Matthews correlation coefficient

The **Matthews correlation coefficient** (MCC) is used in machine learning as a measure of the quality of binary (two-class) classifications, introduced by biochemist Brian W. Matthews in 1975.[1] Although the MCC is equivalent to Karl Pearson's phi coefficient,[2] which was developed decades earlier, the term MCC is widely used in the field of bioinformatics.

The coefficient takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes.[3] The MCC is in essence a correlation coefficient between the observed and predicted binary classifications; it returns a value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and −1 indicates total disagreement between prediction and observation. The statistic is also known as the phi coefficient. MCC is related to the chi-square statistic for a 2×2 contingency table

$$|\mathrm{MCC}| = \sqrt{\frac{\chi^2}{n}}$$

where *n* is the total number of observations.

While there is no perfect way of describing the confusion matrix of true and false positives and negatives by a single number, the Matthews correlation coefficient is generally regarded as being one of the best such measures.[4] Other measures, such as the proportion of correct predictions (also termed accuracy), are not useful when the two classes are of very different sizes. For example, assigning every object to the larger set achieves a high proportion of correct predictions, but is not generally a useful classification.

The MCC can be calculated directly from the confusion matrix using the formula:

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

In this equation, *TP* is the number of true positives, *TN* the number of true negatives, *FP* the number of false positives and *FN* the number of false negatives. If any of the four sums in the denominator is zero, the denominator can be arbitrarily set to one; this results in a Matthews correlation coefficient of zero, which can be shown to be the correct limiting value.

The MCC can be calculated with the formula:

$$\mathrm{MCC} = \sqrt{PPV \times TPR \times TNR \times NPV} - \sqrt{FDR \times FNR \times FPR \times FOR}$$

using the positive predictive value, the true positive rate, the true negative rate, the negative predictive value, the false discovery rate, the false negative rate, the false positive rate, and the false omission rate.

The original formula as given by Matthews was:[1]

$$N = TN + TP + FN + FP$$
$$S = \frac{TP + FN}{N}$$
$$P = \frac{TP + FP}{N}$$
$$\mathrm{MCC} = \frac{TP/N - S \times P}{\sqrt{PS(1-S)(1-P)}}$$

This is equal to the formula given above. As a correlation coefficient, the Matthews correlation coefficient is the geometric mean of the regression coefficients of the problem and its dual. The component regression coefficients of the Matthews correlation coefficient are Markedness (Δp) and Youden's J statistic (Informedness or Δp').[4][5] Markedness and Informedness correspond to different directions of information flow and generalize Youden's J statistic, the δp statistics and (as their geometric mean) the Matthews Correlation Coefficient to more than two classes.[4]

Some scientists claim the Matthews correlation coefficient to be the most informative single score to establish the quality of a binary classifier prediction in a confusion matrix context.[6]

Contents

Confusion matrix

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 *contingency table* or *confusion matrix*, as follows:

		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$		

Multiclass case

The Matthews correlation coefficient has been generalized to the multiclass case. This generalization was called the R_K statistic (for K different classes) by the author, and defined in terms of a $K \times K$ confusion matrix C [12] .[13]

$$\mathrm{MCC} = \frac{\sum_k \sum_l \sum_m C_{kk} C_{lm} - C_{kl} C_{mk}}{\sqrt{\sum_k (\sum_l C_{kl})(\sum_{k'|k' \neq k} \sum_{l'} C_{k'l'})} \sqrt{\sum_k (\sum_l C_{lk})(\sum_{k'|k' \neq k} \sum_{l'} C_{l'k'})}}$$

When there are more than two labels the MCC will no longer range between -1 and +1. Instead the minimum value will be between -1 and 0 depending on the true distribution. The maximum value is always +1.

Advantages of MCC over accuracy and F1 score

As explained by Davide Chicco in his paper *"Ten quick tips for machine learning in computational biology"* (BioData Mining, 2017) and by Giuseppe Jurman in his paper *"The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation"* (BMC Genomics, 2020), the Matthews correlation coefficient is more informative than F1 score and accuracy in evaluating binary classification problems, because it takes into account the balance ratios of the four confusion matrix categories (true positives, true negatives, false positives, false negatives)[6] [10].

The former article explains, for *Tip 8*:

In order to have an overall understanding of your prediction, you decide to take advantage of common statistical scores, such as accuracy, and F1 score.

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

(Equation 1, accuracy: worst value = 0; best value = 1)

$$\text{F1 score} = \frac{2TP}{2TP + FP + FN}$$

(Equation 2, F1 score: worst value = 0; best value = 1)

However, even if accuracy and F1 score are widely employed in statistics, both can be misleading, since they do not fully consider the size of the four classes of the confusion matrix in their final score computation.

Suppose, for example, you have a very imbalanced validation set made of 100 elements, 95 of which are positive elements, and only 5 are negative elements (as explained in Tip 5). And suppose also you made some mistakes in designing and training your machine learning classifier, and now you have an algorithm which always predicts positive. Imagine that you are not aware of this issue.

By applying your only-positive predictor to your imbalanced validation set, therefore, you obtain values for the confusion matrix categories:

TP = 95, FP = 5; TN = 0, FN = 0.

These values lead to the following performance scores: accuracy = 95%, and F1 score = 97.44%. By reading these over-optimistic scores, then you will be very happy and will think that your machine learning algorithm is doing an excellent job. Obviously, you would be on the wrong track.

On the contrary, to avoid these dangerous misleading illusions, there is another performance score that you can exploit: the Matthews correlation coefficient [40] (MCC).

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

(Equation 3, MCC: worst value = −1; best value = +1).

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
false negative (FN)
eqv. with miss, Type II error

---

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

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

specificity, selectivity or true negative rate (TNR)

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

precision or positive predictive value (PPV)

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

negative predictive value (NPV)

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

miss rate or false negative rate (FNR)

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

fall-out or false positive rate (FPR)

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

false discovery rate (FDR)

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

false omission rate (FOR)

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

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

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

---

accuracy (ACC)

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

balanced accuracy (BA)

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

F1 score
is the harmonic mean of precision and sensitivity

$$\text{F}_1 = 2 \cdot \frac{\text{PPV} \cdot \text{TPR}}{\text{PPV} + \text{TPR}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

Matthews correlation coefficient (MCC)

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

Fowlkes–Mallows index (FM)

By considering the proportion of each class of the confusion matrix in its formula, its score is high only if your classifier is doing well on both the negative and the positive elements.

In the example above, the MCC score would be undefined (since TN and FN would be 0, therefore the denominator of Equation 3 would be 0). By checking this value, instead of accuracy and F1 score, you would then be able to notice that your classifier is going in the wrong direction, and you would become aware that there are issues you ought to solve before proceeding.

Consider this other example. You ran a classification on the same dataset which led to the following values for the confusion matrix categories:

TP = 90, FP = 4; TN = 1, FN = 5.

In this example, the classifier has performed well in classifying positive instances, but was not able to correctly recognize negative data elements. Again, the resulting F1 score and accuracy scores would be extremely high: accuracy = 91%, and F1 score = 95.24%. Similarly to the previous case, if a researcher analyzed only these two score indicators, without considering the MCC, they would wrongly think the algorithm is performing quite well in its task, and would have the illusion of being successful.

On the other hand, checking the Matthews correlation coefficient would be pivotal once again. In this example, the value of the MCC would be 0.14 (Equation 3), indicating that the algorithm is performing similarly to random guessing. Acting as an alarm, the MCC would be able to inform the data mining practitioner that the statistical model is performing poorly.

For these reasons, we strongly encourage to evaluate each test performance through the Matthews correlation coefficient (MCC), instead of the accuracy and the F1 score, for any binary classification problem.

> — Davide Chicco, Ten quick tips for machine learning in computational biology[6]

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

informedness or bookmaker informedness (BM)

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

markedness (MK) or deltaP

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

Sources: Fawcett (2006),[7] *Powers (2011),*[4] *Ting (2011),*[8] *and CAWCR*[9] *Chicco & Jurman (2020)*[10]*. Tharwat (2018)*[11]*.*

Note that the F1 score depends on which class is defined as the positive class. In the first example above, the F1 score is high because the majority class is defined as the positive class. Inverting the positive and negative classes results in the following confusion matrix:

TP = 0, FP = 0; TN = 5, FN = 95

This gives an F1 score = 0%.

The MCC doesn't depend on which class is the positive one, which has the advantage over the F1 score to avoid incorrectly defining the positive class.

See also

- Cohen's kappa
- Cramér's V, a similar measure of association between nominal variables.
- F1 score
- Phi coefficient
- Fowlkes–Mallows index

References

1. Matthews, B. W. (1975). "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". *Biochimica et Biophysica Acta (BBA) - Protein Structure*. **405** (2): 442–451. doi:10.1016/0005-2795(75)90109-9 (https://doi.org/10.1016%2F0005-2795%2875%2990109-9). PMID 1180967 (https://pubmed.ncbi.nlm.nih.gov/1180967).
2. Cramer, H. (1946). *Mathematical Methods of Statistics*. Princeton: Princeton University Press, p. 282 (second paragraph). ISBN 0-691-08004-6
3. Boughorbel, S.B (2017). "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5456046). *PLOS One*. **12** (6): e0177678. Bibcode:2017PLoSO..1277678B (https://ui.adsabs.harvard.edu/abs/2017PLoSO..1277678B). doi:10.1371/journal.pone.0177678 (https://doi.org/10.1371%2Fjournal.pone.0177678). PMC 5456046 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5456046). PMID 28574989 (https://pubmed.ncbi.nlm.nih.gov/28574989).
4. Powers, David M W (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (http://www.flinders.edu.au/science_engineering/fms/School-CSEM/publications/tech_reps-research_artfcts/TRRA_2007.pdf) (PDF). *Journal of Machine Learning Technologies*. **2** (1): 37–63.
5. Perruchet, P.; Peereman, R. (2004). "The exploitation of distributional information in syllable processing". *J. Neurolinguistics*. **17** (2–3): 97–119. doi:10.1016/s0911-6044(03)00059-9 (https://doi.org/10.1016%2Fs0911-6044%2803%2900059-9).
6. Chicco D (December 2017). "Ten quick tips for machine learning in computational biology" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5721660). *BioData Mining*. **10** (35): 35. doi:10.1186/s13040-017-0155-3 (https://doi.org/10.1186%2Fs13040-017-0155-3). PMC 5721660 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5721660). PMID 29234465 (https://pubmed.ncbi.nlm.nih.gov/29234465).
7. 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).
8. Ting, Kai Ming (2011). *Encyclopedia of machine learning* (https://link.springer.com/referencework/10.1007%2F978-0-387-30164-8). Springer. ISBN 978-0-387-30164-8.
9. 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.
10. 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** (6). 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).
11. Tharwat A (August 2018). "Classification assessment methods". *Applied Computing and Informatics*. doi:10.1016/j.aci.2018.08.003 (https://doi.org/10.1016%2Fj.aci.2018.08.003).
12. Gorodkin, Jan (2004). "Comparing two K-category assignments by a K-category correlation coefficient". *Computational Biology and Chemistry*. **28** (5): 367–374. doi:10.1016/j.compbiolchem.2004.09.006 (https://doi.org/10.1016%2Fj.compbiolchem.2004.09.006). PMID 15556477 (https://pubmed.ncbi.nlm.nih.gov/15556477).
13. Gorodkin, Jan. "The Rk Page" (http://rk.kvl.dk/introduction/index.html). *The Rk Page*. Retrieved 28 December 2016.