Immunological Bioinformatics

Ole Lund
Morten Nielsen
Claus Lundegaard
Can Keşmir
Søren Brunak

The MIT Press
Cambridge, Massachusetts
London, England

	Predicted positive	Predicted negative	Total
Actual positive	TP	FN	AP
Actual negative	FP	TN	AN
Total	PP	PN	N

Table 4.2: Classification of predictions. TP: true positives (predicted positive, actual positive); TN: true negatives (predicted negative, actual negative); FP: false positives (predicted positive, actual negative); FN: false negatives (predicted negative, actual positive).

4.10 Performance Measures for Prediction Methods

A number of different measures are commonly used to evaluate the performance of predictive algorithms. These measures differ according to whether the performance of a real-valued predictor (e.g., binding affinities) or a classification is to be evaluated.

In almost all cases percentages of correctly predicted examples are not the best indicators of the predictive performance in classification tasks, because the number of positives often is much smaller than the number of negatives in independent test sets. Algorithms that underpredict a lot will therefore appear to have a high success rate, but will not be very useful.

We define a set of performance measures from a set of data with N predicted values p_i and N actual (or target) values a_i. The value p_i is found using a prediction method of choice, and the a_i is the known corresponding target value. By introducing a threshold t_a, the N points can be divided into actual positives A_P (points with actual values a_i greater than t_a) and actual negatives A_N. Similarly, by introducing a threshold for the predicted values t_p, the points can be divided into predicted positives P_P and predicted negatives P_N. These definitions are summarized in table 4.2 and will in the following be used to define a series of different performance measures.

4.10.1 Linear Correlation Coefficient

The linear correlation coefficient, which is also called Pearson's r, or just the correlation coefficient, is the most widely used measure of the association between pairs of values [Press et al., 1992]. It is calculated as

$$c = \frac{\sum_i (a_i - \overline{a})(p_i - \overline{p})}{\sqrt{\sum_i (a_i - \overline{a})^2}\sqrt{\sum_i (p_i - \overline{p})^2}} , \tag{4.48}$$

where the overlined letters denote average values. This is one of the best measures of association, but as the name indicates it works best if the actual and predicted values when plotted against each other fall roughly on a line. A value of 1 corresponds to a perfect correlation and a value of −1 to a perfect anticorrelation (when the prediction is high, the actual value is low). A value of 0 corresponds to a random prediction.

4.10.2 Matthews Correlation Coefficient

I f all the predicted and actual values only take one of two values (normally 0 and 1) the linear correlation coefficient reduces to the Matthews correlation coefficient [Matthews, 1975]

$$c = \frac{T_P T_N - F_P F_N}{\sqrt{(T_P + F_N)(T_N + F_P)(T_P + F_P)(T_N + F_N)}} = \frac{T_P T_N - F_P F_N}{\sqrt{A_P A_N P_P P_N}} . \qquad (4.49)$$

As for the Pearson correlation, a value of 1 corresponds to a perfect correlation.

4.10.3 Sensitivity, Specificity

Four commonly used measures are calculated by dividing the true positives and negatives by the actual and predicted positives and negatives [Guggenmoos-Holzmann and van Houwelingen, 2000],

Sensitivity Sensitivity measures the fraction of the actual positives which are correctly predicted: $sens = \frac{TP}{AP}$.

Specificity Specificity denotes the fraction of the actual negatives which are correctly predicted: $spec = \frac{TN}{AN}$

PPV The positive predictive value (PPV) is the fraction of the predicted positives which are correct: $PPV = \frac{TP}{PP}$.

NPV The negative predictive value (NPV) stands for the fraction of the negative predictions which are correct: $NPV = \frac{TN}{PN}$.

4.10.4 Receiver Operator Characteristics Curves

One problem with the above measures (except Pearson's r) is that a threshold t_p must be chosen to distinguish between predicted positives and negatives. When comparing two different prediction methods, one may have a better Matthews correlation coefficient than the other. Alternatively, one may have a higher sensitivity or a higher specificity. Such differences may be due to the choice of thresholds and in that case the two prediction methods may

Rank	Prediction	Actual	TPP	FPP	Area
1	0.1	1	0.33	0	0
2	0.3	0	0.33	0.5	0.17
3	0.35	1	0.66	0.5	0.17
4	0.7	1	1.00	0.5	0.17
5	0.88	0	1.00	1	**0.67**

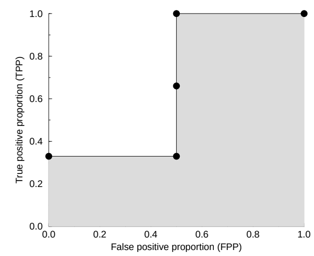

Figure 4.15: Calculation of a ROC curve. The table on the left side of the figure indicates the steps involved in constructing the ROC curve. The pairs of predicted and actual values must first be sorted according to the predicted value. The value in the lower right corner is the A_{ROC} value. In the right panel of the figure is shown the corresponding ROC curve.

be rendered identical if the threshold for one of the methods is adjusted. To avoid such artifacts a nonparametric performance measure such as a receiver operator characteristics (ROC) curve is generally applied.

The ROC curve is constructed by using different values of the threshold t_p to plot the false-positive proportion $FPP = F_P/A_N = F_P/(F_P + T_N)$ on the x-axis against the true positive proportion $TPP = T_P/A_P = T_P/(T_P + F_N)$ on the y-axis [Swets, 1988]. Figure 4.15 shows an example of how to calculate a ROC curve and the area under the curve, A_{ROC}, which is a measure of predictive performance. An A_{ROC} value close to 1 indicates again a very good correlation; a value close to 0 indicates a negative correlation and a value of 0.5, no correlation. A general rule of thumb is that an A_{ROC} value > 0.7 indicates a useful prediction performance, and a value > 0.85 a good prediction. A_{ROC} is indeed a robust measure of predictive performance. Compared with the Matthews correlation coefficient, it has the advantage that it is independent of the choice of t_p. It is still, however, dependent on the choice of a threshold t_a for the actual values. Compared with Pearson's correlation r it has the advantage that it is nonparametric, i.e., that the actual value of the predictions is not used in the calculations, only their ranks. This is an advantage in situations where the predicted and actual values are related by a nonlinear function.