Notes on Performance Metrics for Pattern Detection and Recognition

Let us define:
P: condition positive (P): the number of real positive cases in the data
N: condition negatives (N): the number of real negative cases in the data

TP: true positive, also called a hit
TN: true negative, also called a correct rejection
FP: false positive: also called a false alarm or a Type I error
FN: false negative (FN), also called a miss or a Type II error

Detection functions can be viewed as a form of Information Retrieval operation. Detection functions retrieve positive examples from a large body of possible cases. The classic performance indicators for information retrieval are “Precision” and “Recall”.

Precision and Recall

Precision, also called Positive Predictive Value(PPV) is the fraction of retrieved instances that are relevant to the problem.

$$PP = \frac{TP}{TP + FP}$$

A perfect precision score (PPV=1.0) means that every result retrieved by a search was relevant, but says nothing about whether all relevant documents were retrieved.

Recall, also known as sensitivity (S), hit rate, and True Positive Rate (TPR) is the fraction of relevant instances that are retrieved.

$$S = TPR = \frac{TP}{T} = \frac{TP}{TP + FN}$$

A perfect recall score (TPR=1.0) means that all relevant documents were retrieved by the search, but says nothing about how many irrelevant documents were also retrieved.

Both precision and recall are therefore based on an understanding and measure of relevance. In our case, “relevance” corresponds to “True”.
Precision answers the question “How many of the Positive Elements are True ?”
Recall answers the question “How many of the True elements are Positive”?

In many domains, there is an inverse relationship between precision and recall. It is possible to increase one at the cost of reducing the other.

F-Measure

The F-measures combine precision and recall into a single value. The F measures measure the effectiveness of retrieval with respect to a user who attaches 2 times as much importance to recall as precision.

The F1 score weights recall higher than precision.

F_1 Score:

$$F_1 = \frac{2TP}{2TP + FP + FN}$$

The F1 score is the harmonic mean of precision and sensitivity.
This is the geometric mean divided by the arithmetic mean.

Accuracy:

Accuracy is the fraction of test cases that are correctly classified (T).

$$ACC = \frac{T}{M} = \frac{TP + TN}{M}$$

where M is the quantity of test data.

Note that the terms Accuracy and Precision have a very different meaning in Measurement theory. In measurement theory, accuracy is the average distance from a true value, while precision is a measure of the reproducibility for the measurement.

Matthews Correlation Coefficient

The Matthews correlation coefficient is a measure of the quality of binary (two-class) classifications. This measure was proposed by the biochemist Brian W. Matthews in 1975.

MCC takes into account true and false positives and negatives and is generally regarded as a balanced measure that can be used even if the classes are of very different sizes.

The MCC is in essence a correlation coefficient between the observed and predicted binary classifications

MCC results a value between +1 and -1, where +1 represents a perfect prediction, 0 no better than random prediction and −1 indicates total disagreement between prediction and observation.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The original formula given by matthews was:

M = Total quanitity of test data:

$$M = TN + TP + FN + FP$$

$$S = \frac{TP + FN}{M}$$

$$P = \frac{TP + FP}{M}$$

$$MCC = \frac{TP/M - S \cdot P}{\sqrt{PS(1-S)(1-P)}}$$