Journal of Engineering Research

Volume 6 | Issue 5 Article 1

2022

Confusion Matrix in Binary Classification Problems: A Step-by-Step Tutorial

Mahmoud Fahmy Amin

Vol. 6, No. 5 – 2022 Journal of Engineering Research (ERJ)

Confusion Matrix in Binary Classification Problems: A Step-by-Step Tutorial

Mahmoud M. Fahmy

Professor, Computer and Control Engineering Department, Faculty of Engineering, Tanta University, Tanta, Egypt
e-mail: m.fahmy@f-eng.tanta.edu.eg

***Abstract*: In the field of machine learning, the confusion matrix is a specific table adopted to describe and assess the performance of a classification model (e.g. an artificial neural network) for a set of test data whose actual distinguishing features are known. The learning algorithm is thus of the supervised learning category. For an n-class classification problem, the confusion matrix is square with n rows and n columns. The rows represent the class actual samples (instances) which are the inputs to the classifier, while the columns represent the class predicted samples, the classifier outputs. (The converse is also valid, i.e. the two dimensions 'actual' and 'predicted' can be assigned to columns and rows, respectively). Binary as well as multiple-class classifiers can be dealt with. It is worth noting that the term 'matrix' here has nothing to do with the theorems of matrix algebra; it is regarded just as an information-conveying table. The descriptive word 'confusion' stems from the fact that the matrix clarifies to what extent the model confuses the classes — mislabels one as another. The essential concept was introduced in 1904 by the British statistician Karl Pearson (1857 — 1936).**

***Keywords*— Machine Learning; Confusion matrix; Accuracy; Recall; Specificity; Precision; True Negative; False Positive, Balanced Accuracy.**

1. BINARY CLASSIFICATION

We begin with the basic and relatively simple situation of a binary classifier, where we have two classes (n= 2) and a 2x2 confusion matrix. See Fig. 1. Let the matrix in this figure, as an illustrative example, belong to a medical test conducted on a number of persons (patients) for the presence or absence of a certain disease. The labels 'positive' (+ve) and 'negative' (-ve) are used to identify these two distinct cases, respectively, which are treated as two classes in a classification problem. (Other labels such as '1' and '0', 'yes' and 'no', or 'event' and 'not event' can likewise be used). With such labeling, attention is sometimes focused on the positive class, and its classification outcomes are considered the decisive characteristics of the classifier.

		Predicted +ve	Predicted -ve
Actual	+ve	100	5
	-ve	10	90

Fig. 1. Confusion matrix for binary classifier

The confusion matrix of Fig. 1 tells us that:

- The label 'positive' means the person has the disease, and the label 'negative' means the person does not.
- A total of 205 (= 100 +5+ 10+ 90) persons were tested.
- Out of the 205 persons, the classifier predicted as 'positive' 110 (= 100+ 10) times and as 'negative' 95 (=5 + 90) times (regardless whether the predictions are correct or not).
- In actuality, 105 (= 100 +5) persons in the test set have the disease and 100 (= 10 + 90) persons do not.

More conclusive information can be drawn from the confusion matrix, as elucidated below.

A. Building blocks: TP, TN, FP, and FN

Formally, a comparison of the actual classifications with the predicted classifications reveals that four well-defined outcomes emerge:

- The actual classification is positive and the predicted classification is positive. This outcome is referred to as 'true positive', abbreviated TP, because the positive sample is correctly identified by the classifier.
- The actual classification is negative and the predicted classification is negative. This is a "true negative" (TN) outcome because the negative sample is correctly identified by the classifier.
- The actual classification is negative and the predicted classification is positive. This is a 'false positive' (FP) outcome because the negative sample is incorrectly identified by the classifier as positive.
- The actual classification is positive and the predicted classification is negative. This is a 'false negative' (FN) outcome because the positive sample is incorrectly identified by the classifier as negative.

These four outcomes, with the above interpretation, pertain in fact to the positive class, provided this class is particularly important and deserves emphasis; it accommodates what can be called 'relevant' samples, while the negative class is regarded as 'irrelevant'.

The outcomes TP, TN, FP, and FN are of prime significance and are termed the 'building blocks', since they are employed to formulate all performance measures as will be evident in Section 3.

The building blocks appear naturally as the elements of the confusion matrix, as shown in Figs. 2 and 3. Note that the true outcomes TP and TN occupy the two diagonal cells of the matrix. The false outcomes FP and FN occupying the two off-diagonal cells imply errors; FP is a type I error and FN is a type II error. In our example of ill and healthy persons, FP represents persons who are healthy and classified as ill while, on the contrary, FN represents persons who are ill and classified as healthy. The latter case (type II error) is normally more dangerous than the former (type I error).

Returning to Fig. 1, the building blocks are seen to be TP = 100, TN= 90, FP =10, FN =5. Ideally, FP and FN would both be of zero values, representing a perfect classifier.

		Predicted +ve	Predicted -ve
Actual	+ve	TP	FN
	-ve	FP	TN

Fig. 2. Building blocks of positive class as elements of 2X2 confusion matrix

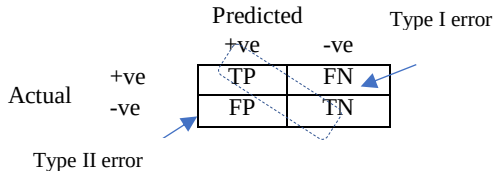


Fig. 3. True outcomes in diagonal cells and false outco mes in off-diagonal cells

But, in practice, there is a challenge of how to minimize FP and FN (i.e. maximize TP and TN). Bear in mind that the building blocks are all whole positive numbers (counts); they cannot be fractions or percentages.

It is to be noted that the positive and negative classes can be interchanged, so that the confusion matrix appears as in Fig. 4. In comparison with Fig. 2, we find that TP and TN are merely interchanged and so are FP and FN.

		Predicted -ve	Predicted +ve
Actual	-ve	TN	FP
	+ve	FN	TP

Fig. 4. Interchanging positive and negative classes

Furthermore, we can write:

Number of positive samples in the test set,

$$N_+ = TP + FN \quad (1)$$

Number of negative samples in the test set,

$$N_- = FP + TN \quad (2)$$

Total number of tested samples,

$$N = TP + FN + FP + TN = N_+ + N_- \quad (3)$$

Number of samples predicted as positive,

$$P_+ = TP + FP \quad (4)$$

Number of samples predicted as negative,

$$P = FN + TN = N - P_+ \quad (5)$$

Example 1

Consider a set of 12 persons, numbered as 1 through 12. Persons 1 through 8 suffer from the covid disease and belong to class 1, while persons 9 through 12 are covid-free and belong to class 0. A binary classifier for this set made 9 correct predictions and 3 incorrect ones. Persons 1 and 2 were predicted as covid-free and person 9 was predicted as having covid.

(a) Determine the building blocks for class 1.

(b) Construct the confusion matrix of the classifier.

Solution

The classification situation is illustrated in Fig. 5. Labels '1' and '0' for the two classes corresponding to 'positive' and 'negative', respectively. From Fig. 5, the building blocks for class 1 are

TP = 6, TN = 3, FP = 1, FN = 2

The confusion matrix of the classifier is shown in Fig. 6.

		Predicted 1	Predicted 0
Actual	1	6	2
	0	1	3

Fig. 6. Confusion matrix for Example 1

Example 2

A set of 1000 pens contains 650 pens of the Parker brand and the remaining pens are of other brands. A binary classifier correctly identified the 650 Parker pens anal incorrectly identified 57 non-Parker pens as Parker.

(a) How many non-Parker pens were correctly identified?

(b) Construct the confusion matrix of the classifier.

Solution

There are two classes: Parker class (positive) and non-Parker class (negative). We also have

$$N = 1000 \quad , \quad N_+ = 650$$
$$TP = 650 \quad , \quad FP = 57$$

From Eq. (3),

$N_- = N - N_+ = 1000 - 650 = 350$

From Eq. (1),

$FN = N_+ — TP = 650 - 650 = 0$

From Eq. (2),

$TN = N_- - FP = 350 - 57 = 293$

That is, the number of non-Parker pens correctly identified is 293. The confusion matrix, based on the building blocks for the Parker class, is shown in Fig. 7.

		Predicted Parker	Predicted Non-Parker
Actual	Parker	650	0
	Non-Parker	57	293

Fig. 7. Confusion matrix for Example 2

B. Building blocks for individual classes

When the two classes handled by a binary classifier are nearly of the same importance, the two sets of their building blocks, with foreseen interrelations, are to be equally studied. We identify the individual classes with arbitrary labels, and no preference is given to one class over the other. Figure 8 shows a confusion matrix for two classes labeled A and B, where it is seen that:

Person's number	1	2	3	4	5	6	7	8	9	10	11	12
Actual classification	1	1	1	1	1	1	1	1	0	0	0	0
Predicted classification	0	0	1	1	1	1	1	1	1	0	0	0
Outcome	FN	FN	TP	TP	TP	TP	TP	TP	FP	TN	TN	TN

Fig. 5. Outcomes for Example 1

		Predicted A	Predicted B	
Actual	A	215	25	N_A=240
	B	40	190	N_B=230

Fig. 8. A confusion matrix of binary classifier with classes A and B

Number of tested class-A samples,

$$N_A = 215 + 25 = 240$$

Number of tested class-B samples,

$$N_B = 40 + 190 = 230$$

Total number of tested samples,

$$N = N_A + N_B = 240 + 230 = 470$$

Here, the descriptors 'positive' and 'negative' do not appear, but their intended meanings are implicit. If we consider class A, we understand that:

- Class-A samples correctly classified are TP_A, true positives for class A; $TP_A = 215$.
- Class-B samples correctly classified are TN_A, true negatives for class A; $TN_A = 190$.
- Class-B samples incorrectly classified as class A are FP_A, false positives for class A; $FP_A = 40$.
- Class-A samples incorrectly classified as class B are FN_A, false negatives for class A; $FN_A = 25$.

Considering class B, on the other hand, we understand that:

- Class-B samples correctly classified are TP_B, true positives for class B; $TP_B = 190$.
- Class-A samples correctly classified are TN_B, true negatives for class B; $TN_B = 215$.
- Class-A samples incorrectly classified as class B are FP_B, false positives for class B; $FP_B = 25$.
- Class-B samples incorrectly classified as class A are FN_B, false negatives for class B; $FN_B = 40$.

For easy reference and remembrance, the building blocks for classes A and B are represented symbolically in Fig. 9. The directed symbol A → A, for example, means when the input to the classifier is A, the classifier output is A.

A little thought ensures that:

$$\left.\begin{aligned} TP_A &= TN_B \\ TN_A &= TP_B \\ FP_A &= FN_B \\ FN_A &= FP_B \end{aligned}\right\} \quad (6)$$

which are intrinsic relationships between the building blocks of class A and those of class B. Note that the symbols P and N are just interchanged when transferring from class A to class B and vice versa. This implies an interesting result that once the building blocks of one class are determined, the building blocks of the other class are readily known with no additional calculations. In Fig. 8, we already have

$$TP_A = TN_B = 215 \;,\; TN_A = TP_B = 190$$
$$FP_A = FN_B = 40 \;,\; FN_A = FP_B = 25$$

It is also obvious from relationships (6) that, for classes A and B, the sum of true positives is equal to the sum of true negatives,

$$TP_A + TP_B = TN_A + TN_B \quad (7)$$

and the sum of false positives is equal to the sum of false negatives ,

$$FP_A + FP_B = FN_A + FN_B \quad (8)$$

From another perspective, under conditions (6), the confusion matrix of a binary classifier with classes A and B can take either of the two forms shown in Fig. 10. In Fig.10a, the first row (column) is assigned to class A and, in Fig. 10b, the first row (column) is assigned to class B. The two forms are of course equivalent; they convey the same pieces of information.

The confusion matrix in Fig. 8 can thus take an alternative (equivalent) form of Fig. 11, by interchanging classes A and B. From either form, we immediately realize that:

- 215 class_A samples are correctly classified.
- 190 class_B samples are correctly classified.
- 40 class-B samples are incorrectly classified as class A.
- 25 class_A samples are incorrectly classified as class B.

TP_A : A → A	TP_B : B → B
TN_A : B → B	TN_B : A → A
FP_A : B → A	FP_B : A → B
FN_A : A → B	FN_B : B → A

Class A Class B

Fig. 9. Symbolic representation of building blocks for two classes

		Predicted B	Predicted A	
Actual	B	190	40	N_B=230
	A	25	215	N_A=240

Fig. 11. Another form for confusion matrix of Fig. 8

		Predicted A	Predicted B
Actual	A	$TP_A = TN_B$	$FN_A = FP_B$
	B	$FP_A = FN_B$	$TN_A = TP_B$

(a)

		Predicted B	Predicted A
Actual	B	$TP_B = TN_A$	$FN_B = FP_A$
	A	$FP_B = FN_A$	$TN_B = TP_A$

(b)

Fig. 10. Two forms for confusion matrix of binary classifier through interchange of classes

Example 3

A binary classifier has the confusion matrix of Fig. 12 for classes K and L.

(a) For class K, how many samples are correctly classified and how many are incorrectly classified?
(b) Repeat part (a) for class L.
(c)
(d) How many class-K samples are tested?
(e) How many class-L samples are tested?
(f) Determine the building blocks for classes K and L.
(g) Construct another equivalent form for the classifier confusion matrix.

		Predicted K	Predicted L
Actual	K	510	70
	L	100	660

Fig. 12. Confusion matrix for Example 3

Solution

Number of class-K samples correctly classified,
$TP_K = 510$ $(= TN_L)$
Number of class-K samples incorrectly classified,
$FN_K = 70$ $(= FP_L)$
Number of class-L samples correctly classified,
$TP_L = 660$ $(= TN_K)$
Number of class-L samples incorrectly classified,
$FN_L = 100$ $(= FP_K)$
Number of tested class-K samples,
$N_K = 510 + 70 = 580$
Number of tested class-L samples,
$N_L = 100 + 660 = 760$

The building blocks for classes K and L are given in Fig. 13. Another equivalent form for the confusion matrix is shown in Fig. 14, obtained. From Fig. 12 by interchanging classes K and L.

	TP	TN	FP	FN
Class K	510	660	100	70
Class L	660	510	70	100

Fig. 13. Building blocks for classes K and L in Example 3

		Predicted L	Predicted K
Actual	L	660	100
	K	70	510

Fig. 14. Another form for confusion matrix in Example 3

		Predicted A	Predicted B	
Actual	A	990	10	N_A=1000
	B	48	2	N_B=50

Fig. 15. Confusion matrix for two imbalanced datasets

2. PERFORMANCE MEASURES FOR BINARY CLASSIFICATION

Based on the confusion matrix, we define a group of different performance measures (metrics) for the evaluation of binary classification models. The most-widely used measures are discussed in Subsections 3.1 through 3.6. Generally, as the value of the measure gets larger, the classifier becomes better.

A. Accuracy

The accuracy of a binary classification model is the ratio of the number of correctly classified samples (true outcomes) to the total number of tested samples. Referring to Fig. 2, the model accuracy is

$$\text{Accuracy} = \frac{TP+TN}{N} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (9)$$

In Fig.1, for example, since TP=100, TN=30, and N = 20S, then

$$\text{Accuracy} = \frac{100+90}{205} = 0.927 \qquad (92.7\%)$$

This indicates that 92.7% of the tested samples are correctly classified or, equivalently, the classification error is 7.3%.
In terms of two classes A and B, the model accuracy takes the forms

$$\text{Accuracy} = \frac{TP_A+TN_A}{N} = \frac{TP_B+TN_B}{N} \qquad (10a)$$

which can also be written as

$$\text{Accuracy} = \frac{TP_A+TP_B}{N} = \frac{TN_A+TN_B}{N} \qquad (10b)$$

In view of relationships (6) and Fig. (10), the four (apparently different) forms of Eqs.(10) are the same in value. It is interesting to think in a like manner of the accuracy of the individual classes. For class A, $\text{Accuracy}_A = \frac{TP_A+TN_A}{N}$ and for class B, $\text{Accuracy}_B = \frac{TP_B+TN_B}{N}$. This implies that the model accuracy is the same as the accuracy of either of the two classes.

In Fig. 8, $TP_A = TN_B = 215$, $TN_A = TP_B = 190$, and $N = 470$. Therefore,

$$\text{Accuracy} = \text{Accuracy}_A = \text{Accuracy}_B = (215 + 190)/470 = 0.862$$

In spite of the formality of the accuracy measure, it is unfortunately reliable only if the two classes have balanced datasets. Two datasets are said to be balanced when they have nearly the same number of samples. Otherwise, the datasets are imbalanced and the accuracy measure can be misleading. To demonstrate, suppose class A has $N_A = 1000$ samples and class B has $N_B = 50$ samples (only 5% of class A). Here the classification model will be 'biased' to class A which has the majority of samples. The confusion matrix in this case may have the form of Fig. 15.

The model accuracy, by Eq. (10), is calculated as

$$\text{Accuracy} = (990 + 2)/1050 = 0.945 \ (94.5\ \%)$$

which can be judged as an acceptably high level of accuracy; 992 samples are correctly classified out of 1050 samples. However, when we examine the outcomes of the individual classes, we find out that while 990 class-A samples are

correctly classified out of 1000 samples with a percentage as high as 99% (taken as TP_A/N_A), only two class-B samples are correctly classified out of 50 samples with a very low percentage of 4% (TP_B/N_B). These results warn us that the 94.5 % accuracy cannot be relied upon; it deceptively describes the classification reliability of individual classes with dataset imbalance. One other measure, called balanced accuracy, will be specified in Subsection 3.5 for imbalanced datasets.

Example 4

A binary classification model is used for two classes A and B. The number of samples classified as class A is 169 and the number of samples classified as class B is 157. The type I and type II errors are recorded to be 39 samples and 46 samples, respectively.

(a) What is the percentage of correctly classified samples in class A? class B?

(b) Determine the accuracy of the model.

Solution

The data given is represented in the confusion matrix of Fig. 16, and we have

		Predicted A	Predicted B	
Actual	A	TP_A	$FN_A = 46$	N_A
	B	$FP_A = 39$	TN_A	N_B
		$P_A = 169$	$P_B = 157$	

Fig. 16. Confusion matrix for Example 4

$N = P_A + P_B = 169 + 157 = 326$

$TP_A = P_A - FP_A = 169 - 39 = 130$

$TN_A = P_A - FN_A = 157 - 46 = 111 \quad (= TP_B)$

$N_A = TP_A + FN_A = 130 + 46 = 176$

$N_B = FP_A + TN_A = 39 + 111 = 150$

Percentage of correctly classified samples in class A,

$$\frac{TP_A}{N_A} * 100 = \frac{130*100}{176} = 73.9\% \qquad (11)$$

Percentage of correctly classified samples in class B,

$$\frac{TP_B}{N_B} * 100 = \frac{111*100}{150} = 74\% \qquad (12)$$

Model accuracy, by Eq. (10),

$$\frac{TP_A+TN_A}{N} = \frac{130+111}{326} = 0.739\% \quad (73.9\%) \quad (13)$$

The difference in the values of (11), (12), and (13) is really slight. The reason is that the datasets of the two classes are balanced.

B. Precision

The precision is the ratio of the number of samples correctly classified as positive to the number of all samples classified as positive. Considering the first column of Fig. 2, we have

$$\text{Precision} = \frac{TP}{P_+} = \frac{TP}{TP + FP} \qquad (14)$$

For example, in Fig. 1, where TP=100 and FP = 10,

$$\text{Precision} = \frac{100}{100+10} = 0.909 \qquad (90.9\ \%)$$

In an ideal case when FP = 0, the precision reaches its maximum value of 1.0. This means that all samples predicted as positive actually belong to the positive class (TP = P_+); the type I error is of zero value. See Fig. 17. Strictly speaking, expression (14) is the precision of the positive class.

		Predicted +ve	Predicted -ve
Actual	+ve	$TP = P_+$	FN
	-ve	$FP = 0$	TN
		P_+	P_-

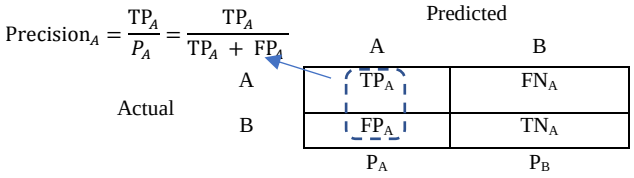
Fig. 17. Maximum precision

For two classes A and B, we write

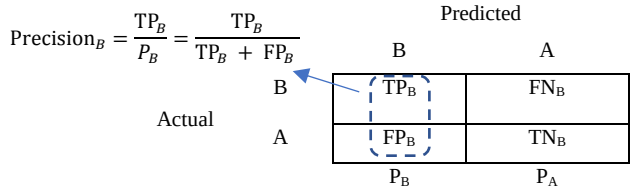
$$\text{Precision}_A = \frac{TP_A}{P_A} = \frac{TP_A}{TP_A + FP_A} \qquad (15)$$

$$\text{Precision}_B = \frac{TP_B}{P_B} = \frac{TP_B}{TP_B + FP_B} \qquad (16a)$$

See the two forms of confusion matrix in Fig. 18. In expression (15), two class-A building blocks TP_A and FP_A (first column in Fig. 18a) are used and, similarly, two class-B building blocks TP_B and FP_B (first column in Fig. 18b) are used in expression (16a). In words, the precision of a certain class is the ratio of the number of samples of the class correctly classified as belonging to this class (true positives) to the number of all samples classified, correctly or incorrectly, as belonging to the same class (true positives plus false positives).



(a) Class A



(b) Class B

Fig. 18. Precisions of classes A and B as obtained from two forms of confusion matrix

By virtue of relationships (6), Precision_B in (16a) can also be expressed in terms of two class-A building blocks TN_A and FN_A (second column in Fig. 18a) as

$$\text{Precision}_B = \frac{TN_A}{TN_A + FN_A} \qquad (16b)$$

That is, one form of confusion matrix, as that in Fig. 18a, can give us both Precision_A and Precision_B by considering the two columns of the matrix, respectively, as illustrated in Fig. 19. Similar arguments apply to the other form in Fig. 18b.

In Fig. 8, $TP_A = 215$, $FP_A = 40$, $FN_A = 25$, and $TN_A = 190$.

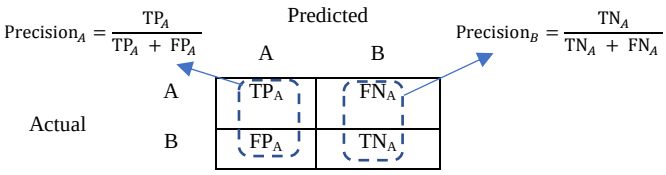


Fig. 19. Precisions of classes A and B as obtained from one form of confusion matrix

Therefore,

$$\text{Precision}_A = 215/(215 + 40) = 0.843$$

$$\text{Precision}_B = 190/(190 + 25) = 0.884$$

A crucial question is: What is the precision of the binary classification model as a whole? This is determined through some sort of averaging of the precisions of the two individual classes. There are three methods to define an average precision; namely,

- Macro-average
- Micro-average
- Weighted-average

The values calculated from these methods generally differ from each other, especially for imbalanced datasets, depending on the individual class precisions.

The macro-average precision of a model with classes A and B is

$$\text{Precision}_{macro} = \frac{\text{Precision}_A + \text{Precision}_B}{2} \qquad (17)$$

i.e. the arithmetic average (mean) of the two precisions, with equal weights of unity.

The micro-average precision is

$$\text{Precision}_{micro} = \frac{\text{TP}_A + \text{TP}_B}{\text{TP}_A + \text{TP}_B + \text{FP}_A + \text{FP}_B} \qquad (18a)$$

where the true positives and false positives for class A are amalgamated with their counterparts for class B. Since the four-term sum in the denominator of expression (18a) is equal to N, we can also write

$$\text{Precision}_{micro} = \frac{\text{TP}_A + \text{TP}_B}{N} \qquad (18b)$$

It is to be noted in the meantime that expression (18b) is the same as the model accuracy defined in (10), and thus

$$\text{Precision}_{\text{micro}} = \text{Accuracy} \qquad (18c)$$

The weighted-average precision is

$$\text{Precision}_{\text{weighted}} = \frac{\text{N}_\text{A}(\text{Precision}_\text{A}) + \text{N}_\text{B}(\text{Precision}_\text{B})}{\text{N}} \qquad (19a)$$

or, by Eqs. (15) and (16a),

$$\text{Precision}_{\text{weighted}} = \frac{\frac{\text{N}_\text{A}}{\text{P}_\text{A}}(\text{TP}_\text{A}) + \frac{\text{N}_\text{B}}{\text{P}_\text{B}}(\text{TP}_\text{B})}{\text{N}} \qquad (19b)$$

where Precision$_A$ and Precision$_B$ are weighted by N$_A$ and N$_B$, respectively.

In Fig. 8, classes A and B have balanced datasets. Since Precision$_A$ = 0.843 and Precision$_B$ = 0.884,

$$\text{Precision}_{macro} = \frac{0.843 + 0.884}{2} = 0.864$$

Since TP$_A$ = 215, TP$_B$ = 190, and N= 470,

$$\text{Precision}_{micro} = \frac{215 + 190}{470} = 0.862 \qquad (= \text{Accuracy})$$

Since N$_A$ = 240 and N$_B$ = 230,

$$\text{Precision}_{\text{weighted}} = \frac{240(0.843) + 230(0.884)}{470} = 0.863$$

Example 5

From the confusion matrix of Fig. 15, determine the macro-, micro-, and weighted-average precisions of the classification model.

Solution

The classes A and B in Fig. 15 have imbalanced datasets. We have

$$\text{Precision}_A = 990/(990 + 48) = 0.954$$

$$\text{Precision}_B = 2/(2 + 10) = 0.167$$

Using Eqs. (17), (18b), and (19), we obtain

$$\text{Precision}_{\text{macro}} = (0.954 + 0.167)/2 = 0.561$$

$$\text{Precision}_{\text{micro}} = (990 + 2)/1050 = 0.945$$

$$\text{Precision}_{\text{wieghted}} = [1000(0.954) + 50(0.167)]/1050 = 0.917$$

C. Recall (Sensitivity)

The recall (also termed sensitivity) is the ratio of the number of samples correctly classified as positive to the number of all actual positive samples. From the first row of Fig. 2, we have

$$\text{Recall} = \frac{\text{TP}}{\text{N}_+} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad (20)$$

In Fig. 1, where TP =100 and FN =5,

$$\text{Recall} = 100/(100+5) = 0.952 \qquad (95.2\ \%)$$

In an ideal case when FN=0, the recall attains its maximum value of 1.0, meaning that all actual samples of the positive class are correctly classified (TP=N$_+$), with zero type II error. See Fig. 20. Specifically, expression (20) is the recall of the positive class.

		Predicted +ve	Predicted -ve	
Actual	+ve	TP = N$_+$	FN=0	N$_+$
	-ve	FP	TN	N$_-$

Fig. 20. Maximum recall

For two classes A anal B, we write

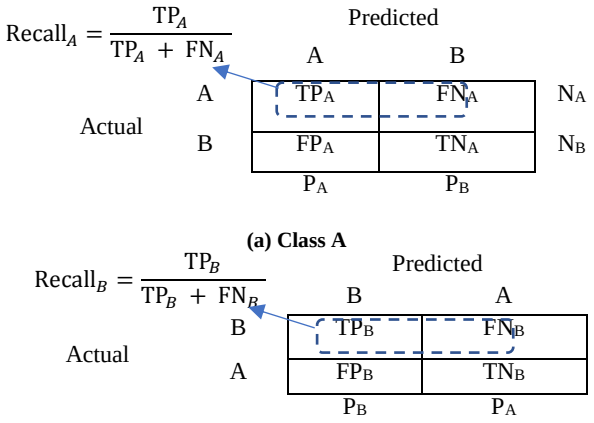
$$\text{Recall}_A = \frac{\text{TP}_A}{N_A} = \frac{\text{TP}_A}{\text{TP}_A + \text{FN}_A} \qquad (21)$$

$$\text{Recall}_B = \frac{\text{TP}_B}{N_B} = \frac{\text{TP}_B}{\text{TP}_B + \text{FN}_B} \qquad (22a)$$

See Fig. 21. In expression (21), two class-A building blocks TP$_A$ and FN$_A$ (first row in Fig. 21a) are used and, similarly, two class-B building blocks TP$_B$ and FN$_B$ (first row in Fig. 21b) are used in expression (22a). The recall of a certain class is thus the ratio of the number of samples of the class correctly classified as belonging to this class (true positives) to the number of all actual samples of the same class (true positives plus false negatives).

By relationships (6), Recall$_B$ in (22a) can also be expressed in terms of two class-A building blocks TN$_A$ and FP$_A$ (second row in Fig. 21a) as

$$\text{Recall}_B = \frac{\text{TN}_A}{\text{TN}_A + \text{FP}_A} \qquad (22b)$$



(a) Class A

(b) Class B

Fig. 21 Recalls of classes A and B as obtained from two forms of confusion matrix

hat is, both Recall$_A$ and Recall$_B$ can be obtained from the form of confusion matrix in Fig. 21a alone, by considering the two rows of the matrix, respectively, as Fig. 22 illustrates. Similar arguments apply Fig. 21b.

In Fig. 8, TP$_A$ = 215, FN$_A$ = 25, TN$_A$ = 190, and FP$_A$ = 40 and therefore

$$\text{Recall}_A = 215/(215 + 25) = 0.896$$

$$\text{Recall}_B = 190/(190 + 40) = 0.826$$

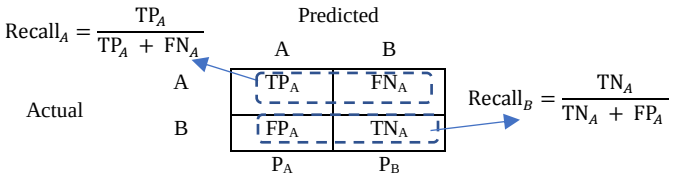
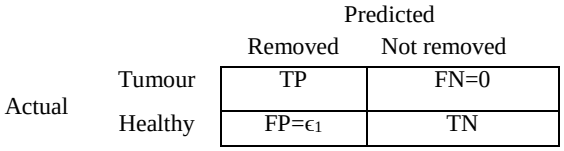


Fig. 22. Recalls of classes A and B as obtained from one form of confusion matrix

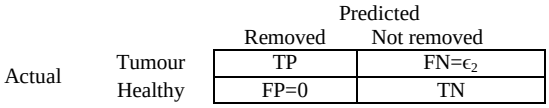
Often, there exists an inverse relationship between precision and recall in the sense that it is possible to increase one at the cost of decreasing the other. Brain surgery provides a comprehensible situation of the implied trade-off. Consider a surgeon removing cancer tumour from a patient's brain. The surgeon is keen to remove all tumour cells because any such cells left would regenerate the tumour. At the same time, the surgeon should avoid removing any healthy cells not ta cause the patient to suffer from impaired brain functions. Nevertheless, in the careful endeavor to ensure that all tumour cells have been removed, the surgeon mistakenly may remove some (a small number ϵ_1) of healthy cells. This is a case of decreasing precision and increasing recall.

On the other hand, the surgeon is keen to ensure that no healthy cells have been removed, but by mistake, some (ϵ_2) tumour cells may not be removed. This is a case of decreasing recall and increasing precision. That is to say, low precision (high recall) guarantees the removal of all tumour cells but gives an opportunity for some healthy cells to be removed also.

B contrast, high in precision (low recall) guarantees that all healthy cells are not removed but some tumour cells may not be removed as well. See Fig. 23 for a corresponding confusion matrix, where two classes are identified: A Tumour (positive) class which has the tumour cells and a healthy (negative) class which has the healthy cells.



(a) Case 1: Low precision, high recall



(b) Case 2: High precision, low recall

Fig. 23 Trade-off between precision and recall

Here, TP is the number of tumour cells correctly removed, FP is the number of healthy cells incorrectly removed, FN is the number of tumour cells incorrectly not removed, and TN is the number of healthy cells correctly not removed. Figure 23a represents case 1, that of low precision anal high recall (FP = ϵ_1, FN = 0; non-zero type I error), while Fig. 23b represents case 2, that of high precision and low recall (FP = 0 , FN = ϵ_2; non-zero type II error).

Paying attention to the Tumour class, Fig. 23a gives

$$\text{Precision(case 1)}_{\text{tumour}} = \frac{\text{TP}}{\text{TP} + \epsilon_1} \quad (<1)$$

$$\text{Recall(case 1)}_{\text{tumour}} = 1 \quad \text{(maximum)}$$

and Fig. 23b gives

$$\text{Precision(case 2)}_{\text{tumour}} = 1 \quad \text{(maximum)}$$

$$\text{Recall(case 2)}_{\text{tumour}} = \frac{\text{TP}}{\text{TP} + \epsilon_2} \quad (<1)$$

LE is conceivable that precision is an indication of 'quality' and recall is an indication of 'quantity', as implied by the definitions of precision in Eq. (14) and recall in Eq. (20). In the example of brain surgery, Fig. 23a shows that the precision is the number of tumour cells removed out of the total number of cells removed. This indicates the quality of surgery success. The recall, on the other hand, is the number of tumour cells removed out of the total number of tumour cells. This is the quantity of successful surgery results.

Moving on to the recall of the classification model, we define the macro-, micro-, and weighted-average recalls. In line with Eqs. (17), (18), and (19) for average precisions of two classes A and B, we have

$$\text{Recall}_{macro} = \frac{\text{Recall}_A + \text{Recall}_B}{2} \quad (23)$$

$$\text{Recall}_{micro} = \frac{\text{TP}_A + \text{TP}_B}{\text{TP}_A + \text{TP}_B + \text{FN}_A + \text{FN}_B} = \frac{\text{TP}_A + \text{TP}_B}{N} \quad (24)$$

$$\text{Recall}_{\text{weighted}} = \frac{\text{N}_\text{A}(\text{Recall}_\text{A}) + \text{N}_\text{B}(\text{Recall}_\text{B})}{\text{N}} \quad (25)$$

It is, however, to be noted that

$$\text{Recall}_{micro} = \text{Recall}_{weighted} \quad (26)$$

as is deduced by substituting for Recall$_A$ and Recall$_B$ from Eqs. (21) and (22a), respectively, into Eq.(25)

$$\text{Recall}_{weighted} = \frac{\text{N}_A(\frac{TP_A}{N_A}) + \text{N}_B(\frac{TP_B}{N_B})}{\text{N}} = \frac{\text{TP}_A + \text{TP}_B}{N} = \text{Recall}_{micro}$$

It is also evident from Eqs. (24) and (18b) that

$$\text{Recall}_{micro} = \text{Precision}_{micro} \quad (27)$$

and moreover by Eq. (8c),

$$\text{Recall}_{micro} = \text{Accuracy} \quad (28)$$

Combining Eqs. (26), (27), and (28), we can write

$$\text{Accuracy} = \text{Precision}_{micro} = \text{Recall}_{micro} = \text{Recall}_{weighted} \quad (29)$$

In Fig. 8, since $\text{Recall}_A = 0.896$ and $\text{Recall}_B = 0.826$,

$$\text{Recall}_{macro} = \frac{0.896 + 0.826}{2} = 0.861$$

Since $TP_A = 215$, $TP_B = 190$, and $N = 470$,

$$\text{Recall}_{micro} = \frac{215 + 190}{470} = 0.862$$

Since $N_A = 240$ and $N_B = 230$,

$$\text{Recall}_{weighted} = \frac{240(0.896) + 230(0.826)}{470} = 0.862$$

Equation (29) is already satisfied, where

$$\text{Accuracy} = \text{Precision}_{micro} = \text{Recall}_{micro} = \text{Recall}_{weighted} = 0.862$$

Example 6

For the confusion matrix of Fig. 15, determine the macro-, micro, and weighted-average recalls of the classification model.

Solution

From Eqs. (21) and (22a),

$\text{Recall}_A = 990/(990 + 10) = 0.99$

$\text{Recall}_B = 2/(2 + 48) = 0.04$

From Eqs. (23), (24), and (27),

$\text{Recall}_{macro} = (0.99 + 0.04)/2 = 0.515$

$\text{Recall}_{micro} = (990 + 2)/1050 = 0.945$

$\text{Recall}_{weighted} = \text{Recall}_{micro} = 0.945$

We remark that two other expressions, related to recall, are used as performance measures. These are TPR (true positive rate) and FNR (false negative rate). The TPR is the same thing as recall, Eq. (20),

$$\text{TPR} = \frac{\text{TP}}{\text{N}_+} = \frac{\text{TP}}{\text{TP + FN}} = \text{Recall} \quad (30)$$

and the FNR is

$$\text{FNR} = 1 - \text{TPR} = \frac{\text{FN}}{\text{N}_+} = \frac{\text{FN}}{\text{TP + FN}} \quad (31)$$

i.e. the ratio of the number of samples incorrectly classified as negative to the number of all actual positive samples.

Example 7

In Example 6, determine

(a) TPR and FNR of each of the two classes A and B.

(b) TPR_{macro} and FNR_{macro} of the classification model.

Solution

Using definitions (30) and (31) and results of Example 6, we obtain for class A,

$\text{TPR}_A = \text{Recall}_A = 0.99$

$\text{FNR}_A = 1 - \text{TPR}_A = 1 - 0.99 = 0.01$

and for class B,

$\text{TPR}_B = \text{Recall}_B = 0.04$

$\text{FNR}_B = 1 - \text{TPR}_B = 1 - 0.04 = 0.96$

For the classification model,

$\text{TPR}_{macro} = \text{Recall}_{macro} = 0.515$

$\text{FNR}_{macro} = 1 - \text{TPR}_{macro} = 1 - 0.515 = 0.485$

D. Specificity

The specificity is the ratio of the number of samples correctly classified as negative to the number of all actual negative samples. From the second row of Fig. 2, we have

$$\text{Specificity} = \frac{\text{TN}}{\text{N}_-} = \frac{\text{TN}}{\text{TN + FP}} \quad (32)$$

In Fig.1, where TN =90 and FP = 10,

Specificity = 90/(90 + 10) = 0.9

In an ideal case when FP = 0 (zero type I error), the specificity has its maximum value of 1.0, meaning that all actual samples of the negative class are correctly classified (TN = N_). Remember that the same condition FP = 0, Fig. 17, makes the precision also at its maximum value of 1.0. See Fig. 24. Expression (32) is in fact the specificity of the positive class.

For two classes A and B,

$$\text{Specificity}_A = \frac{\text{TN}_A}{\text{N}_B} = \frac{\text{TN}_A}{\text{TN}_A + \text{FP}_A} \quad (33)$$

$$\text{Specificity}_B = \frac{\text{TN}_B}{\text{N}_A} = \frac{\text{TN}_B}{\text{TN}_B + \text{FP}_B} \quad (34a)$$

		Predicted +ve	Predicted -ve	
Actual	+ve	TP	FN	N_+
	-ve	FP=0	TN= N_-	N_-

Fig. 24. Maximum specificity (maximum precision)

See Fig.25. In expression (33), two class-A building blocks TN_A and FP_A (second row in Fig. 25a) are used, and two class-B building blocks TN_B and FP_B (second row in Fig. 25b) are used in expression (34a). The specificity of one class is thus the ratio of the number of samples of the other class correctly classified as belonging to the other class (true negatives) to the number of all actual samples of the other class (true negatives plus false positives).

By relationships (6), Specificity_B in (34a) can also be expressed in terms of two class-A building blocks TP_A and FN_A (first row in Fig. 25a) as

$$\text{Specificity}_B = \frac{\text{TP}_A}{\text{N}_A} = \frac{\text{TP}_A}{\text{TP}_A + \text{FN}_A} \quad (34b)$$

Therefore, both Specificity_A and Specificity_B can be obtained from one form of confusion matrix, that of Fig. 25a, by considering the two rows of the matrix, respectively. See Fig. 26. Similar arguments apply to Fig. 25b.

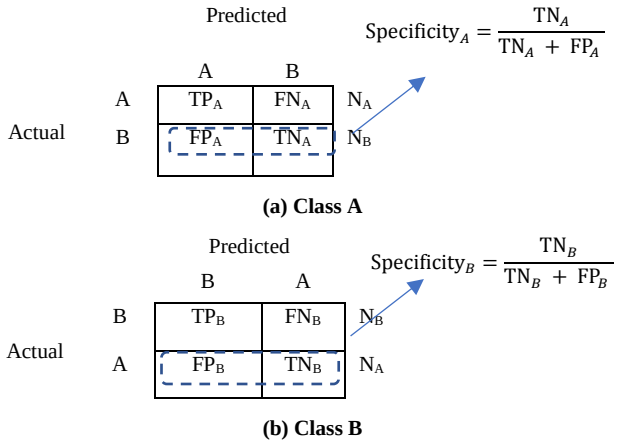


Fig. 25. Specificities of classes A and B as obtained from two forms of confusion matrix

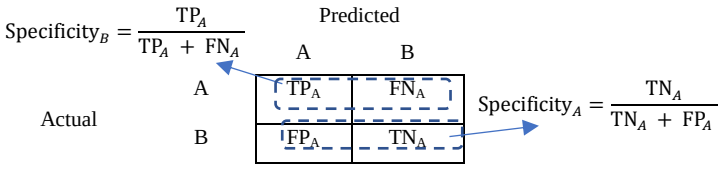


Fig. 26. Specificities of classes A and B as obtained from one form of confusion matrix

By comparison, it is clear that Eqs. (33) and (22b) are identical and so are Eqs. (34b) and (21), providing the results

$$\text{Specificity}_A = \text{Recall}_B \quad (35)$$

$$\text{Specificity}_B = \text{Recall}_A \quad (36)$$

i.e. the specificity of one class is nothing but the recall of the other class.

In Fig. 8, $\text{Recall}_A = 0.896$ and $\text{Recall}_B = 0.826$ and therefore

$$\text{Specificity}_A = 0.826 \quad , \quad \text{Specificity}_B = 0.896$$

The macro-, micro-, and weighted-average specificities are defined in analogy to both average precisions and average recalls. The first two average specificities take several forms based on previously derived relationships. We have

$$\text{Specificity}_{macro} = \frac{\text{Specificity}_A + \text{Specificity}_B}{2} = \frac{\text{Recall}_A + \text{Recall}_B}{2} \quad (37)$$

implying that

$$\text{Specificity}_{macro} = \text{Recall}_{macro} \quad (38)$$

$$\text{Specificity}_{micro} = \frac{\text{TN}_A + \text{TN}_B}{\text{TN}_A + \text{TN}_B + \text{FP}_A + \text{FP}_B} = \frac{\text{TN}_A + \text{TN}_B}{N} = \frac{\text{TP}_A + \text{TN}_A}{N} \quad (39)$$

implying that

$$\text{Specificity}_{micro} = \text{Recall}_{micro} = \text{Precision}_{micro} \quad (40)$$

It turns out that $\text{Specificity}_{micro}$ is to be incorporated in Eq. (29), so that we can write

$$\text{Accuracy} = \text{Precision}_{micro} = \text{Recall}_{micro}$$

$$= \text{Recall}_{weighted} = \text{Specificity}_{micro} \quad (41)$$

The weighted-average specificity is

$$\text{Specificity}_{\text{weighted}} = \frac{\text{N}_\text{A}(\text{Specificity}_\text{A}) + \text{N}_\text{A}(\text{Specificity}_\text{B})}{\text{N}} \quad (42a)$$

or, by Eqs. (33) and (34a),

$$\text{Specificity}_{weighted} = \frac{\frac{\text{N}_\text{A}}{\text{N}_\text{B}}(\text{TN}_\text{A}) + \frac{\text{N}_\text{B}}{\text{N}_\text{A}}(\text{TN}_\text{B})}{\text{N}} \quad (42b)$$

Example 8

For the confusion matrix of Fig. 15, determine the macro-, micro-, and weighted-average specificities of the classification model.

Solution

Using results of Example 6, we obtain

$\text{Specificity}_\text{A} = \text{Recall}_\text{B} = 0.04$

$\text{Specificity}_\text{B} = \text{Recall}_\text{A} = 0.99$

$\text{Specificity}_\text{macro} = \text{Recall}_\text{macro} = 0.515$

$\text{Specificity}_\text{micro} = \text{Recall}_\text{micro} = 0.945$

From Eq. (42),

$\text{Specificity}_\text{weighted} = [1000(0.04) + 50(0.99)]/1050 = 0.085$

In addition to TPR and FNR expressed along with recall at the end of Subsection 3.3, we here define TNR (true negative rate) and FPR (false positive rate). The TNR is the same thing as specificity, Eq. (32),

$$\text{TNR} = \frac{\text{TN}}{\text{N}_-} = \frac{TN}{\text{TN} + \text{FP}} = \text{Specificity} \quad (43)$$

and the FPR is

$$\text{FPR} = 1 - \text{TNR} = \frac{FP}{\text{N}_-} = \frac{\text{FP}}{\text{TN} + \text{FP}} \quad (44)$$

i.e. the ratio of the number of samples incorrectly classified as positive to the number of all actual negative samples.

Example 9

In Example 6, determine
(a) TNR and FPR of each of the two classes A and B.
(b) TNR_micro and FPR_micro of the classification model.

Solution

Using definitions (43) and (44) and results of Example 8, we obtain for class A,

$\text{TNR}_\text{A} = \text{Specificity}_\text{A} = 0.04$

$\text{FPR}_\text{A} = 1 - \text{TNR}_\text{A} = 1 - 0.04 = 0.96$

and for class B,

$\text{TNR}_\text{B} = \text{Specificity}_\text{B} = 0.99$

$\text{FPR}_\text{B} = 1 - \text{TNR}_\text{B} = 1 - 0.99 = 0.01$

For the classification model,

$\text{TNR}_\text{micro} = \text{Specificity}_\text{micro} = 0.945$

$\text{FPR}_\text{micro} = 1 - \text{TNR}_\text{micro} = 1 - 0.945 = 0.055$

E. Balanced accuracy

In Subsection (*A*), we emphasized the fact that the accuracy measure of a binary classifier can be misleading when the datasets of the two classes are imbalanced. A performance measure, known as balanced accuracy, is thus introduced. It combines recall and specificity in the form

$$\text{Balanced accuracy} = \frac{\text{Recall} + \text{Specificity}}{2} \quad (45)$$

i.e. the arithmetic average of recall and specificity. Remember that recall, TP/N_+, deals with only the positive class while specificity, TN/N_-, deals with only the negative class. A combination of these two measures proves advantageous, especially for imbalanced datasets.

In Fig. 1, where Recall = 0.952 and Specificity = 0.9,

Balanced accuracy = (0.952 + 0.9)/2 = 0.926

For two classes A and B, the balanced accuracy of the model is the arithmetic average of recall and specificity of either class A or class B;

$$\text{Balanced accuracy} = \frac{Recall_A + \text{Specificity}_A}{2} = \frac{Recall_B + \text{Specificity}_B}{2} \quad (46)$$

Make sure that the two expressions in Eq. (46), by Eqs. (35) and (36), are identical.

In Fig. 8, where $\text{Recall}_A = \text{Specificity}_B = 0.896$ and $\text{Recall}_B = \text{Specificity}_A = 0.826$,

Balanced accuracy = (0.896 + 0.826)/2 = 0.861

Equation (46) can alternatively be written as

$$\text{Balanced accuracy} = \frac{Recall_A + Recall_B}{2} = \frac{\text{Specificity}_A + \text{Specificity}_B}{2} \quad (47)$$

This provides a noticeable result that the balanced accuracy is the same as the macro-average recall of classes A and B or the macro-average specificity of the two classes,

$$\text{Balanced accuracy} = \text{Recall}_{macro} = \text{Specificity}_{macro} \quad (48)$$

Example 10

For the confusion matrix of Fig. 15, determine the balanced accuracy of the classification model.

Solution

Using the value of the macro-average recall, or the macro-average specificity, in the solution of Example 8, Eq. (48) yields

Balanced accuracy = 0.515

A comparison between balanced accuracy and accuracy is in order. Consider a binary classifier with the confusion matrix of Fig. 27, where the datasets of classes A and B are balanced ($N_A = 195$, $N_B = 192$). The accuracy, by Eq. (10a), is

Accuracy = (185 + 10)/(195 +192) = 0.943

The recalls of classes A and B, by Eqs. (21) and (22b), are

$\text{Recall}_A = 185/195 = 0.949$

$\text{Recall}_B = 180/192 = 0.938$

Therefore, the balanced accuracy, by Eq. (47), is

		Predicted A	Predicted B	
Actual	A	185	10	$N_A = 195$
	B	12	180	$N_B = 192$

Fig. 27. Confusion matrix with balanced datasets

Balanced accuracy = (0.945 + 0.938)/2 = 0.944

The values of accuracy and balanced accuracy are seen to be approximately the same. The reason is that the datasets of the two classes are balanced.

However, for a binary classifier with the confusion matrix of Fig. 28, where the datasets of the two classes are imbalanced ($N_A = 10$, $N_B = 190$), we have

Accuracy = (0 + 190)/(10 +190) = 0.95 (95%)

		Predicted A	Predicted B	
Actual	A	0	10	$N_A = 10$
	B	0	190	$N_B = 190$

Fig. 28. Confusion matrix with imbalanced datasets

The accuracy is calculated to be of a high value (95%), giving an impression that the classifier performs quite properly. But this is far from reality. Although the classifier correctly predicts all samples of class B ($FN_B = 0$), it does not correctly predict any sample of class A ($TP_A = 0$). The classifier has a deficiency in performance, not detected by accuracy. In other words, the 95% accuracy is misleading and cannot be relied upon. Balanced accuracy can be taken into account instead. Since

$\text{Recall}_A = 0/10 = 0$

$\text{Recall}_B = 190/190 = 1$

then

Balanced accuracy = (0 + 1)/2 = 0.5 (50%)

which is considerably less than the value of accuracy and may thus be reliable. The difference in the values of accuracy and balanced accuracy is due to the imbalance of datasets.

F. F_β measure and F1 score

The precision and recall are commonly combined to provide a performance measure called F_β measure, defined as

$$F_\beta = \frac{1}{\frac{\beta}{\text{Precision}} + \frac{1-\beta}{\text{Recall}}} = \frac{\text{Recall} \times \text{Precision}}{\beta(\text{Recall}) + (1-\beta)\text{Precision}} \quad (49)$$

This means that F_β is the weighted harmonic average of precision and recall. Here, B is a positive fractional factor, $0 < \beta < 1$, which reflects the importance of precision and recall with respect to each other. The greater β is, the greater importance is given to precision and, conversely, the smaller β is, the greater importance is given to recall. Indeed, there should be a trade-off between precision and recall, relying on the particulars of the classification problem at hand; cf. the example of brain surgery in Subsection 3.3.

Substituting for precision and recall from Eqs. (14) and (20), respectively, into Eq. (49) , F_β is formulate as

$$F_\beta = \frac{TP}{TP + \beta(FP) + (1-\beta)FN} \quad (50)$$

Note the similarity in form among the expressions of precision in Eq. (14), recall in Eq. (20), and F_β in Eq. (50), where in the respective denominators, FP is replaced by FN and both (FP and FN) are replaced by the weighted sum of FP and FN.

The special case

$$\beta = 0.5 \qquad (51)$$

is of particular interest, where precision and recall are of equal weight (importance). The F_β measure under condition (51) is referred to as the F1 score which, by Eq. (49), takes the form

$$F1 = \frac{2}{\frac{1}{\text{Precision}}+\frac{1}{\text{Recall}}} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \qquad (52)$$

That is, F1 is the harmonic average of precession and recall. See Fig. 29 or a graphical representation. The distance h is equal to 0.5 F1, and is less than the smaller of precision and recall. The proof is a simple geometric exercise.

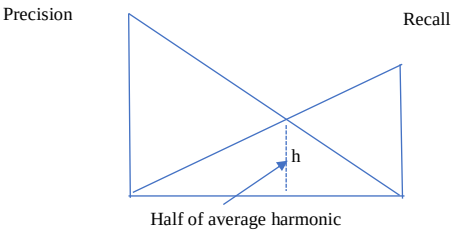


Fig. 29. Harmonic average of precision and recall

Equation (52), in view of Eq. (50) with β =0.5, becomes

$$F1 = \frac{TP}{TP+0.5(FP+FN)} \qquad (53)$$

where the arithmetic average of FP and FN replaces FP in Eq.(14) and FN in Eq. (20).

In Fig. 1, where Precision = 0.909 and Recall = 0.952, Eq. (49) for β = 0.8 (as an example) and Eq. (52) yield

$$F_{\beta=0.8} = \frac{0.952\times 0.909}{0.8(0.952)+(1-0.8)0.909} = 0.917$$

$$F1 = \frac{2\times 0.909\times 0.952}{0.909+0.952} = 0.93$$

The same results are of course produced by the equivalent expressions (50) and (53).

For two classes A and B, we have for class A,

$$F_{\beta A} = \frac{\text{Precision}_A\times\text{Recall}_A}{\beta(\text{Recall}_A)+(1-\beta)\text{Precision}_A}$$

$$= \frac{TP_A}{TP_A+\beta(FP_A)+(1-\beta)FN_A} \qquad (54)$$

$$F1_A = \frac{2\times\text{Precision}_A\times\text{Recall}_A}{\text{Precision}_A+\text{Recall}_A}$$

$$= \frac{TP_A}{TP_A+0.5(FP_A+FN_A)} \qquad (55)$$

and similar expressions apply to class B. In certain classification problems, $F_{\beta A}$, and $F_{\beta B}$ as well as $F1_A$, and $F1_B$, pertaining to the individual classes, can be useful in their own right.

For the classification model, we have

$$F_{\beta model} = \frac{\text{Precision}_{model}\times\text{Recall}_{model}}{\beta(\text{Recall}_{model})+(1-\beta)\text{Precision}_{model}} \qquad (56)$$

$$F1_{model} = \frac{2\times\text{Precision}_{model}\times\text{Recall}_{model}}{\text{Precision}_{model}+\text{Recall}_{model}} \qquad (57)$$

Equations (56) and (57) represent the macro-, micro, or weighted-average F_β and F1 of the model, respectively, where Precision_{model} is correspondingly the macro-, micro-, or weighted-average precision of the model, and Recall_{model} is defined in a similar way. We should always take Eqs. (56) and (57) into account when we calculate the average F_β and F1 for the model. For example, $F1_{macro}$ is not the arithmetic average of $F1_A$ and $F1_\beta$ but it is, by definition, the harmonic average of Precision_{macro} and Recall_{macro}.

For the micro-average, $F_{\beta micro}$ reduces to $F1_{micro}$,

$$F_{\beta micro} = F1_{micro} \qquad (58)$$

and the effect of β ceases to exist. In this case,

$$F_{\beta micro} = F1_{micro} = \text{Precision}_{micro} = \text{Recall}_{micro} \qquad (59)$$

which follows in view of Eq. (27). Remember the fact that the harmonic average of two equal values is the same as either value.

Aggregating Eqs. (41) and (53) leads to

$$\text{Accuracy} = \text{Precision}_{micro} = \text{Recall}_{micro}$$

$$= \text{Recall}_{weighted} = \text{Specificity}_{micro}$$

$$= F_{\beta micro} = F1_{micro} = \frac{TP_A+TP_B}{N} \qquad (60)$$

and we find out (remarkably) that seven measures are defined. by one and the same expression, $\frac{TP_A+TN_A}{N}$.

Example 11

For the confusion matrix of Fig 15, determine

(a) F1 score of class A and that of class B.

(b) Macro-, micro-, and weighted-average F1 scores of the classification model.

Solution

Collecting results of Examples 5 and 6,

Precision_A = 0.954, Precision_B = 0.167, Recall_A = 0.99, Recall_B = 0.04, Precision_{macro} = 0.561 , Recall_{macro} = 0.515, Precision_{micro} = Recall_{micro} = $\text{Recall}_{weighted}$ = 0.945, $\text{Precision}_{weighted}$ = 0.917

From Eq. (55),

$$F_{1A} = \frac{2\times 0.954\times 0.99}{0.954+0.99} = 0.972$$

$$F_{1B} = \frac{2\times 0.167\times 0.04}{0.167+0.04} = 0.065$$

From Eqs. (57) and (59),

$$F1_{macro} = \frac{2\times\text{Precision}_{macro}\times\text{Recall}_{macro}}{\text{Precision}_{macro}+\text{Recall}_{macro}}$$

$$= \frac{2\times 0.561\times 0.515}{0.561+0.515} = 0.537$$

$$F1_{micro} = \text{Precision}_{micro} = 0.945$$

$$F1_{weighted} = \frac{2\times\text{Precision}_{weighted}\times\text{Recall}_{weighted}}{\text{Precision}_{weighted}+\text{Recall}_{weighted}}$$

$$= \frac{2\times 0.917\times 0.945}{0.917+0.945} = 0.931$$

III. SUMMARY OF RESULTS FOR BINARY CLASSIFICATION

Table 1 lists the expressions of the performance measures and their interrelationships for binary classification with two classes A and B. The subscript 'class' in rows 5 and 6 symbolizes either class A or class B, and the subscript 'model'

in rows 17 and 16 symbolizes either macro, micro-, or weighted-average.

Table 1. Performance measures for binary classification with two classes A and B

#	Measure
1	$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{N}}$
2	$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$
3	$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$ (Sensitivity)
4	$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$
5	$\text{F}_{\beta\text{class}} = \frac{\text{Precision}_{\text{class}} \times \text{Recall}_{\text{class}}}{\beta(\text{Recall}_{\text{class}}) + (1-\beta)\text{Precision}_{\text{class}}} = \frac{\text{TP}_{\text{class}}}{\text{TP}_{\text{class}} + \beta(\text{FP}_{\text{class}}) + (1-\beta)\text{FN}_{\text{class}}}$
6	$\text{F1}_{\text{class}} = \frac{2 \times \text{Precision}_{\text{class}} \times \text{Recall}_{\text{class}}}{\text{Precision}_{\text{class}} + \text{Recall}_{\text{class}}} = \frac{\text{TP}_{\text{class}}}{\text{TP}_{\text{class}} + 0.5(\text{FP}_{\text{class}} + \text{FN}_{\text{class}})}$
7	$\text{Balanced Accuracy} = \frac{\text{Recall} + \text{Specificity}}{2}$
8	$\text{Precision}_{macro} = \frac{\text{Precision}_A + \text{Precision}_B}{2}$
9	$\text{Recall}_{macro} = \frac{\text{Recall}_A + \text{Recall}_B}{2}$
10	$\text{Specificity}_{macro} = \frac{\text{Specificity}_A + \text{Specificity}_B}{2}$
11	$\text{Precision}_{micro} = \frac{\text{TP}_A + \text{TP}_B}{2} = \text{Accuracy}$
12	$\text{Recall}_{micro} = \text{Precision}_{micro}$
13	$\text{Specificity}_{micro} = \text{Precision}_{micro}$
14	$\text{Precision}_{\text{weighted}} = \frac{N_A(\text{Precision}_A) + N_A(\text{Precision}_B)}{\text{N}} = \frac{\frac{\text{N}_\text{A}}{\text{P}_\text{A}}(\text{TP}_\text{A}) + \frac{\text{N}_\text{B}}{\text{P}_\text{B}}(\text{TP}_\text{B})}{\text{N}}$
15	$\text{Recall}_{\text{weighted}} = \frac{\text{N}_\text{A}(\text{Recall}_\text{A}) + \text{N}_\text{A}(\text{Recall}_\text{B})}{\text{N}} = \frac{\text{TP}_A + \text{TP}_B}{N} = \text{Accuracy}$
16	$\text{Specificity}_{\text{weighted}} = \frac{\text{N}_\text{A}(\text{Specificity}_\text{A}) + \text{N}_\text{B}(\text{Specificity}_\text{B})}{\text{N}} = \frac{\frac{\text{N}_\text{A}}{\text{N}_\text{B}}(\text{TN}_\text{A}) + \frac{\text{N}_\text{B}}{\text{N}_\text{A}}(\text{TN}_\text{B})}{\text{N}}$
17	$\text{F}_{\beta\text{model}} = \frac{\text{Precision}_{\text{model}} \times \text{Recall}_{\text{model}}}{\beta(\text{Recall}_{\text{model}}) + (1-\beta)\text{Precision}_{\text{model}}}$
18	$\text{F1}_{\text{model}} = \frac{2 \times \text{Precision}_{\text{model}} \times \text{Recall}_{\text{model}}}{\text{Precision}_{\text{model}} + \text{Recall}_{\text{model}}}$
19	$\text{Specificity}_A = \text{Recall}_B$ $\text{Specificity}_B = \text{Recall}_A$
20	$\text{Specificity}_{macro} = \text{Recall}_{macro}$
21	$\text{Balanced accuracy} = \frac{\text{Recall}_\text{A} + \text{Specificity}_\text{B}}{2} = \text{Recall}_{\text{macro}} = \frac{\text{Specificity}_\text{A} + \text{Specificity}_\text{B}}{2} = \text{Specificity}_{\text{macro}}$
22	$\text{F}_{\beta\text{micro}} = \text{F1}_{\text{micro}} = \text{Precision}_{\text{micro}}$
23	$\text{Accuracy} = \text{Precision}_{\text{micro}} = \text{Recall}_{\text{micro}} = \text{Recall}_{\text{weighted}} = \text{Specificity}_{\text{micro}} = \text{F}_{\beta\text{micro}} = \text{F1}_{\text{micro}} = \frac{\text{TP}_A + \text{TP}_B}{\text{N}}$

Acknowledgement

I would like to Thank Prof. Dr. Amany Sarhan for help valuable help in writing and editing this tutorial and incorporating the references used in this tutorial, hoping that this material can help in a deeper understanding of this important topic.

REFERENCES

1. David M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation," Journal of machine learning technologies, volume 2, issue 1, pp-37-63, 2011.
2. Hasnae Zerouaouib and Ali Idri, "Deep hybrid architectures for binary classification of medical breast cancer images," Biomedical Signal Processing and Control," volume 71, Part B, January 2022.
3. I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. Data Mining. Morgan Kaufmann, fourth edition, 2017.
4. Kai Ming Ting, "Precision and Recall," Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA, 2011.
5. Kanvinde, Nandan, Abhishek Gupta, and Raunak Joshi. "Binary classification for high dimensional data using supervised non-parametric ensemble method." arXiv preprint arXiv:2202.07779, 2022.
6. W. Siblini, J. Fréry, L. He-Guelton, F. Oblé, and Y. Q. Wang, "Master Your Metrics with Calibration. In Berthold, M., Feelders, A., and G., K., editors, Advances in Intelligent Data Analysis XVIII. IDA 2020. Springer, Cham. Lecture Notes in Computer Science, vol. 12080.
7. https://www.turing.com/kb/precision-recall-method