Performance Measures in Binary Classification

Matthias Kohl*

Department of Mechanical and Process Engineering, Furtwangen University, Jakob-Kienzle-Str. 17, D-78054 VS-Schwenningen, Germany

Abstract: We give a brief overview over common performance measures for binary classification. We cover sensitivity, specificity, positive and negative predictive value, positive and negative likelihood ratio as well as ROC curve and AUC.

Keywords: Sensitivity, specificity, positive predictive value, negative predictive value, positive likelihood ratio, negative likelihood ratio, prevalence, ROC curve, AUC, informative diagnostic test.

PERFORMANCE MEASURES

In many cases diagnostic tests are performed to distinguish between two groups reflecting presence or absence of a relevant medical condition. In this setup let us assume a group of N patients with true status y_1, ..., y_N where $y_i = 1$ represents presence and $y_i = 0$ absence of the medical condition. A diagnostic test T yields results t_1, ..., t_N where $t_i = 1$ represents a positive and $t_i = 0$ a negative test.

The simplest approach to measure the performance of test T is to use the *probability of misclassification* (PMC)

$$PMC = \frac{\text{cardinality of } \{i = 1,\ldots,N \,|\, y_i \neq t_i\}}{N}$$

respectively, the *accuracy* (ACC) = 1 - PMC

However, such a single performance measure may be misleading, as there are two possibilities for a correct respectively, wrong decision of the diagnostic test that are the correct respectively, wrong prediction of the presence or absence of the medical condition [1]. Thus, a pair of criteria should be used to obtain an exact description of the performance. In general, the results of a test can be summarized by the so called *confusion matrix*.

The confusion matrix whose structure is presented in Table **1** includes the information on the *prevalence* (Pr) for the considered group.

$$\Pr = \frac{TP + FN}{TP + FN + TN + FP}$$

Table 1: Confusion Matrix for Binary Classification

		Test result	
		0	**1**
True situation	0	True negative (TN)	False positive (FP)
	1	False negative (FN)	True positive (TP)

In addition, it is the basis for the definition of various performance measures. The percentage of correct positive tests for patients having the medical condition is called *sensitivity* (Se), whereas the percentage of correct negative tests for patients not having the medical condition is called *specificity* (Sp).

$$Se = \frac{TP}{TP + FN} \qquad SP = \frac{TN}{TN + FP}$$

The accuracy can also be expressed as a weighted sum of sensitivity and specificity

$$ACC = \Pr * Se + (1 - \Pr) * SP$$

The *positive predictive valu*e (PPV) is the probability that a patient with a positive test has the medical condition and the *negative predictive value* (NPV) is the probability that a patient with a negative test does not have the medical condition.

$$PPV = \frac{TP}{TP + FP} \qquad NPV = \frac{TN}{TN + FN}$$

The *positive likelihood ratio* (PLR) tells how likely patients with the medical condition are to have a positive test compared to patients without the medical condition. The *negative likelihood ratio* (NLR) tells how likely patients with the medical condition are to have a negative result compared to patients without the medical condition.

*Address corresponding to this author at the Department of Mechanical and Process Engineering, Furtwangen University, Jakob-Kienzle-Str. 17, D-78054 VS-Schwenningen, Germany; Tel: +49 (0) 7720 307-4746; Fax: +49 (0) 7720 307-4727; E-mail: Matthias.Kohl@hs-furtwangen.de

E-ISSN: 1929-6029/12

$$PLR = \frac{Se}{1-Sp} \qquad NLR = \frac{1-Se}{Sp}$$

During the development of a diagnostic test it is standard to use sensitivity and specificity for assessing the performance of the test. However, if there are two or more tests which have to be compared PLR and NLR should be chosen [2, 3].

It is important to note that both pairs of performance measures do not depend on the prevalence of the selected group which may be different from the intended-use population, whereas PPV and NPV depend on prevalence.

$$PPV = \frac{Pr * Se}{Pr * Se + (1-\mathrm{Pr}) * (1-Sp)}$$

$$NPV = \frac{(1-Pr) * Sp}{(1-\mathrm{Pr}) * Sp + Pr * (1-Se)}$$

The information provided by PPV and NPV is of great importance for physicians and patients [4]. In real-world applications where prevalence is often below 10% the diagnostic test must aim at substantially high values for sensitivity and specificity in order to be of utility otherwise PPV and NPV will be unacceptably low.

ROC CURVE

Let us assume a diagnostic test T giving not only 0 and 1 but a whole range of values where large values of T are more likely for patients having the medical condition and small values of T are more likely for patients not having the medical condition. Hence, for the final diagnostic test, which should only return 0 or 1, we have to select a threshold to distinguish between the two categories. In this setup sensitivity and specificity are the most frequently used performance measures and are displayed by so-called receiver operating characteristic (ROC) curves. For each threshold for the values of T one obtains a sensitivity and specificity value. Plotting these values leads to the ROC-curve of the diagnostic test T.

In Figure **1** we have selected a threshold leading to a sensitivity and a specificity of 0.75. Decreasing the threshold will increase the sensitivity and decrease the specificity, whereas increasing the threshold will decrease the sensitivity and increase the specificity. The line Se = 1 – Sp reflects a diagnostic test which is not *informative*, i.e. not better than chance. If there is a second test with a sensitivity and specificity lying in region A, then it has a higher PLR and a lower NLR which means that it is better. If the second test has a sensitivity and specificity lying in region B respectively, region C, then PLR is smaller and NLR is smaller respectively, PLR is larger and NLR is larger. Thus, it is not clear and it depends on the actual situation which classifier performs better. Finally, if the second classifier's sensitivity and specificity lie in region D, it has a lower PLR and a higher NLR, i.e. it is performing worse [2].

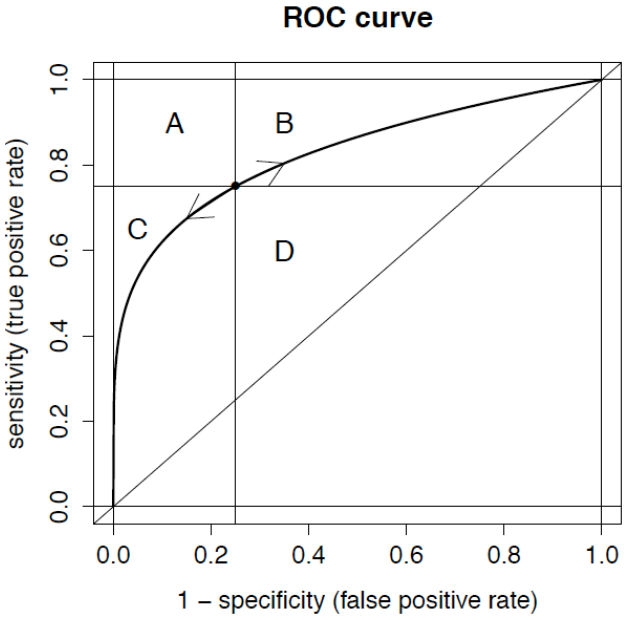

Figure 1: ROC curve for a diagnostic test.

ROC curves are often summarized by the area under the ROC curve (AUC) where an AUC of 0.5 means that the diagnostic test is not better than chance in predicting the categories, whereas values larger than 0.5 indicate a result better than chance. If the AUC is smaller than 0.5 the labels of the categories are misplaced and should be switched leading to an AUC greater than 0.5. Having an AUC larger than 0.5 the diagnostic test is informative; equivalent characterizations in terms of the introduced performance measures are: Se + Sp > 1, PPV + NPV > 1 , PPV > Pr, NPV < 1 – Pr, PLR > 1, or NLR < 1.

ACKNOWLEDGEMENT

We would like to thank two anonymous referees for valuable comments on the manuscript.

APPENDIX OF SYMBOLS

PMC = probability of misclassification

ACC = accuracy

TN = true negative

FN = false negative

TP = true positive

TN = true negative

Pr = prevalence

Se = sensitivity

Sp = specificity

PPV = positive predictive value

NPV = negative predictive value

PLR = positive likelihood ratio

NLR = negative likelihood ratio

ROC = receiver operating characteristic

AUC = area under the ROC curve

REFERENCES

[1] Sokolova M and Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag 2009; 45: 427-37. http://dx.doi.org/10.1016/j.ipm.2009.03.002

[2] Biggerstaff BJ. Comparing diagnostic tests: a simple graphic using likelihood ratios. Statist Med 2000; 19: 649-63. http://dx.doi.org/10.1002/(SICI)1097-0258(20000315)19:5<649::AID-SIM371>3.0.CO;2-H

[3] Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. BMJ 2004; 329(7458): 168-69. http://dx.doi.org/10.1136/bmj.329.7458.168

[4] Altman DG, Bland JM. Diagnostic tests 2: predictive values. BMJ 1994; 309(6947): 102. http://dx.doi.org/10.1136/bmj.309.6947.102

Received on 15-09-2012 Accepted on 01-10-2012 Published on 08-10-2012

http://dx.doi.org/10.6000/1929-6029.2012.01.01.08