single classification is then represented by a point in this 2D coordinate space which is referred to as ROC space. In the ROC space, the upper left point (0, 1) is most wanted because it gives 100% percent of true positives and zero false positives. It can be called as "ROC Heaven" and, correspondingly, the point (1, 0) is the least wanted point that can be called "ROC Hell" [3]. The diagonal line represents a random classifier which always gives 50% of true positive rate and 50% false positive rate. Each discreet classifier can be presented by a single point according to its TPR and FPR in the ROC space. Different ROC profiles will be more or less desirable under different class distributions and different error cost functions. More details about basic ROC space properties are available in [3].

Consider a probabilistic classifier with two classes '+' and '-'. We can sort the instances according to the probabilities of belonging to class +. Different classification

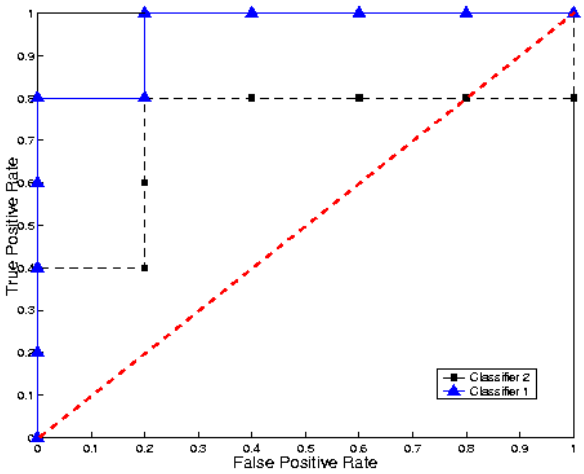

Figure 2. ROC curves for the classifiers CLS_1 and CLS_2.

results will be given according to the varying threshold T based on:

$$\forall i : \begin{cases} \vec{x}_i \to \{+\} & if : P(+ \mid \vec{x}_i) \geq T \\ \vec{x}_i \to \{-\} & Otherwise \end{cases}$$

where we normally set $T = 0.5$ [1] when we calculate accuracy for a probability estimation model. If we vary the value of T through [0, 1], it will result a continuous curve in ROC space which is referred to as a ROC curve. In another words, A classifier results in a ROC curve, which aggregates its behavior for all possible decision thresholds. The quality of the classifier can be measured by the area under the curve of ROC (AUC), which measures how well the classifier separates the two classes without reference to a decision threshold. In other words, AUC represents the quality of ranking of examples by this classifier. Given k instances, there are only k+1 possible thresholds. A practical method is as follows: (1) rank test instances on decreasing membership scores. (2) Starting in (0, 0), if the next instance in the ranking list is positive then move $1/P$ up, if it is negative then move $1/N$ to the right. Given the two classifiers CLS_1 and CLS_2 in table 2, the ROC curves drawn by the above method are shown in figure 2. According to Hand and Till [4], the AUC value for a binary classification problem with two classes {+, −} can be calculated by:

$$AUC = \frac{\sum_{i=1}^{P} r_i - P(P+1)/2}{PN} \tag{2}$$

[1] The optimal threshold for a probabilistic classifier depends on the class distribution and misclassification costs. The membership scores are not calibrated estimate of probabilities in most cases [9]. Therefore, assigning T=0.5 (e.g., for Naïve Bayes classifier) is a misleading in many machine learning literatures.

Table 1. Two classifiers (CLS_1 and CLS_2) with the same accuracy but different AUC values. This table is inspired by a similar table in [5].

Examples	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	$\mathbf{x}_9$	$\mathbf{x}_{10}$
CLS_1	-	-	-	-	+	-	+	+	+	+
r_i for LCS_1					5		7	8	9	10
CLS_2	+	-	-	-	-	+	+	-	+	+
r_i for LCS_2	1					6	7		9	10

where P and N are the number of positive and negative examples, respectively. r_i is the rank of ith positive example in the ranking list according to the probabilities of the class +. For example, the AUC values for classifier 1 and 2 listed in table 2 are:

$$AUC_{(CLS_1)} = \frac{(5+7+8+9+10)-5\times(5+1)/2}{5\times 5} = \frac{24}{25}$$

$$AUC_{(CLS_2)} = \frac{(1+6+7+9+10)-5\times(5+1)/2}{5\times 5} = \frac{18}{25}$$

We may notice that both classifier 1 and 2 have the same accuracy 80% (8 of 10 examples are correctly classified) and thus they are equally good in accuracy. However, the intuition tells us that classifier 1 is better than classifier 2 since classifier 1 gives a better overall ranking. This can be seen from AUC measure but not the accuracy measure. Ling *et al.* [5] mathematically proved that the AUC measure is consistent and more discriminating than the accuracy measure. The method for calculating AUC for multi-class problems is given in [4], however, in this paper, only two-class problems are considered.

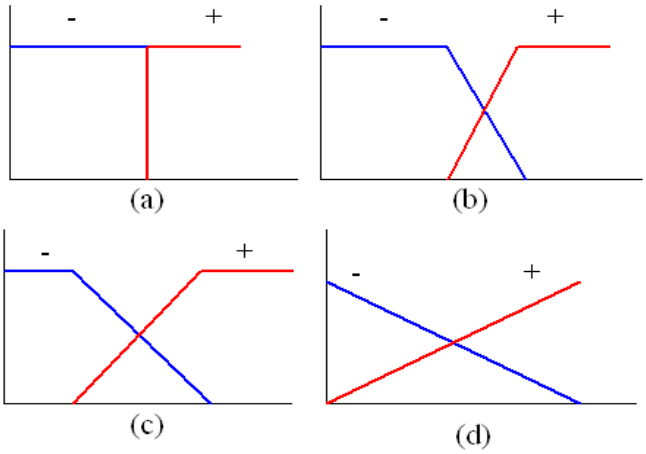

Figure 3. Different degrees of overlapping between two fuzzy labels that are used as class labels.

3. ROC Analysis for Prediction

Consider a prediction problem that the output space or target attribute t is numeric. For each instance $\mathbf{x}$ (a multidimensional vector, it also written as $\vec{x}$ in equations) to predict its target value t (i.e. $\mathbf{x}_i \rightarrow t_i$). Suppose we discretize the output universe with m fuzzy sets: $F_1, \ldots, F_m$. We can consider each fuzzy set as a single class label that has weights denoted by and each instance can be mapped to a representation as follows:

$$\vec{x}_i \rightarrow < F_1 : \xi_1, \ldots, F_m : \xi_m >$$

where,

$$\sum_{i=1}^{m} \xi_i = 1$$

We then can use an arbitrary probabilistic classifier to obtain a series of conditional probabilities on target fuzzy sets given a test instance $\mathbf{x}$: $P(F_1|\mathbf{x}), \ldots, P(F_m|\mathbf{x})$. The estimate of t, denoted $\hat{t}$ to be the expected value:

$$\hat{t} = \int_{\Omega} tp(t \mid \vec{x})dt \tag{3}$$

where:

$$p(t \mid \vec{x}) = \sum_{j=1}^{m} p(t \mid F_j)P(F_j \mid \vec{x}) \tag{4}$$

and

$$p(t \mid F_j) = \frac{M_t(F_j)}{\int_{\Omega_t} M_t(F_j)dt} \tag{5}$$

where $M_x(F_j)$ is the membership of x belonging to fuzzy set of labels F_j. So that we can obtain:

$$\hat{t} = \sum_{j=1}^{m} P(F_j \mid \vec{x})E(t \mid F_j) \tag{6}$$

where:

$$E(t \mid F_j) = \frac{\int_{\Omega_t} tM_t(F_j)dt}{\int_{\Omega_t} M_t(F_j)dt} \tag{7}$$

where the process of calculating $E(x_t|Fj)$ is also called defuzzification in some other literatures.

From above we can see that, by fuzzifying the continuoustarget attribute t into intervals that could be considered as class labels, any probabilistic classifiers can be extended to a prediction model. However, we need to notice that the class labels not discreet but overlapped each other and there are many different degrees of overlapping. For example, figure 3 shows four different possible overlapping. In this paper, we only consider the simplest case that $m = 2$, where one fuzzy label is represented by − and the other by +. In the following paper, unless otherwise stated, we will use the fuzzy labels with 50% overlapping (figure 3-d), it satisfies:

$$\forall i : P(- \mid \vec{x}_i) + P(+ \mid \vec{x}_i) = 1$$

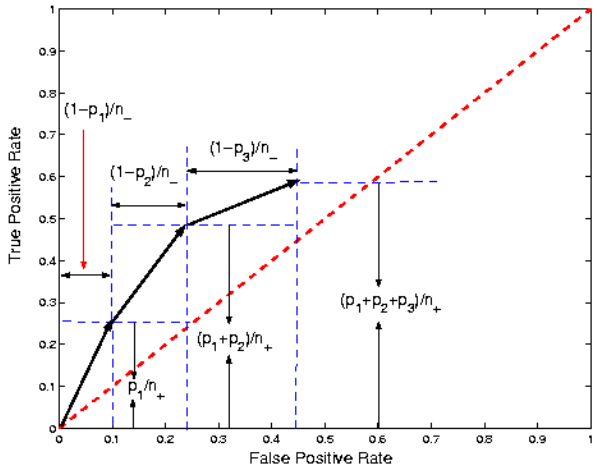

Figure 4. An illustration of drawing ROC curve and AUC calculation by adding a new point.

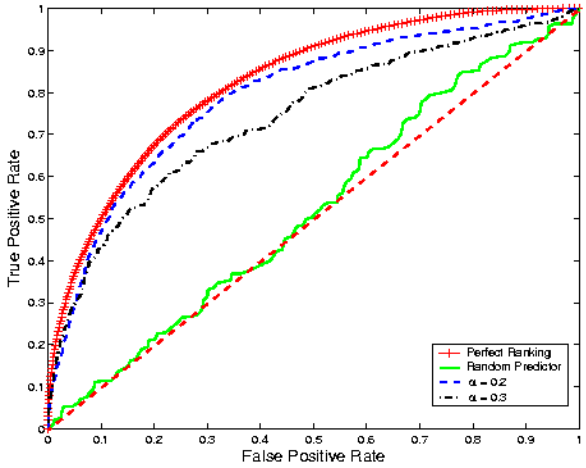

Figure 5. ROC curves for a perfect-ranking predictor, a random predictor, and two predictors obtained by perfect ranking predictor corrupted by different levels of noise.

The basic difference between such predictors and normal probabilistic classifiers is that the class labels are overlapped each other. For a particular instance, it has original membership probabilities of positives from fuzzy discretization $P(+|\mathbf{x})$ and predicted class probabilities $\hat{P}(+|\mathbf{x}_i)$ from classifiers In the following context, unless otherwise stated, we will focus on the membership probabilities of positives and we write $\hat{P}(+|\mathbf{x}_i)$ as p_i and $\hat{P}(+|\mathbf{x}_i)$ as $\hat{p}_i$. For example, given the original membership scores and predicted scores, how can we draw the ROC curve? A simple and practical method based on discreet class labels is proposed as follows:

- *Given a test set of size L, rank the instances on decreasing predicted membership scores of the 'positive' class $\hat{p}_i$, where $i \in \{1, 2, \ldots, L\}$.*
- *$TP_0 = 0$, $FP_0 = 0$*
- *for $i = 1 : L$, Do:*
  $TP_i = TP_{i-1} + p_i/n_+$, $FP_i = FP_{i-1} + (1 - p_i)/n_-$
- *Starting from (0, 0), for $i = 1 : L$, draw the curve by joining (FP_{i-1}, TP_{i-1}) and (FP_i, TP_i) successively.*

where n_+ is the sum of positive parts on all examples, it is obtained by:

$$n_+ = \sum_{i=1}^{L} p(i|+) = \sum_{i=1}^{L} p(i) \tag{8}$$

Similarly, we can obtain:

$$n_- = \sum_{i=1}^{L} p(i|-) = \sum_{i=1}^{L} 1 - p(i) = 1 - n_+$$

Figure 5 shows a set ROC curves on a real-world prediction problem: the marked curve is a perfect ranking, which means that given a ranked list in a decreasing manner based on $\hat{p}_i$ (i.e. $\hat{p}_1 \geq \ldots \geq \hat{p}_L$), the relation $p_1 \geq \ldots \geq p_L$ holds. The curves marked with α= 0.2 represents a perfect ranking predictor corrupted by a uniform distributed noise in the range of [0, 0.2], denoted by U[0, 0.2]. So that the predicted probabilities are:

$$\forall i \quad \hat{p}_i = p_i \pm \varepsilon \qquad \varepsilon \in U[0, \alpha]$$

The random classifier is a random guess that follows:

$$\forall i \quad \hat{p}_i \in U[0,1]$$

As we can see from those curves, they exhibit similar properties as with discreet labels, the only difference is that the maximum value for prediction is not 1. This will be discussed in details in the next section.

4. AUC Value for Prediction

Figure 4 gives an illustration of drawing ROC curve for such prediction problems. We need to notice that the optimal point is not (1,0) for predictions (i.e. AUC value is always less than 1). The reason for this is because we use

overlapping fuzzy labels. ROC analysis reflects the separation of positive and negative examples by a classifier. In this case, no matter how good a classifier is, it still cannot completely separate the positives and negatives because they are overlapped to each other. The different overlapping degrees will result in different maximum AUC values. Figure 6 depicts the ROC curves with maximum AUC values on the fuzzy labels with different overlapping degrees shown in figure 3. In the legend, the AUC values that are calculated by the method that will be discussed in the following part of this section.

Consider the ranking list on decreasing membership scores in the way we draw the ROC curves. The first example of the ranking list is the one with the highest predicted score with original score of p_1. By adding this example to the ROC space, the area under the ROC curve is a triangle with side lengths of $(1-p_1)/n_-$ and p_1/n_+, respectively (see figure 4). So that

$$AUC_1 = \frac{p_1(1-p_1)}{2n_+n_-} \tag{9}$$

By adding a new example with score extended and a new area in trapezoidal the current AUC becomes:

$$AUC_2 = \frac{1}{n_+n_-}\left[\frac{p_1(1-p_1)}{2} + (1-p_2)\frac{p_1+(p_1+p_2)}{2}\right]$$

Similarly, by adding the third point:

$$AUC_3 = \frac{1}{n_+n_-}\left[\frac{p_1(1-p_1)}{2} + (1-p_2)\frac{p_1+(p_1+p_2)}{2} + (1-p_3)\frac{(p_1+p_2)+(p_1+p_2+p_3)}{2}\right]$$

By successively adding the kth example ($k \neq 1$), we can obtain:

$$AUC_k = \frac{1}{2n_+n_-}\sum_{i=1}^{k}(1-p_i)\left(2\sum_{j=1}^{i-1}p_j + p_i\right) \tag{10}$$

Equation 10 can be rearranged and the AUC value for prediction on a test set with L examples is:

$$AUC = \frac{1}{2n_+n_-}\left[\sum_{i=1}^{L}p_i(1-p_i) + 2C\right] \tag{11}$$

where,

$$C = \sum_{i=2}^{L}\sum_{j=1}^{i-1}p_j(1-p_i) \tag{12}$$

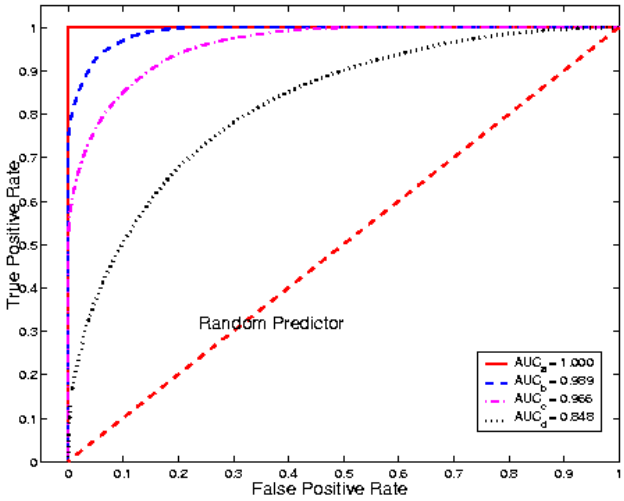

Figure 6. ROC curve with maximum AUC values given fuzzy labels on different degrees of overlapping in figure 3.

Consider the equation 11, the first term $\sum_{i=1}^{L}p_i(1-p_i)$ is invariant to different rankings. Now we only consider the term C to investigate the relation between AUC value and example ranking. Term C can be separated into two terms $A_i = 1 - p_i$ and $B_i = \sum_{j=1}^{i-1}(1-p_j)$ so that $C = \sum_i A_iB_i$. Suppose we have the following ranking of examples according to the B terms:

R_p : … … $1-p_k$, $1-p_{k+1}$ … …

if we swap the positions of these two examples to:

R_x : … … $1-p_{k+1}$, $1-p_k$ … …

Suppose $p_k \geq p_{k+1}$, such swapping is referred to as bad swapping, because R_p is more desirable than R_x for a better ranking. The swapping will result in a change in AUC values, if we define:

$$D(R_p) = \sum_{i=2}^{k+1}A_iB_i = (1-p_k)T + (1-p_{k+1})(T+p_k)$$

$$D(R_x) = \sum_{i=2}^{k+1}A_iB_i = (1-p_{k+1})T + (1-p_k)(T+p_{k+1})$$

where $T = \sum_{j=1}^{k-1}p_j$ and according to equation 12 we obtain:

$$C(R_p) = D(R_p) + \sum_{i=k+2}^{L}\sum_{j=1}^{i-1}p_j(1-p_i)$$

$$C(R_x) = D(R_x) + \sum_{i=k+2}^{L}\sum_{j=1}^{i-1}p_j(1-p_i)$$

Table 2. AUC vales with different rankings by exchanging examples from the perfect ranking, where n_i=1-p_i.

	n_1	n_2	n_3	n_4	n_5	$\frac{p_k - p_{k+1}}{n_+ n_-}$	AUC
R_p	0.2	0.4	0.6	0.8	1.0	0	0.8333
R_{x1}	0.2	0.4	0.8	0.6	1.0	0.0333	0.8000
R_{x2}	0.2	0.8	0.4	0.6	1.0	0.0667	0.7333
R_{x3}	0.2	0.8	0.4	1.0	0.6	0.0667	0.6667

The latter terms for $C(R_P)$ and $C(Rx)$ have identical values. Therefore, according to equation 11, we can calculate the change of AUC values by exchange these two examples as follows:

$$AUC(R_p) - AUC(R_x) = \frac{1}{n_+ n_-}[C(R_p) - C(R_x)]$$

$$= \frac{1}{n_+ n_-}\left[D(R_p) - D(R_x)\right] = \frac{p_k - p_{k+1}}{n_+ n_-} \geq 0$$

where the equality holds when $p_k = p_{k+1}$. If we suppose $p_k \leq p_{k+1}$, such a swapping then becomes a good swapping, the AUC values will be increased by the same value. For example, we start from a perfect ranking R_p shown in table 4. We can obtain:

$$n_- = \sum_i n_i = 0.2 + 0.4 + 0.6 + 0.8 + 1 = 3$$

and $n_+ = 5-3 = 2$. By swapping 0.8 and 0.6, we obtain the change in AUC as follows:

$$\frac{p_k - p_{k+1}}{n_+ n_-} = \frac{(1-0.6)-(1-0.8)}{2 \times 3} = 0.0333$$

So that the new AUC value for the rearranged list is

$$AUC(R_{x1}) = AUC(R_p) - 0.0333 = 0.8000$$

Based on the new ranking list R_{x_1}, swap 0.8 and 0.4, we then can obtain a new ranking list R_{x_2}, such that:

$$\frac{p_k - p_{k+1}}{n_+ n_-} = \frac{(1-0.4)-(1-0.8)}{2 \times 3} = 0.0667$$

$$AUC(R_{x_2}) = AUC(R_{x_1}) - 0.0667 = 0.7333$$

Similarly, we can obtain another bad ranking list R_{x_3} by such swapping (i.e. bad swapping) and the AUC values will keep decreasing.

5. Conclusions

In this paper, we extended ROC analysis that is commonly used in classification to prediction. A method of drawing ROC curves for prediction is proposed and some of important properties of such ROC are discussed. By introducing the method for calculating AUC values for prediction, we also investigate the relation between the AUC values and the ranking of examples. In particular, a quantitative analysis of AUC value by swapping two neighboring instances is given. However, in this paper, we only consider a very simple case that the predictor is obtained by defuzzification of probabilistic classifiers. The future research focus on extending this framework to multi-classes (more than 2 fuzzy labels) and study the relation between AUC and some other measures used in prediction such as mean squared error and average error.

Acknowledgements

The author thanks Prof. Peter Flach for useful discussions on ROC analysis that inspired the research presented in this paper.

References

[1] T. Fawcett, "ROC graphs: notes and practical considerations for data mining researchers", HP Technical Report HPL-2003-4, HP Laboratories, 2003.

[2] P. A. Flach, "The geometry of ROC space: understanding machine learning metrics through ROC isometrics", Proceedings of the ICML-04, 2004.

[3] P. A. Flach, "The many faces of ROC analysis in machine learning", ICML-2004 Tutorial, Notes available at:*http://www.cs.bris.ac.uk/flach/ICML04tutorial/index.html*

[4] D. Hand and R. J. Till, "A simple generalisation of the area under the ROC curve for multiple class classification problems", Machine Learning, Vol. 45, 171-186, 2001.

[5] C. X. Ling, J. Huang and H. Zhang, "AUC: a statistically consistent and more discriminating measure than accuracy", Proceedings of IJCAI-03, 2003.

[6] F. Provost, T. Fawcett, and R. Kohavi, "The case against accuracy estimation for comparing induction algorithms," in J. Shavlik, editor, Proceedings of ICML-98, pp. 445-453, 1998.

[7] F. Provost and T. Fawcett, "Robust classification for imprecise environments", Machine Learning. Vol. 42, 203-231, 2001.

[8] F. Provost and P. Domingos, "Tree induction for probability-based ranking",Machine Learning. 52, 199-215, 2003.

[9] [9] B. Zadrozny and C. Elkan, "Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers", Proceedings of ICML-01, 2001.