Extending F1 metric, probabilistic approach

Mikołaj Sitarz sitarz@refaba.com
Refaba
2022 Kraków

Abstract

This article explores the extension of well-known F_1 score used for assessing the performance of binary classifiers. We propose the new metric using probabilistic interpretation of precision, recall, specificity, and negative predictive value. We describe its properties and compare it to common metrics. Then we demonstrate its behavior in edge cases of the confusion matrix. Finally, the properties of the metric are tested on binary classifier trained on the real dataset.

Keywords: machine learning, binary classifier, F_1, MCC, precision, recall

1. Background

The F_1 metric – as described by Sasaki (2007) – is commonly used to evaluate the performance of binary machine learning classifiers. Calculated as a harmonic mean of precision and recall (see section 2) gains an advantage over less complex metrics like accuracy. Especially when used against imbalanced datasets. The properties of F_1 and methods of maximizing the expected metric value, were precisely described, and analyzed from the theoretical and experimental point of view by Lipton et al. (2014).

F_1 is often criticized as an evaluation metric. The main axis of that critique is lack of the dependency on *true negatives* - pointed among the others by Powers (2020) and Hand and Christen (2018). Another of its drawbacks is asymmetry – it may give different score when the dataset labeling is changed (positives labeled as negatives and negatives labeled as positives). These facts make it unreliable as a metric in certain cases.

As a cure for these F_1 problems – MCC is often pointed to – like presented by Chicco and Jurman (2020). On the other hand sometimes, researchers prefer to use them both "cooperating" – like Cao et al. (2020). To this last aspect, we will return later in our article.

2. Common metrics

2.1 Basic and composite metrics

Let us first list the basic building blocks of which the binary classifier metrics are composed:

- TP - true positives (positive samples classified as positive),
- FN - false negatives (positive samples misclassified as negative),

- TN - true negatives (negative samples classified as negative),
- FP - false positives (negative samples misclassified as positive),

Based on them, the *precision* and *recall* (also called *sensitivity*) metrics were defined:

$$\mathrm{PREC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$\mathrm{REC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Then F_1 is defined as a harmonic mean of two above:

$$\mathrm{F}_1 = \frac{2}{\frac{1}{\mathrm{PREC}} + \frac{1}{\mathrm{REC}}} = \frac{\mathrm{TP}}{\mathrm{TP} + \frac{1}{2}(\mathrm{FP} + \mathrm{FN})}$$

We will also use another metric called *specificity*:

$$\mathrm{SPEC} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

Precision and *recall* are more popular in machine learning publications, while the medical ones usually prefer *recall* and *specificity*. The latter two are the base components of *Youden index* (also known as *informedness*) defined by Youden (1950):

$$\mathrm{J} = \mathrm{REC} + \mathrm{SPEC} - 1$$

We also want to include another metric called - *negative predictive value* - NPV in this overview:

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}$$

Finally, NPV is a component of *markedness* (see: Powers (2020)):

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

The presented list obviously is not comprehensive - it does not exhaust all the metrics in use. However, we will focus on them in our discussion.

Summarizing: in the further sections we will be considering the following plain metrics: PREC, REC, SPEC, NPV, and following composite metrics: F_1, J, MK.

2.2 Matthews correlation coefficient

On a separate note, the MCC metric deserves attention. *Matthews correlation coefficient* known under several other forms, in its present form defined by Matthews (1975). MCC is a Pearson correlation coefficient calculated for two binary sequences: the original sample values (positives and negatives) and the values returned by classifier. In terms of TP, FN, TN, FP values it can be calculated as follows:

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

3. Probabilistic approach – focusing on conditional probabilities

The probabilistic interpretation of F_1, PREC and REC has been comprehensively presented by Goutte and Gaussier (2005). We want to attack the problem from the opposite side and start with the definition of 4 conditional probabilities which values we want to maximize when designing a binary classifier:

- $P(+ \mid C+)$ – the probability that the sample is positive, provided the classifier result was positive.
- $P(C+ \mid +)$ – the probability that the classifier result will be positive, provided the sample is positive.
- $P(C- \mid -)$ – the probability that the classifier result will be negative, provided the sample is negative.
- $P(- \mid C-)$ – the probability that the sample is negative, provided the classifier result was negative.

Given all these conditional probabilities, we can require a valid classifier to produce results for which each is close to 1. Basing on this requirement, we are building the new metric P_4, demanding it to have the following properties:

1. The metric value is limited to the given range: $\mathrm{P}_4 \in [0, 1]$.
2. When any of the four conditional probabilities tends to zero, P_4 metric also tends to zero regardless of the values of the other probabilities.
3. When all the four conditional probabilities tend to one, P_4 metric also tends to one.

Now let us quantitatively describe each of these probabilities:

$$P(+ \mid C+) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} = \mathrm{PREC} \tag{1}$$

$$P(C+ \mid +) = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = \mathrm{REC} \tag{2}$$

$$P(C- \mid -) = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} = \mathrm{SPEC} \tag{3}$$

$$P(- \mid C-) = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}} = \mathrm{NPV} \tag{4}$$

Coming back to the composite metrics mentioned before: F_1 captures only probabilities (1) and (2), J is based on (2) and (3), while MK depends on (1) and (4). We have not yet met a metric that refers directly to all four probabilities at once. This fact imposes a certain desire to combine all of them into a single measure. So, let us finally define P_4 as a harmonic mean of all the four conditional probabilities:

$$\mathrm{P}_4 = \frac{4}{\frac{1}{\mathrm{PREC}} + \frac{1}{\mathrm{REC}} + \frac{1}{\mathrm{SPEC}} + \frac{1}{\mathrm{NPV}}}$$

Thus, we get:

$$\mathrm{P}_4 = \frac{4 \cdot \mathrm{TP} \cdot \mathrm{TN}}{4 \cdot \mathrm{TP} \cdot \mathrm{TN} + (\mathrm{TP} + \mathrm{TN}) \cdot (\mathrm{FP} + \mathrm{FN})}$$

The newly defined P_4 metric satisfies all the three requirements we defined above. This is due to the properties of the harmonic mean. What is more, for the requirements 2 and 3 the inverse implication is also true:

4. When P_4 metric tends to zero, at least one of the conditional probabilities is close to zero.

5. When P_4 metric tends to one, all the probabilities are close to one.

P_4 is also symmetrical with respect to dataset labels swapping (similarly to the MCC metric), as opposed to F_1 – see appendix A. In the coming sections, we will take a closer look at the newly defined metric and compare its properties with those of the commonly known metrics.

4. Edge cases

4.1 Confusion matrix

It is essential that when analyzing the performance of a classifier, it should not be considered in isolation from the population to which it applies. The same classifier used on two populations having different sample distributions – will lead to the different performance metric values. Therefore, a convenient way to present the performance of a classifier with respect to a given population is a confusion matrix:

	Actual positive TP+FN	Actual negative FP+TN
Classified positive TP + FP	TP	FP
Classified negative FN + TN	FN	TN

We will use the shorten version of confusion matrix, in our demonstration:

$$\mathbf{C} = \begin{bmatrix} \mathrm{TP} & \mathrm{FP} \\ \mathrm{FN} & \mathrm{TN} \end{bmatrix}$$

To show the properties of P_4 against other metrics, we will present four examples of confusion matrices. For each of them, one of the conditional probabilities (1), (2), (3), (4) is close to 0, while the others are moderately close to 1. In each of the cases presented, we use a simulated classifier and a population of 10000 samples.

Because some of the metrics presented above, are ranged in $[-1, 1]$ than in $[0, 1]$, we must scale them first, to compare with the other ones. Thus, we will be using:

$$\mathrm{MCC}' = (\mathrm{MCC} + 1)/2$$

$$\mathrm{J}' = (\mathrm{J} + 1)/2$$

$$\mathrm{MK}' = (\mathrm{MK} + 1)/2$$

4.2 Case 1 - "alarming precision"

This is a classic case in which F_1 shines. The population is highly imbalanced in favor of negative samples. The classifier's performance on positive samples is 90%, the same on negative ones. Thus, we have the following confusion matrix:

$$\mathbf{C_1} = \begin{bmatrix} 45 & 995 \\ 5 & 8955 \end{bmatrix}$$

In this case, the four conditional probabilities are as follows:

Conditional Probability	Value
$P(+ \mid C+)$	0.0433
$P(C+ \mid +)$	0.9000
$P(C- \mid -)$	0.9000
$P(- \mid C-)$	0.9994

And now let us look at how these affect the values of our metrics:

Metric	Value
P_4	0.1519
F_1	0.0826
MCC'	0.5924
J'	0.9000
MK'	0.5214

We can identify three groups of metrics in the table above:

- "Close to zero" (yellow) group – having two members: P_4 and F_1.
- "Middle of the range" (white) group – MCC' and MK' – still reacting OK on the situation.
- "Ignoring" (red) group – metric J' – no proper reaction on the low conditional probability: $P(+ \mid C+)$.

4.3 Case 2 - "alarming negative predictive value"

This case can be simply obtained from "Case 1" by re-labeling the samples - naming the "positives" as "negatives" and vice versa. Thus, we have the following confusion matrix:

$$\mathbf{C_2} = \begin{bmatrix} 8955 & 5 \\ 995 & 45 \end{bmatrix}$$

Probabilities:

Conditional Probability	Value
$P(+ \mid C+)$	0.9994
$P(C+ \mid +)$	0.9000
$P(C- \mid -)$	0.9000
$P(- \mid C-)$	0.0433

That gives the following metric values:

Metric	Value
P_4	0.1519
F_1	0.9471
MCC′	0.5924
J′	0.9000
MK′	0.5214

As we can see, F_1 is the only metric that changed its value after the label swap. This clearly shows the problem with its asymmetry. Contrasting to the previous case, F_1 is completely not noticing one of the key probabilities being close to zero. The other metrics considered have not changed compared to the "Case 1".

4.4 Case 3 - "alarming recall"

This case represents a typical situation when the classifier is over-predicting in favor of negative results, resulting a particularly superior performance on the negative samples and deficient performance on the positive samples. The population contains 10% positive samples. So, there we have the confusion matrix:

$$\mathbf{C_3} = \begin{bmatrix} 50 & 9 \\ 950 & 8991 \end{bmatrix}$$

Conditional probabilities:

Conditional Probability	Value
$P(+ \mid C+)$	0.8475
$P(C+ \mid +)$	0.0500
$P(C- \mid -)$	0.9990
$P(- \mid C-)$	0.9044

Values of the metrics:

Metric	Value
P_4	0.1718
F_1	0.0944
MCC'	0.5960
J'	0.5245
MK'	0.8759

As we see F_1 is back in the league. MCC' and J' playing well. And a red card is given to a player of the visiting team: MK'.

4.5 Case 4 - "alarming specificity"

The last case represents the inversion of "Case 3". The classifier over-predicts in favor of positive results – having particularly superior performance on positive results and deficient performance on the negative ones. The population contains 10% negative samples. So, let us look at the confusion matrix, probabilities, and the metrics:

$$\mathbf{C_4} = \begin{bmatrix} 8991 & 950 \\ 9 & 50 \end{bmatrix}$$

Conditional Probability	Value
$P(+ \mid C+)$	0.9044
$P(C+ \mid +)$	0.9990
$P(C- \mid -)$	0.0500
$P(- \mid C-)$	0.8475

Metric	Value
P_4	0.1718
F_1	0.9494
MCC'	0.5960
J'	0.5245
MK'	0.8759

MK' again occupies the "red" group, this time together with F_1 as a companion. P_4 obviously – works as designed.

4.6 Summary

As we have seen in the four edge cases: none of the considered, existing so far compound metrics (F_1, J, MK), guarantees correct behavior in all of them. MCC as a correlation coefficient here represents a separate category and stands out positively against its background. And even though it does not reach values near its minimum in edge cases, its performance should still be considered satisfactory.

What is not surprising, however, is the behavior of the newly defined P_4 metric itself – it reaches a correspondingly low value every time, and this is due to the very assumptions on which it was based.

5. P_4 compared to other metrics

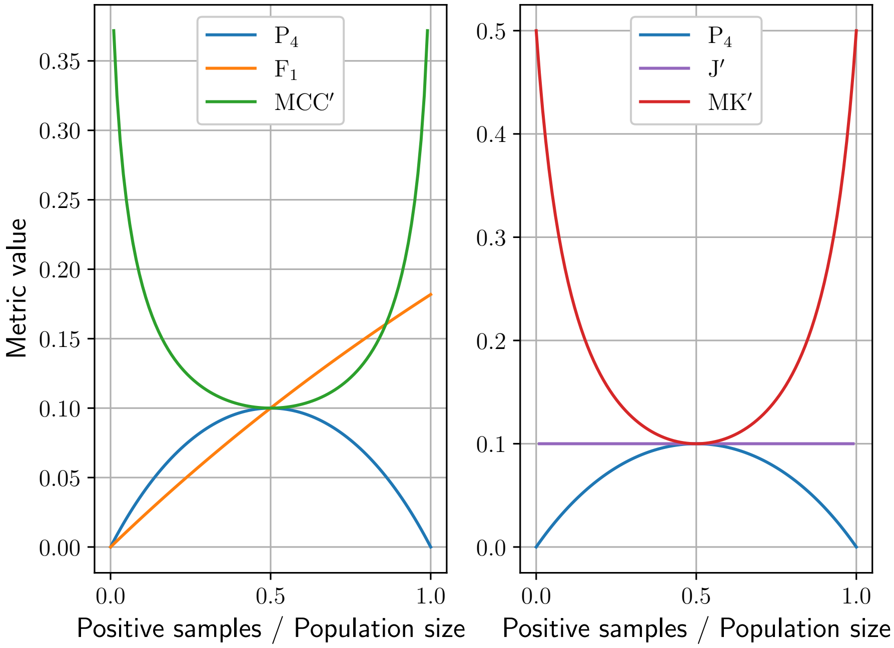
In the following subsections we will again compare P_4 against four traditional metrics: MCC, F_1, J and MK. This time, covering quasi-continuous range of cases. As in previous cases, we use a simulated classifier and a fixed population size of 10000.

5.1 Metrics vs population balance

In the experiment we fixed the following parameters:

- The ratio of *true positives* to the *actual positives* (TPR – *true positive rate*) is fixed and equals 0.1.
- The ratio of *true negatives* to the *actual negatives* (TNR – *true negative rate*) is fixed and equals 0.1.

Thus, using our rather poor "classifier", we are observing how the values of each metric change as a function of: *actual positives* to population size ratio. The result can be seen on the charts below:



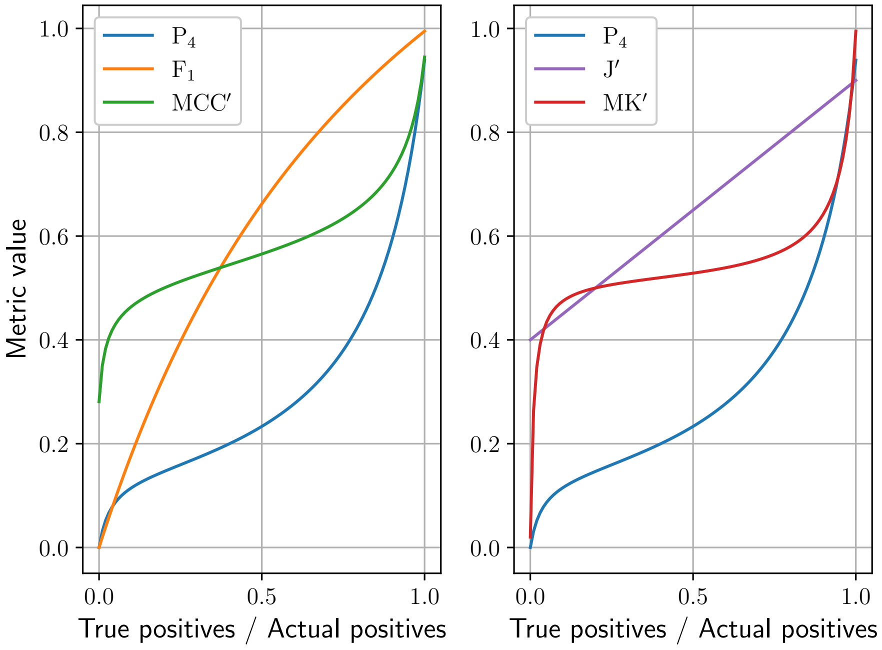
On the left chart, attention is drawn to the symmetrical shape determined by MCC′ and P_4 curves, while F_1 follows its own path. On the right chart we have a similar symmetrical shape as in the previous one, but this time P_4 is accompanied by MK'. Youden index is not sensitive to the population balance change. The example presented here is distant from the results obtained from the classifiers encountered on a daily basis but has the advantage of capturing differences between the metrics studied.

5.2 Metrics vs true positive rate

Let us see a bit more realistic example now. The following parameters are fixed now:

- The ratio of *true negatives* to the *actual negatives* is fixed and equals 0.8.
- The ratio of *actual positives* to the population size is also fixed and equals 0.95.

Then we are observing how the metric changes as a function of TPR (*true positive rate*).



In this case, the course of P_4 and MCC' metrics is similar in nature, however for the first half of TPR's range the difference between them is roughly ≈ 0.3. They reach full agreement at the end of the plot. At this point the difference between them and F_1 equals 0.05 – which is *actual negatives* to the population size ratio.

5.3 Summary

The charts presented here do not exhaust all the relationships between the metrics being discussed. Many of the aspects are left undiscussed due to the short form of this paper. We are also not analyzing the base building blocks of the composite metrics: PREC, REC, SPEC, NPV, because the results they give are simpler than the presented ones. The analysis of *accuracy* is skipped because of the same reason. The following conclusions can be drawn from the charts:

- P_4 and MCC behave differently in extreme case, however they tend to behave similarly in the more real-world case.
- F_1 is oversimplified compared with P_4 and MCC.
- J' behaves linear in presented cases

6. P_4 metric in use

To demonstrate the properties and usefulness of P_4, we will check how it behaves on a real dataset. To achieve this goal the technique derived from the "Receiver operating characteristic" method will be used.

We will use well-known Breast Cancer Wisconsin dataset provided by UCI Machine Learning Repository (http://archive.ics.uci.edu/ml), with a help of Scikit-Learn package – Pedregosa et al. (2011). It's a set consisting of 569 samples, 30 dimensions. The samples contain various characteristics of biological cell nuclei (radius, texture, symmetry etc.) and the cancer binary classification: malignant/benign.

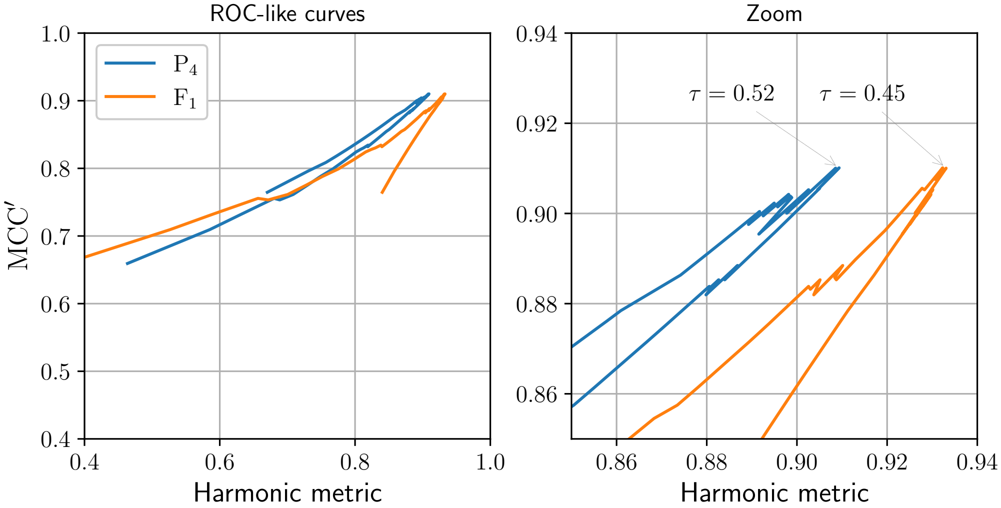
6.1 Receiver operating characteristic

"Receiver operating characteristic" (ROC) - is commonly used technique for assessing the trade-off between *recall* and *specificity*. It's used together with the classifiers that, as a result – give a probability of being positive for the sample – like for example logistic regression. To obtain an answer "positive"/"negative", we must decide on a specific probability threshold τ above which we consider the sample positive. Using ROC technique – we are starting from $\tau_0 = 0$ and iterating, increasing it by $\Delta\tau$ until $\tau_n = 1$ is reached. For each τ_i we calculate confusion matrix $\mathbf{C_i}$ and thus the *precision-recall* pair. Then we plot the curve on the REC vs SPEC chart. That plot gives us an insight into the characteristic of the classifier-dataset pair and allows choosing optimal τ threshold.

This method has been creatively adapted by Cao et al. (2020). Instead of *precision-recall* pair, MCC-F_1 pair has been used in their case, allowing more unambiguous result and easier selection of the optimum.

6.2 MCC-F_1 and MCC-P_4 curves

We will use the same technique as mentioned above but also including the P_4 metric in place of F_1 – comparing two curves: MCC-F_1 and MCC-P_4. We chose the *Support Vector Machine* classifier with probabilistic output (see Cortes and Vapnik (1995), Platt et al. (1999)) and the Gaussian kernel. The result – two ROC-like curves – can be seen in the chart below:



The resulting graphs are similar in both cases, however for P_4 the curve has a smaller opening angle. The critical section is captured in the close-up on the graph to the right. The curves give different optimal threshold values: $\tau = 0.52$ for MCC-P_4 curve and $\tau = 0.45$ for MCC-F_1. From the perspective of this single experiment – the behavior of P_4 metric is as expected. The difference between the two results, is since P_4 includes two additional components comparing to F_1 – conditional probabilities: $P(C- \mid -)$ and $P(- \mid C-)$.

7. Conclusions

The definition of the new P_4 metric presented, broadens the range of available tools for evaluating binary classifiers. It represents one step further in the direction indicated by F_1. The main advantages of P_4 are that it zeroes out when at least one of the key four conditional probabilities also zeroes out, and that reaching a value close to 1 requires that all four probabilities also go to 1.

We realize that evaluating the performance of binary classifiers is a complex problem, and we cannot expect a single metric to be the ultimate gold standard here. Some situations may require that selected conditional probabilities be considered more significant than others. And this, in turn, will require the development of weights like those known from F_β.

The key differences between P_4 and MCC are a different probabilistic interpretation and a guarantee that P_4 will zero out under certain conditions. Finally, their values belong to other ranges: $\mathrm{MCC} \in [-1, 1]$ and $\mathrm{P}_4 \in [0, 1]$. The last one may be perceived as a little easier to interpret. Despite these facts, P_4 appears to be much closer to MCC than the other composite metrics. In a situation when one uses the F_1 however, we can frankly recommend its replacement with the P_4.

8. Acknowledgments

We would like to thank the founders of *SciHub* web service. Without its help, the creation of this article would not have been possible.

A. Symmetry

In this appendix, we will prove the symmetry of P_4 with respect to dataset labels swapping. By label swapping we mean renaming labels from positives to negatives and vice versa.

1. P_4 is defined as the harmonic mean of PREC, REC, SPEC and NPV.
2. Harmonic mean is a commutative operation.
3. Dataset label swapping causes the following changes to the confusion matrix:
   (a) TP becomes TN
   (b) TN becomes TP
   (c) FP becomes FN
   (d) FN becomes FP
4. After this changes to the confusion matrix: PREC becomes NPV, NPV becomes PREC (see definitions in section 2).
5. Similarly, REC becomes SPEC, SPEC becomes REC
6. Swapping the order of the arguments of the harmonic mean does not change its value (see point 2) – what ends the proof.

References

Chang Cao, Davide Chicco, and Michael M Hoffman. The mcc-f1 curve: a performance evaluation technique for binary classification. *arXiv preprint arXiv:2006.11278*, 2020.

Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3): 273–297, 1995.

Cyril Goutte and Eric Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval*, pages 345–359. Springer, 2005.

David Hand and Peter Christen. A note on using the f-measure for evaluating record linkage algorithms. *Statistics and Computing*, 28(3):539–547, 2018.

Zachary Chase Lipton, Charles Elkan, and Balakrishnan Narayanaswamy. Thresholding classifiers to maximize f1 score, 2014. URL https://arxiv.org/abs/1402.1892.

B.W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2): 442–451, 1975. ISSN 0005-2795. doi: https://doi.org/10.1016/0005-2795(75)90109-9. URL https://www.sciencedirect.com/science/article/pii/0005279575901099.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

David M. W. Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. 2020. doi: 10.48550/ARXIV.2010.16061. URL https://arxiv.org/abs/2010.16061.

Yutaka Sasaki. The truth of the f-measure. *Teach Tutor Mater*, 01 2007.

William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.