Empirical Evaluation of Predictive Models: A keynote at ECIR 2022

Peter Flach
University of Bristol
United Kingdom
Peter.Flach@bristol.ac.uk

Abstract

I give a brief overview of my recent keynote at the 2022 European Conference on Information Retrieval that was held in Stavanger, Norway. I pay particular attention to some basic questions involving the F-score that appear to lead to confusion. I also settle a question raised at the conference by reconstructing an account from Van Rijsbergen's classic text *Information Retrieval*.

1 Background

Evaluation of predictive models is of primary importance in machine learning and information retrieval. Empirical evaluation is an act of measurement, and as such very common but perhaps not as straightforward as one might expect. For example, it is almost always the case that there is a discrepancy between what is of interest (e.g., mathematical ability of a student, or population performance of a model) and what is directly measurable (performance of the student or model on a specific set of questions or labelled data points, with many contextual aspects influencing performance). One would also often like to have a causal account of what is observed: e.g., an explanation why one student or algorithm outperforms another, possibly in counterfactual form (if the test were manipulated in a particular way, the performance difference would disappear).

Even more fundamental issues arise when one considers measurement scales and how to combine different quantities into aggregate measurements. We are all familiar with this: e.g., classification accuracy can be seen as a weighted arithmetic mean of true positive and negative rates, with class prevalences as weights; the F-score is commonly defined as the harmonic mean of precision and recall; and some people prefer the geometric mean to aggregate precision and recall. Choosing a different mean amounts to a change of scale (e.g., the log of the geometric mean is the arithmetic mean of the logs) and as such admissible, even if the change of scale requires justification. But mixing different means (and hence scales) can easily lead to incoherence: taking expectations involves the arithmetic mean, which implies for example that the area under the precision-recall curve bears no direct relationship to an aggregate F-score (while the area under the ROC curve does relate to aggregate accuracy by traversing the operating points on the curve in a particular way [Hernández-Orallo et al., 2012]).

2 The Keynote

I started my talk with a general introduction[1] which involved audience participation on the following questions:

Q1. Why is F_1 the harmonic mean of precision and recall?

1. It's a choice, it could equally well have been an arithmetic or geometric mean.
2. It corresponds to averaging the mistakes a classifier makes.
3. Another reason.

Q2. When is F_1 preferred over accuracy-based measures (micro/macro-accuracy)?

1. When we have many more negatives than positives.
2. When true negatives don't add value.
3. Another reason.

Q3. If we use $F_{1/2}$, then...

1. Precision gets twice the weight of recall.
2. Precision gets four times the weight of recall.
3. Neither.

Interestingly, none of these questions received a very clear majority answer from the audience. My own answers are the second alternative in each case.

For the first question, this is most easily seen by rearranging terms from the usual harmonic mean definition to obtain

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

which is my preferred definition of F_1 as it clearly shows how the two kinds of mistakes (false positives and false negatives) are arithmetically averaged. Using the geometric mean – preferred by some – corresponds employing logarithmic scales for precision and recall, since the log of the geometric mean is the arithmetic mean of the logs. This emphasises smaller values, but would need to be justified.

It is well-known that instance-averaged accuracy suffers from over-emphasising the majority class for highly imbalanced data. However, this can be remedied to switching to macro-accuracy, which is an unweighted arithmetic mean of per-class recall. So class imbalance by itself is insufficient justification for using F-score. As I showed in the talk, F-score can be seen as accuracy on a modified confusion matrix, replacing the number of true negatives with the number of true positives (or 0, if one wanted to obtain the Jaccard index instead). In this way we don't downplay the false positives (mistakes on the negative class), but only the true negatives (as Google doesn't make money on not returning a web page I'm not interested in).

[1] The slides I used at the keynote are viewable within a web browser at https://flach.github.io/slides/2022_ecir/.

The third question hinges on whether one uses β^2 or β in the definition of F_β, where a value of β different from 1 allows one to put more emphasis on either precision or recall. Van Rijsbergen's original definition uses $\frac{1}{\beta^2+1}$ and $\frac{\beta^2}{\beta^2+1}$ for the weights of precision and recall, respectively, in the weighted harmonic mean, the ratio of which is $1/\beta^2$ or 4 for $\beta = 1/2$. The use of the square here has always puzzled me – I have thought a bit more about this since and will discuss further in the next section.

The talk then continued by giving a short overview of ROC curves, a topic I have researched for over 20 years. Among the many advantages of ROC curves are the following:

linear interpolation: any point on a straight line between thresholds (or classifiers) A and B can be achieved by making a suitably biased random choice between them, leading to the ROC convex hull (ROCCH);

area under curve: AUROC estimates the probability that a randomly selected positive is ranked before a randomly selected negative, and is moreover linearly related to expected classification performance if thresholds are set to make a particular proportion of positive predictions [Hernández-Orallo et al., 2012];

calibration: slopes of ROCCH segments are empirical likelihood ratios associated with intervals of classifier scores, and can be used to obtain calibrated probabilities (isotonic regression); in particular, if a perfectly calibrated classifier assigns score c to an instance, then it is on the decision boundary for $acc_c = 2c\pi tpr + 2c(1-\pi)fpr$, where π is the prevalence of positives and $acc_{1/2}$ is standard accuracy [Silva Filho et al., 2021].

ROC curves use two of the three degrees of freedom in a normalised confusion matrix (true and false positive rates), with the remaining one (class ratio) fixing the slope of iso-accuracy lines. Instead of false positive rate we could use precision, so there is clearly a point-to-point correspondence between ROC curves and precision-recall curves. However, PR curves don't have any of the nice properties of ROC curves: for example, interpolation is not linear, and AUPR is not a coherent measure of aggregated performance.

Following Flach and Kull [2015], I went into some detail in my talk how this can be overcome by using non-linearly scaled versions of precision, recall and F-score. The key idea is to first take reciprocals (e.g., $prec = TP/(TP+FP)$ becomes $1/prec = 1 + FP/TP$); clip the latter to the $[1, 1/\pi]$ interval to exclude overly small values of precision and recall; and linearly map this back to the unit interval. The resulting measures quantify the *gain* over the baseline always-positive classifier (e.g., $precG = 1 - \frac{\pi}{1-\pi}FP/TP$). F-gain is now simply the arithmetic mean of precision gain and recall gain, and the resulting Precision-Recall-Gain curves are almost entirely "ROC-like". In particular, AUPRG can be interpreted as an aggregate F1-score, similar to how AUROC can be interpreted as an aggregate accuracy.

After this I briefly discussed some new results that relate AUPRG to a *weighted* ranking score, potentially providing a well-founded alternative to measures such as normalised discounted cumulative gain. I ended the keynote talking about the need for more sophisticated measurement models [Flach, 2019] that include latent variables [Chen et al., 2019; Song and Flach, 2021] and allow causal explanations of performance differences.

3 After the Conference

At the Q&A following my talk, David Lewis asked me whether my proposed measures satisfied this property ascribed to F-score by Keith Van Rijsbergen:"[β] measures the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall as precision". I knew the quote but had never properly understood it, so couldn't answer the question then. After the conference I looked it up in in Chapter 7 of the online version of Van Rijsbergen's book[2] and found that the text continues as follows: "The simplest way I know of quantifying this is to specify the P/R ratio at which the user is willing to trade an increment in precision for an equal loss in recall. *Definition 6.* The relative importance a user attaches to precision and recall is the P/R ratio at which $\partial E/\partial R = \partial E/\partial P$, where $E = E(P, R)$ is the measure of effectiveness based on precision and recall."

My elaboration of this goes as follows. Take a weighted harmonic mean of precision and recall as effectiveness measure:

$$E = \frac{1}{\alpha/P + (1-\alpha)/R}$$

The partial derivatives with respect to P and R can be written as follows:

$$\frac{\partial E}{\partial P} = E^2\frac{\alpha}{P^2} \qquad \frac{\partial E}{\partial R} = E^2\frac{1-\alpha}{R^2}$$

Setting these equal gives

$$P/R = \sqrt{\frac{\alpha}{1-\alpha}} = 1/\beta$$

So if $\beta = 1/2$, as in my Q3, then precision is **twice as high as recall** – at the particular point where "the user is willing to trade an increment in precision for an equal loss in recall" (equal partial derivatives). On the other hand, re-expressing the weights used in the harmonic mean we obtain $\alpha = 1/(\beta^2+1)$ and $1-\alpha = \beta^2/(\beta^2+1)$, and hence precision has **four times the weight** of recall. While there is no contradiction between these two interpretations of β, I would argue that the latter is considerably more transparent. I would furthermore suggest that $\alpha \in [0, 1]$ is more interpretable as a parameter.

So what about David Lewis' original question? We define FG_β as a weighted arithmetic mean of precision gain and recall gain, parametrised in the same way as F_β. The partial derivatives are hence α and $1-\alpha$, which are only equal for $\alpha = 1/2$ or $\beta = 1$. As this is independent of the values of precision/recall gain, we don't derive a similar ratio as in Van Rijsbergen's account – but the transparent interpretation of α as a weight in the mean carries over all the same.

Acknowledgments

Part of this work was funded through a project with the Alan Turing Institute.[3] The ideas in this paper were developed in collaboration with Su Whan Baek, Yu Chen, Tom Diethe, Cèsar Ferri, José Hernández-Orallo, Meelis Kull, Miquel Perello-Nieto, Ricardo Prudencio, Raúl Santos-Rodriguez, Telmo Silva Filho, Hao Song, and many others.

[2] http://www.dcs.gla.ac.uk/Keith/Chapter.7/Ch.7.html

[3] https://www.turing.ac.uk/research/research-projects/measurement-theory-data-science-and-ai

References

Yu Chen, Telmo Silva Filho, Ricardo Prudencio, Tom Diethe, and Peter Flach. β^3-IRT: A new item response model and its applications. In *22nd International Conference on Artificial Intelligence and Statistics*, pages 1013–1021, 2019. URL https://proceedings.mlr.press/v89/chen19b.html.

Peter Flach. Performance evaluation in machine learning: the good, the bad, the ugly, and the way forward. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9808–9814, 2019. URL https://ojs.aaai.org//index.php/AAAI/article/view/5055.

Peter Flach and Meelis Kull. Precision-recall-gain curves: PR analysis done right. In *Advances in Neural Information Processing Systems*, pages 838–846, 2015. URL http://people.cs.bris.ac.uk/~flach/PRGcurves/.

José Hernández-Orallo, Peter Flach, and Cèsar Ferri. A unified view of performance metrics: translating threshold choice into expected classification loss. *Journal of Machine Learning Research*, 13:2813–2869, 2012. URL https://www.jmlr.org/papers/v13/hernandez-orallo12a.html.

Telmo Silva Filho, Hao Song, Miquel Perello-Nieto, Raul Santos-Rodriguez, Meelis Kull, and Peter Flach. Classifier calibration: How to assess and improve predicted class probabilities: a survey. *arXiv preprint arXiv:2112.10327*, 2021. URL https://arxiv.org/abs/2112.10327.

Hao Song and Peter Flach. Efficient and robust model benchmarks with item response theory and adaptive testing. *International Journal of Interactive Multimedia & Artificial Intelligence*, 6(5), 2021. URL https://www.ijimai.org/journal/bibcite/reference/2901.