QUANTITATIVE METHODS IN ECONOMICS
Vol. XV, No. 2, 2014, pp. 382 – 391

ESTIMATING THE ROC CURVE AND ITS SIGNIFICANCE FOR CLASSIFICATION MODELS' ASSESSMENT

Krzysztof Gajowniczek, Tomasz Ząbkowski
Department of Informatics, Faculty of Applied Informatics and Mathematics,
Warsaw University of Life Sciences - SGGW
e-mail: krzysztof_gajowniczek@sggw.pl, tomasz_zabkowski@sggw.pl
Ryszard Szupiluk
Warsaw School of Economics
e-mail: rszupi@sgh.waw.pl

Abstract: Article presents a ROC (receiver operating characteristic) curve and its application for classification models' assessment. ROC curve, along with area under the receiver operating characteristic (AUC) is frequently used as a measure for the diagnostics in many industries including medicine, marketing, finance and technology. In this article, we discuss and compare estimation procedures, both parametric and non-parametric, since these are constantly being developed, adjusted and extended.

Keywords: ROC curve, AUC, classification models' assessment

INTRODUCTION

Plotting the ROC curve is a popular way for discriminatory accuracy visualization of the binary classification models and the area under this curve (AUC) is a common measure of its exact evaluation. ROC methodology is derived from signal detection theory developed during the II World War where it was used to determine if an electronic receiver is able to distinguish between the signal and the noise. Nowadays, it has been used for the diagnostics in medical imaging and radiology[Hanley and McNeil 1982], psychiatry, manufacturing inspection systems, finance and database marketing.

The ROC analysis is useful for the following reasons: (1) evaluation of the discriminatory ability of a continuous predictor to correctly assign into a two-group classification; (2) an optimal cut-off point selection to least misclassify the two-group class; (3) compare the efficacy of two (or more) predictors.

Many parametric and non-parametric estimation methods have been proposed for estimating the ROC curve and its associated summary measures. In this study, we focus on three methods which have been mostly employed in practical applications. In the following sections of the article we introduce notation and the basic concepts of the ROC curve and AUC measure. The further sections are devoted to one parametric and two non-parametric methods of ROC and AUC estimation. The paper ends with a simulation study and short discussion in the last section.

MEASURES OF BINARY CLASSIFICATION PERFORMANCE

Determination of the ROC curve and the area under the curve is related to the classification matrix construction (Table 1) and calculation of sensitivity and specificity measures.

Table 1. Classification matrix

		Predicted value	
		Positive (P)	Negative (N)
Real value	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

Source: own preparation

ROC curve is a set of points: $(x, y) : x = 1 - specificity, y = sensitivity$ where for a particular decision threshold value u sensitivity and specificity is determined. Sensitivity is a ratio of true positive cases to all real positive cases:

$$se = \frac{TP}{TP + FN} \tag{1}$$

whilst specificity determines the share of true negatives cases to all real negative cases:

$$sp = \frac{TN}{FP + TN}. \tag{2}$$

The interpretation of these measures is as follows. Sensitivity is the ability of the classifier to detect instances of a given class (the conditional probability of classification for the selected class, provided that the object actually belongs to it). In turn, specificity determines the extent to which the decision classifier of belonging to the selected class is characterized by the class (supplement conditional probability of classification for the selected class, provided that the object of this class should not be).

It should be noted that the output values generated by the model (e.g. neural network, logistic functions) belong to a certain range, therefore, the threshold should be determined on the basis of which the assignment is made of the cases to particular classes. When determining the value of the decision threshold u in the range [0, 1], and setting $f(x)$ such that:

$$f(x) = \begin{cases} 0 \ for \ x < u \\ 1 \ for \ x \geq u \end{cases} \tag{3}$$

a set of points can be obtained, which allows to plot the ROC curve.

In order to present the mechanism of the ROC curve plotting the following example will be shown. Table 2 contains example with 10 observations sorted in descending order of a classifier probability (so-called scoring model) with the actual classification of the observations (1 or 0). The next columns in the table include the settings of the actual and predicted classifications (TP, TP + FN, TN, TN + FP). SE column shows the sensitivity in accordance with formula (1), and the SP column - specificity determined by the formula (2).

Table 2. Mechanism of the ROC curve plotting

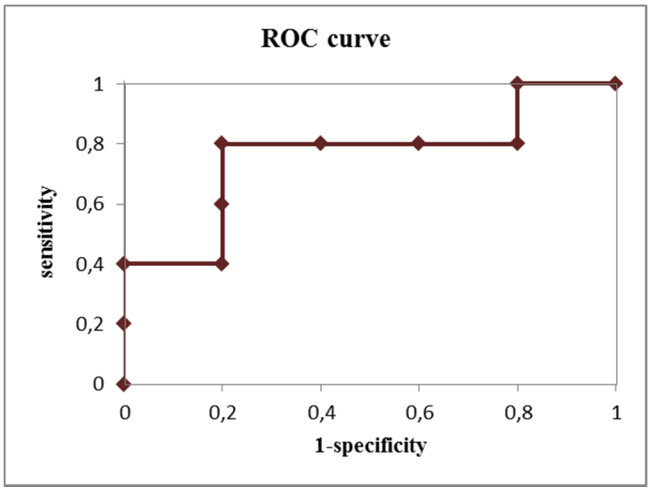
No. obs.	Classifier probability	True class	TP	TP+FN	SE	TN	TN+FP	SP	1-SP
1	0.90	1	1	5	0.2	5	5	1	0
2	0.85	1	2	5	0.4	5	5	1	0
3	0.75	0	2	5	0.4	4	5	0.8	0.2
4	0.70	1	3	5	0.6	4	5	0.8	0.2
5	0.55	1	4	5	0.8	4	5	0.8	0.2
6	0.45	0	4	5	0.8	3	5	0.6	0.4
7	0.40	0	4	5	0.8	2	5	0.4	0.6
8	0.35	0	4	5	0.8	1	5	0.2	0.8
9	0.25	1	5	5	1	1	5	0.2	0.8
10	0.10	0	5	5	1	0	5	0.0	1

Source: own preparation

The ROC curve for the data presented in Table 2 has the following form (Figure 1). The ROC curve was determined based on 10 observations only, therefore this curve has a discrete character. In case of a larger number of observations, the curve would be more smooth.

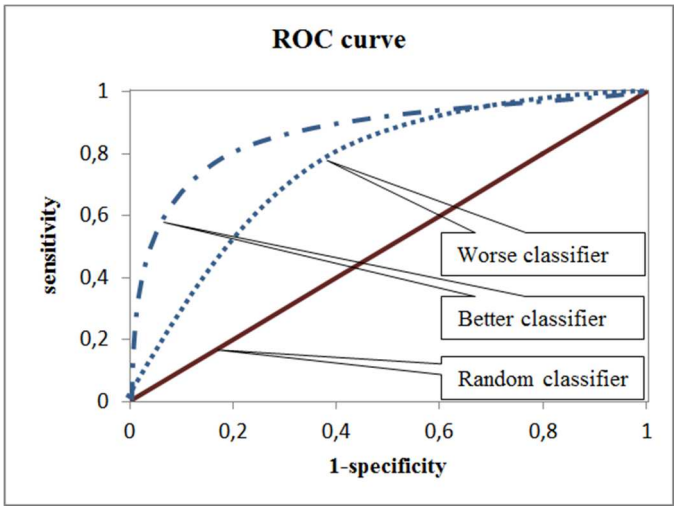
For the purpose of interpretation and comparison of multiple curves, two possible variants of the ROC curve are shown in Figure 2.

Figure 1. ROC curve for the data in Table 2



Source: own preparation

Figure 2. The ROC curve and its possible variants



Source: own preparation

Curve, which coincides with the diagonal curve, has no classification ability. The more the curve is convex and approaching the upper left corner, the better the discrimination has particular model. Highest (perfect) correctness puts classifier in (0,1).

Comparing ROC curves on the graph may be subject to error, especially when comparing a large number of models. Therefore, several ROC curve summary measures of the discriminatory accuracy of a test have been proposed in

the literature, such as the area under the curve (AUC) or the Youden index $\max_c = \{Se(u) + Sp(u) - 1\}$ [Youden 1950].

THE AUC ESTIMATION

One of the main feature associated with the ROC curve, is that curve is increasing and invariant under any monotonic increasing transformation of the considered variables. In general AUC is given by

$$AUC = \int_0^1 ROC(u)\, du \tag{4}$$

Moreover, let X_p and X_n denote the class marker for positive and negative cases, respectively. It could be shown that $AUC = P(X_p > X_n)$. This can be interpreted as the probability that in a randomly selected pair of positive and negative observations the classifier probability is higher for the positive case.

Since the ROC curve measures the inequality between the good and the bad score distributions, it seems reasonable to show a relation between the ROC curve and the Lorenz curve. Twice the area between the Lorenz curve and the diagonal line at 45 degree corresponds to the Gini concentration index. This leads to an interesting interpretation of the AUC measure in terms of the Gini coefficient: $Gini = 2AUC - 1$.

Parametric estimation

A simple parametric approach is to assume the X_p and X_n are independent normal variable with $X_p \sim N(\mu_p, \sigma_p^2)$ and $X_n \sim N(\mu_n, \sigma_n^2)$. Then the ROC curve can be summarized as follow:

$$ROC(u) = \Phi\left(a + b\Phi^{-1}(u)\right) \qquad u \in [0,1] \tag{5}$$

where $a = (\mu_p - \mu_n)/\sigma_p$, $b = \sigma_n/\sigma_p$ and Φ indicates the standard normal distribution function $X \sim N(0,1)$. Furthermore,

$$AUC = \Phi\left(\frac{\mu_p - \mu_n}{\sqrt{\sigma_p^2 + \sigma_n^2}}\right) \text{ or equivalently } AUC = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) \tag{6}$$

and can be estimated by substituting sample means and standard deviations into all above mentioned formulas.

In practical applications the assumption of normality is untenable, therefore transformation such as the log or the Box–Cox is often suggested [Zou and Hall

$$X^{\lambda} = \begin{cases} X^{\lambda} - 1 \big/ \lambda & for\ \lambda \neq 0 \\ \log(X) & for\ \lambda = 0 \end{cases} \tag{7}$$

2000], and the estimator (6) is then applied to the transformed data. Based on the observations on the positive and negative cases, an appropriate likelihood function can be constructed and maximized giving $\hat{\lambda}$, the maximum likelihood of estimate λ.

Non-parametric estimation

a)

The area under the empirical ROC curve is equal to the Mann–Whitney U statistic [Mann and Whitney 1947] which is usually computed to test whether the levels on some quantitative variable X in one population P tend to be greater than in second population N, without actually assuming how are they distributed in these two population. This measure provides an unbiased non-parametric estimator for the AUC [Faraggi and Reiser 2002]:

$$AUC = \frac{1}{N_p N_n} \sum_{i=1}^{N_p} \sum_{j=1}^{N_n} I\left(X_{pi}, X_{nj}\right) \text{ with } I = \begin{cases} 1 & for\ x_{pi} > x_{nj} \\ \frac{1}{2} & for\ x_{pi} = x_{nj} \\ 0 & otherwise \end{cases} \tag{8}$$

where N_p, N_n are the number of positive and negative cases respectively. Unfortunately, this estimator in some situation is not recommended, because it conceptually requires all $N_p N_n$ comparison and when we are dealing with large number of observations, computational time could be long. Sometimes in (8) sigmoid function is used instead of indicator function [Calders and Jaroszewicz 2007].

b)

When calculating the area under the curve it should be noted that the probabilistic classifiers give the values of the output vector other than the zero and one. Therefore, having m cases classification $o_1,...,o_m$ belonging to a set classes $C = \{C_1, C_2\}$ according to the decision threshold u, sorted so that $0 = se(C_1, o_1) \leq ... \leq se(C_1, o_m) = 1$ and $1 = sp(C_1, o_1) \geq ... \geq sp(C_1, o_m) = 0$ the area under the curve could be calculated via trapezoidal integration:

$$AUC = -\frac{1}{2}\sum_{i=2}^{m}(sp_i se_{i-1} - sp_{i-1} se_i) \tag{9}$$

where $se_i = se(C_1, o_i)$ represents the sensitivity of the classification i-th case to the class C_1, $sp_i = sp(C_1, o_i)$ is the specificity of the classification i-th case to the class C_1. The trapezoidal approach systematically underestimates the AUC, because of the way all of the points on the ROC curve are connected with straight lines rather than smooth concave curves.

To overcome the lack of smoothness of the empirical estimator, [Zou et al. 1997] used kernel methods to estimate the ROC curve, which were later improved by [Lloyd 1998]. Kernel density estimators are known to be simple, versatile, with good theoretical and practical properties.

TESTING DIFFERENCES BETWEEN TWO ROC CURVES

To compare classification algorithms by comparing the area under the ROC curves, is used the following procedure described by [Bradley 1997][Hanley and McNeil 1983]. We consider the following set of hypotheses

$$\begin{aligned} H_0 &: AUC_1 = AUC_2 \\ H_1 &: AUC_1 \neq AUC_2 \end{aligned} \tag{10}$$

to evaluate it, the following test statistic is used

$$z = \frac{A\hat{U}C_1 - A\hat{U}C_2}{\sqrt{SE^2\left(A\hat{U}C_1\right) + SE^2\left(A\hat{U}C_2\right)}} \tag{11}$$

which has the standardize normal distribution $N(0,1)$, and where

$$SE\left(A\hat{U}C\right) = \sqrt{\frac{\theta(1-\theta) + (n_1 - 1)\left(Q_1 - \theta^2\right) + (n_2 - 1)\left(Q_2 - \theta^2\right)}{n_1 n_2}} \tag{12}$$

$$Q_1 = \frac{\theta}{2-\theta},\ Q_2 = \frac{2\theta^2}{1+\theta} \tag{13}$$

where n_1 and n_2 are the number of negative and positive examples respectively and θ is the true area under the ROC curve (but in practice only the estimator $A\hat{U}C$ is used).

SIMULATION STUDY

In order to check the performance of the selected AUC estimator, we conducted the simulations based on the data used for telecom customer churn modelling (the loss of customers moving to some other company). The data is a collection of "Cell2Cell: The Churn Game" [Neslin 2002] derived from the Center of Customer Relationship Management at Duke University, located in North Carolina in the United States. They constitute a representative slice of the entire database, belonging to an anonymous company operating in the sector of mobile telephony in the United States.

The data contains 71047 observations, wherein each observation corresponds to the individual customer. For each observation 78 variables are assigned, of which 75 potential explanatory variables are used for models construction. All explanatory variables are derived from the same time period, except the binary dependent variable (the values 0 and 1) labeled as "churn", which has been observed in the period from 31 to 60 days later than the other variables. In the collection there is an additional variable "calibrat" to identify the learning sample and test sample, comprising 40000 and 31047 observations. Learning sample contains 20000 cases classified as churners (leavers) and 20000 cases classified as non-churners. In the test sample, which is used to check the quality of the constructed model, there is only 1.96% of people who quit. Such a small percentage of the class highlighted can be often found in the business practice.

In this study similar set of modelling techniques has been used as in [Gajowniczek and Ząbkowski 2012]. These were artificial neural networks, classification trees, boosting classification trees, logistic regression and discriminant analysis.

After estimation the $\widehat{\lambda}$ parameter by power transformation, we observed that most of the distributions (Table 3) have not the normal distribution based on Shapiro-Wilk normality test at $\alpha = 0.01$. As stated in [Krzyśko et al. 2008], X_p, X_n may not have a normal distribution, but the reasoning based on the ROC curve built for a normal distribution may give good results, because the ROC curves do not count individual distribution, but the relationship between the distributions.

Table 3. Tests for normality

	Negative cases (churn=1)		Positive cases (churn=0)	
	p-value	$\widehat{\lambda}$	p-value	$\widehat{\lambda}$
Artificial neural network (SANN)	2.88E-18	1.18	0.6694	1.37
Boosting classification trees (Boosting)	1.02E-22	1.16	0.2164	1.41
Logistic regression (Logit)	1.01E-08	0.77	0.0380	0.71
Classification trees (C&RT)	7.93E-51	1.13	1.20E-16	1.55
Discriminant analysis (GDA)	2.27E-08	0.79	0.0181	0.80

Source: own preparation

Very small differences can be seen in Table 4 among non-parametric AUC estimates. The biggest difference in AUC can be observed in case of classification trees. This is due to the fact that C&RT assigns observations to the leafs. Within each leaf there is the same probability of belonging to the positive class. Therefore, when there are only few leafs in the tree then we don't expect the distribution of probabilities to meet the assumption of normality.

Table 4. AUC estimation using different techniques

	Mann-Whitney (non-parametric)	Trapezoidal integration (non-parametric)	Normal assumption (parametric)
Artificial neural network (SANN)	0.6242784	0.6242784	0.6864752
Boosting classification trees (Boosting)	0.6632097	0.6632097	0.7045478
Logistic regression (Logit)	0.6189685	0.6189685	0.6072612
Classification trees (C&RT)	0.6215373	0.6227865	0.752052
Discriminant analysis (GDA)	0.6190384	0.6190384	0.6288627

Source: own preparation

Table 5 show the critical levels (p-values) for testing differences between two ROC curves based on Mann-Whitney estimation. The hypothesis of equality of the areas under the ROC curve could be reject when p-value are smaller than accepted level of significance. It can be observed that, at the significance level $\alpha = 0.05$, the areas under the curves for the SANN, Logit, C&RT, GDA are not significantly different. Only the AUC measures for Boosting significantly differs from the other methods.

Table 5. P-values for the differences between two AUC measure

	SANN	Boosting	Logit	C&RT	GDA
SANN	1.00000000	0.02417606	0.75930305	0.93139089	0.76237611
Boosting		1.00000000	0.01042189	0.01925184	0.01054391
Logit			1.00000000	0.82563864	0.99678138
C&RT				1.00000000	0.82878156
GDA					1.00000000

Source: own preparation

CONCLUSIONS

The aim of this study was to compare the accuracy of commonly used ROC curve estimation methods taking into account different classification techniques. We show that non-parametric methods give convergent results in terms of the AUC measure while parametric approach tends to give the higher values of AUC, except the Logit. In practical applications, for parametric methods of ROC estimation the assumption of normality is untenable, therefore, non-parametric methods should be utilized.

The simulation experiment suggest that the non-parametric ROC estimation using trapezoidal rule is a reliable method when the distributions of the predictive outcome are skewed and that it provides a smooth ROC. Finally, this approach of estimation is not difficult nor computationally time consuming.

Acknowledgments

The study is cofounded by the European Union from resources of the European Social Fund. Project PO KL „Information technologies: Research and their interdisciplinary applications", Agreement UDA-POKL.04.01.01-00-051/10-00.

REFERENCES

Bradley A.P. (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognition, vol. 30, No. 7, pp. 1145-1159.

Calders T., Jaroszewicz S. (2007) Efficient AUC Optimization for Classification, Proceedings of The 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'07), pp. 42-53.

Faraggi D.. Reiser B. (2002) Estimation of the area under the ROC curve, Statistics in Medicine, vol. 21, pp. 3093–3096.

Gajowniczek K., Ząbkowski T. (2012) Problemy modelowania rezygnacji klientów w telefonii komórkowej, Metody Ilościowe w Badaniach Ekonomicznych, vol. 13, No 3, pp. 65-79.

Hanley J. A., McNeil B. J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology vol. 143, pp. 29-36.

Hanley J. A., McNeil B. J. (1983) A method of comparing the areas under receiver operating characteristic curves derived from the same cases, Radiology vol. 148, pp. 839-843.

Krzyśko M., Wołyński W., Górecki T., Skorzybut M. (2008) Systemy uczące się, Wydawnictwo Naukowo-Techniczne.

Lloyd C. J. (1998) Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems, Journal of the American Statistical Association, vol. 93, pp. 1356–1364.

Mann H. B., Whitney D. R. (1947) On a test of whether one of two random variables is stochastically larger than the other, The Annals of Mathematical Statistics; vol. 18, pp. 50–60.

Neslin S. (2002) Cell2Cell: The churn game. Cell2Cell Case Notes, Hanover, NH: Tuck School of Business, Dartmoth College, Downloaded from: http://www.fuqua.duke.edu/centers/ccrm/datasets/cell/

Youden W. J. (1950) An index for rating diagnostic tests, Cancer, vol.3, pp. 32–35.

Zou K. H.; Hall W. J., Shapiro D. E. (1997). Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests, Statistics in Medicine, vol. 16, pp. 2143–2156.

Zou K. H., Hall W.J. (2000) Two transformation models for estimating an ROC curve derived from continuous data, Journal of Applied Statistics, vol. 27, pp. 621–631.