WIKIPEDIA

Confusion matrix

In the field of machine learning and specifically the problem of statistical classification, a **confusion matrix**, also known as an error matrix,[9] is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a **matching matrix**). Each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class, or vice versa – both variants are found in the literature.[10] The name stems from the fact that it makes it easy to see whether the system is confusing two classes (i.e. commonly mislabeling one as another).

It is a special kind of contingency table, with two dimensions ("actual" and "predicted"), and identical sets of "classes" in both dimensions (each combination of dimension and class is a variable in the contingency table).

Contents

Example

Given a sample of 12 pictures, 8 of cats and 4 of dogs, where cats belong to class 1 and dogs belong to class 0,

actual = [1,1,1,1,1,1,1,1,0,0,0,0],

assume that a classifier that distinguishes between cats and dogs is trained, and we take the 12 pictures and run them through the classifier. The classifier makes 9 accurate predictions and misses 3: 2 cats wrongly predicted as dogs (first 2 predictions) and 1 dog wrongly predicted as a cat (last prediction).

prediction = [0,0,1,1,1,1,1,1,0,0,0,1]

With these two labeled sets (actual and predictions), we can create a confusion matrix that will summarize the results of testing the classifier:

Predicted class / Actual class	Cat	Dog
Cat	**6**	2
Dog	1	**3**

In this confusion matrix, of the 8 cat pictures, the system judged that 2 were dogs, and of the 4 dog pictures, it predicted that 1 were cats. All correct predictions are located in the diagonal of the table (highlighted in bold), so it is easy to visually inspect the table for prediction errors, as values outside the diagonal will represent them.

In terms of sensitivity and specificity, the confusion matrix is as follows:

Predicted class / Actual class	P	N
P	TP	FN
N	FP	TN

Table of confusion

In predictive analytics, a **table of confusion** (sometimes also called a **confusion matrix**) is a table with two rows and two columns that reports the number of *false positives*, *false negatives*, *true positives*, and *true negatives*. This allows more detailed analysis than mere proportion of correct classifications (accuracy). Accuracy will yield misleading results if the data set is unbalanced; that is, when the numbers of observations in different classes vary greatly. For example, if there were 95 cats and only 5 dogs in the data, a particular classifier might classify all the observations as cats. The overall accuracy would be 95%, but in more detail the classifier would have a 100% recognition rate (sensitivity) for the cat class but a 0% recognition rate for the dog class. F1 score is even more unreliable in such cases, and here would yield over 97.4%, whereas informedness removes such bias and yields 0 as the probability of an informed decision for any form of guessing (here always guessing cat). Confusion matrix is not limited to binary

Terminology and derivations from a confusion matrix

condition positive (P)
the number of real positive cases in the data

condition negative (N)
the number of real negative cases in the data

---

true positive (TP)
eqv. with hit

true negative (TN)
eqv. with correct rejection

false positive (FP)
eqv. with false alarm, type I error or underestimation

false negative (FN)
eqv. with miss, type II error or overestimation

---

sensitivity, recall, hit rate, or true positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = 1 - \mathrm{FNR}$$

specificity, selectivity or true negative rate (TNR)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} = 1 - \mathrm{FPR}$$

precision or positive predictive value (PPV)

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} = 1 - \mathrm{FDR}$$

negative predictive value (NPV)

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}} = 1 - \mathrm{FOR}$$

miss rate or false negative rate (FNR)

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{P}} = \frac{\mathrm{FN}}{\mathrm{FN}+\mathrm{TP}} = 1 - \mathrm{TPR}$$

fall-out or false positive rate (FPR)

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} = 1 - \mathrm{TNR}$$

false discovery rate (FDR)

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TP}} = 1 - \mathrm{PPV}$$

false omission rate (FOR)

$$\mathrm{FOR} = \frac{\mathrm{FN}}{\mathrm{FN}+\mathrm{TN}} = 1 - \mathrm{NPV}$$

prevalence threshold (PT)

$$\mathrm{PT} = \frac{\sqrt{\mathrm{TPR}(-\mathrm{TNR}+1)}+\mathrm{TNR}-1}{(\mathrm{TPR}+\mathrm{TNR}-1)} = \frac{\sqrt{\mathrm{FPR}}}{\sqrt{\mathrm{TPR}}+\sqrt{\mathrm{FPR}}}$$

threat score (TS) or critical success index (CSI)

$$\mathrm{TS} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}}$$

---

accuracy (ACC)

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{P}+\mathrm{N}} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$$

classification and can be used in multi-class classifiers as well.[11]

According to Davide Chicco and Giuseppe Jurman, the most informative metric to evaluate a confusion matrix is the Matthews correlation coefficient (MCC).[12]

Assuming the confusion matrix above, its corresponding table of confusion, for the cat class, would be:

balanced accuracy (BA)

$$\mathrm{BA} = \frac{TPR + TNR}{2}$$

F1 score

is the harmonic mean of precision and sensitivity:

$$\mathrm{F}_1 = 2 \times \frac{\mathrm{PPV} \times \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

Matthews correlation coefficient (MCC)

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

Fowlkes–Mallows index (FM)

$$\mathrm{FM} = \sqrt{\frac{TP}{TP + FP} \times \frac{TP}{TP + FN}} = \sqrt{PPV \times TPR}$$

informedness or bookmaker informedness (BM)

$$\mathrm{BM} = \mathrm{TPR} + \mathrm{TNR} - 1$$

markedness (MK) or deltaP (Δp)

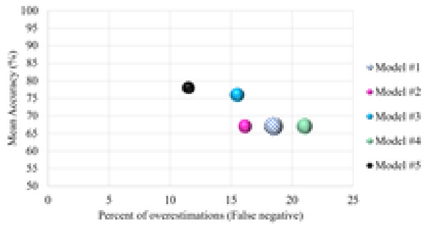
$$\mathrm{MK} = \mathrm{PPV} + \mathrm{NPV} - 1$$

Sources: Fawcett (2006),[1] Piryonesi and El-Diraby (2020),[2] Powers (2011),[3] Ting (2011),[4] CAWCR,[5] D. Chicco & G. Jurman (2020, 2021),[6][7] Tharwat (2018).[8]

Predicted class / Actual class	**Cat**	**Non-cat**
Cat	**6 true positives**	2 false negatives
Non-cat	1 false positive	**3 true negatives**

The final table of confusion would contain the average values for all classes combined.

Let us define an experiment from **P** positive instances and **N** negative instances for some condition. The four outcomes can be formulated in a 2×2 *confusion matrix*, as follows:



Comparing mean accuracy and percent of false negative (overestimation) of five machine learning (multi-class) classification models. Models #1, #2 and #4 have a very similar accuracy but different false negative or overestimation levels.[11]

		Predicted condition		Sources: [13][14][15][16][17][18][19][20]	
	Total population $= P + N$	**Positive (PP)**	**Negative (PN)**	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP), hit**	**False negative (FN), type II error, miss, underestimation**	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
Actual condition	**Negative (N)**	**False positive (FP), type I error, false alarm, overestimation**	**True negative (TN), correct rejection**	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN} = 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2 PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$

Confusion matrices with more than two categories

The confusion matrices discussed above have only two conditions: positive and negative. In some fields, confusion matrices can have more categories. For example, the table below summarises communication of a whistled language between two speakers, zero values omitted for clarity.[21]

Perceived vowel / **Vowel produced**	**i**	**e**	**a**	**o**	**u**
i	**15**		1		
e	1		1		
a			**79**	5	
o			4	**15**	3
u				2	**2**

See also

- Positive and negative predictive values

References

1. Fawcett, Tom (2006). "An Introduction to ROC Analysis" (http://people.inf.elte.hu/kiss/11dwhdm/roc.pdf) (PDF). *Pattern Recognition Letters*. **27** (8): 861–874. doi:10.1016/j.patrec.2005.10.010 (https://doi.org/10.1016%2Fj.patrec.2005.10.010).
2. Piryonesi S. Madeh; El-Diraby Tamer E. (2020-03-01). "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index". *Journal of Infrastructure Systems*. **26** (1): 04019036. doi:10.1061/(ASCE)IS.1943-555X.0000512 (https://doi.org/10.1061%2F%28ASCE%29IS.1943-555X.0000512).
3. Powers, David M. W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (https://www.researchgate.net/publication/228529307). *Journal of Machine Learning Technologies*. **2** (1): 37–63.
4. Ting, Kai Ming (2011). Sammut, Claude; Webb, Geoffrey I. (eds.). *Encyclopedia of machine learning*. Springer. doi:10.1007/978-0-387-30164-8 (https://doi.org/10.1007%2F978-0-387-30164-8). ISBN 978-0-387-30164-8.
5. Brooks, Harold; Brown, Barb; Ebert, Beth; Ferro, Chris; Jolliffe, Ian; Koh, Tieh-Yong; Roebber, Paul; Stephenson, David (2015-01-26). "WWRP/WGNE Joint Working Group on Forecast Verification Research" (https://www.cawcr.gov.au/projects/verification/). *Collaboration for Australian Weather and Climate Research*. World Meteorological Organisation. Retrieved 2019-07-17.
6. Chicco D., Jurman G. (January 2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941312). *BMC Genomics*. **21** (1): 6-1–6-13. doi:10.1186/s12864-019-6413-7 (https://doi.org/10.1186%2Fs12864-019-6413-7). PMC 6941312 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941312). PMID 31898477 (https://pubmed.ncbi.nlm.nih.gov/31898477).
7. Chicco D., Toetsch N., Jurman G. (February 2021). "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7863449). *BioData Mining*. **14** (13): 1-22. doi:10.1186/s13040-021-00244-z (https://doi.org/10.1186%2Fs13040-021-00244-z). PMC 7863449 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7863449). PMID 33541410 (https://pubmed.ncbi.nlm.nih.gov/33541410).
8. Tharwat A. (August 2018). "Classification assessment methods" (https://doi.org/10.1016%2Fj.aci.2018.08.003). *Applied Computing and Informatics*. doi:10.1016/j.aci.2018.08.003 (https://doi.org/10.1016%2Fj.aci.2018.08.003).
9. Stehman, Stephen V. (1997). "Selecting and interpreting measures of thematic classification accuracy". *Remote Sensing of Environment*. **62** (1): 77–89. Bibcode:1997RSEnv..62...77S (https://ui.adsabs.harvard.edu/abs/1997RSEnv..62...77S). doi:10.1016/S0034-4257(97)00083-7 (https://doi.org/10.1016%2FS0034-4257%2897%2900083-7).
10. Powers, David M. W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (https://www.researchgate.net/publication/228529307). *Journal of Machine Learning Technologies*. **2** (1): 37–63. S2CID 55767944 (https://api.semanticscholar.org/CorpusID:55767944).
11. Piryonesi S. Madeh; El-Diraby Tamer E. (2020-03-01). "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index". *Journal of Infrastructure Systems*. **26** (1): 04019036. doi:10.1061/(ASCE)IS.1943-555X.0000512 (https://doi.org/10.1061%2F%28ASCE%29IS.1943-555X.0000512).
12. Chicco D., Jurman G. (January 2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941312). *BMC Genomics*. **21** (1): 6-1–6-13. doi:10.1186/s12864-019-6413-7 (https://doi.org/10.1186%2Fs12864-019-6413-7). PMC 6941312 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941312). PMID 31898477 (https://pubmed.ncbi.nlm.nih.gov/31898477).

13. Fawcett, Tom (2006). "An Introduction to ROC Analysis" (http://people.inf.elte.hu/kiss/11dwhdm/roc.pdf) (PDF). *Pattern Recognition Letters*. **27** (8): 861–874. doi:10.1016/j.patrec.2005.10.010 (https://doi.org/10.1016%2Fj.patrec.2005.10.010).
14. Piryonesi S. Madeh; El-Diraby Tamer E. (2020-03-01). "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index". *Journal of Infrastructure Systems*. **26** (1): 04019036. doi:10.1061/(ASCE)IS.1943-555X.0000512 (https://doi.org/10.1061%2F%28ASCE%29IS.1943-555X.0000512).
15. Powers, David M. W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (https://www.researchgate.net/publication/228529307). *Journal of Machine Learning Technologies*. **2** (1): 37–63.
16. Ting, Kai Ming (2011). Sammut, Claude; Webb, Geoffrey I. (eds.). *Encyclopedia of machine learning*. Springer. doi:10.1007/978-0-387-30164-8 (https://doi.org/10.1007%2F978-0-387-30164-8). ISBN 978-0-387-30164-8.
17. Brooks, Harold; Brown, Barb; Ebert, Beth; Ferro, Chris; Jolliffe, Ian; Koh, Tieh-Yong; Roebber, Paul; Stephenson, David (2015-01-26). "WWRP/WGNE Joint Working Group on Forecast Verification Research" (https://www.cawcr.gov.au/projects/verification/). *Collaboration for Australian Weather and Climate Research*. World Meteorological Organisation. Retrieved 2019-07-17.
18. Chicco D, Jurman G (January 2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941312). *BMC Genomics*. **21** (1): 6-1–6-13. doi:10.1186/s12864-019-6413-7 (https://doi.org/10.1186%2Fs12864-019-6413-7). PMC 6941312 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941312). PMID 31898477 (https://pubmed.ncbi.nlm.nih.gov/31898477).
19. Chicco D, Toetsch N, Jurman G (February 2021). "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7863449). *BioData Mining*. **14** (13): 1-22. doi:10.1186/s13040-021-00244-z (https://doi.org/10.1186%2Fs13040-021-00244-z). PMC 7863449 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7863449). PMID 33541410 (https://pubmed.ncbi.nlm.nih.gov/33541410).
20. Tharwat A. (August 2018). "Classification assessment methods" (https://doi.org/10.1016%2Fj.aci.2018.08.003). *Applied Computing and Informatics*. doi:10.1016/j.aci.2018.08.003 (https://doi.org/10.1016%2Fj.aci.2018.08.003).
21. Rialland, Annie (August 2005). "Phonological and phonetic aspects of whistled languages". *Phonology*. **22** (2): 237–271. CiteSeerX 10.1.1.484.4384 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.484.4384). doi:10.1017/S0952675705000552 (https://doi.org/10.1017%2FS0952675705000552).