ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

On F_β-score and Cost-Consistency in Evaluation of Imbalanced Classification

Aleksi Avela*

Aalto University, School of Science
Department of Mathematics and Systems Analysis
P.O. Box 11100, FI-00076 Aalto, Finland

Abstract. Among many other difficulties of imbalanced classification, evaluation of classifiers is rarely trivial. F_β-score is often recommended as one of the go-to evaluation measures in imbalanced classification, but researchers have voiced their concerns on whether F_β-score in fact is an appropriate measure. In this paper, we introduce a framework of cost-consistency, i.e., whether an evaluation measure is consistent with total classification cost at least for some cost and class imbalance ratio, and show that, with a simple cost structure, F_β-score is not cost-consistent.

1 Introduction

In real-life classification tasks, class and classification cost distributions are rarely balanced – known as the problem of imbalanced data. It is not only learning from imbalanced data that is challenging, but even just evaluating different classifiers on an imbalanced classification task is not trivial. It is well known that accuracy is not an appropriate evaluation measure for imbalanced classification [1], but choosing the most suitable measure is far from easy.

The motivation for this work is that, although conventional evaluation measures are usually defined independent of classification costs, that does not change the fact that in real-life classification tasks there are factual costs – even if they would be (partly) unknown. Thus, it is important to understand how the applied measures behave in relation to total classification cost.

The rest of the paper is organized as follows. In Section 2, we discuss total cost and some conventional evaluation measures. In Section 3, we introduce a framework of cost-consistency, i.e., whether an evaluation measure is consistent with total cost, and evaluate two measures, F_β-score and informedness, under this framework. Finally, Section 4 concludes the work.

2 Evaluation and total cost of imbalanced classification

In this paper, we focus on binary classification. The performance of a binary classifier on a data sample can be summarized by a confusion matrix as shown in Table 1, where tp is the number of true positives, fp the number of false positives, fn the number of false negatives, tn the number of true negatives, and N is the total number of observations in the sample. On the marginals, pp =

*The author gratefully acknowledges support from The Emil Aaltonen Foundation (Nuoren tutkijan apuraha, 230014) and valuable feedback from Pauliina Ilmonen and Matias Heikkilä.

tp + fp and pn = tn + fn are the number of predicted positives and negatives, respectively, and ap = tp + fn and an = tn + fp are the number of actual positives and negatives in the sample, respectively.

Table 1: Binary confusion matrix.

Actual:		**P**	**N**	
Predicted:	**P**	tp	fp	pp
	N	fn	tn	pn
		ap	an	N

In imbalanced classification, the so-called minority class is commonly denoted as the positive class, i.e., ap < an (or even ap $\ll$ an). Moreover, the cost of misclassifying a positive observation negative, C_{FN}, is typically (much) higher than the cost of misclassifying a negative observation positive, C_{FP}. A common issue with imbalanced data is that standard classifiers tend to learn a classification rule with a trivially high accuracy while neglecting the positive class [1].

If the costs are explicitly known, total cost is a rational choice for an evaluation measure. However, in practice, costs are rarely (fully) known and challenging to estimate and can also be observation-dependent or nonlinear [2]. Assuming a simple (i.e., observation-independent and linear) cost structure, any *reasonable* cost system $[C_{\mathrm{FN}}, C_{\mathrm{FP}}, C_{\mathrm{TP}} < C_{\mathrm{FN}}, C_{\mathrm{TN}} < C_{\mathrm{FP}}]$ (see, [3]) can be expressed as a scaled total cost $c \times \mathrm{fn} + \mathrm{fp}$, where $c = \frac{C_{\mathrm{FN}} - C_{\mathrm{TP}}}{C_{\mathrm{FP}} - C_{\mathrm{TN}}}$.

It is easy to show that the actual total cost can be replaced with this simpler scaled cost in evaluation as long as the data sample is fixed for all the evaluated classifiers. Elkan (2001) showed that any reasonable cost system can be transformed into $[C'_{\mathrm{FN}} = \frac{C_{\mathrm{FN}} - C_{\mathrm{TN}}}{C_{\mathrm{FP}} - C_{\mathrm{TN}}}, C'_{\mathrm{FP}} = 1, C'_{\mathrm{TP}} = \frac{C_{\mathrm{TP}} - C_{\mathrm{TN}}}{C_{\mathrm{FP}} - C_{\mathrm{TN}}}, C'_{\mathrm{TN}} = 0]$. The simple scaled cost we defined is actually the difference between this transformed total cost and the transformed total cost of a perfect classification, $C'_{\mathrm{TP}}\mathrm{ap} + C'_{\mathrm{TN}}\mathrm{an} = C'_{\mathrm{TP}}\mathrm{ap} = C'_{\mathrm{TP}}(\mathrm{tp} + \mathrm{fn})$, for a given sample:

$$C'_{\mathrm{FN}}\mathrm{fn} + C'_{\mathrm{TP}}\mathrm{tp} + 1 \times \mathrm{fp} + 0 \times \mathrm{tn} - C'_{\mathrm{TP}}(\mathrm{tp} + \mathrm{fn}) = (C'_{\mathrm{FN}} - C'_{\mathrm{TP}})\mathrm{fn} + \mathrm{fp}$$
$$= \big(\tfrac{C_{\mathrm{FN}} - C_{\mathrm{TN}}}{C_{\mathrm{FP}} - C_{\mathrm{TN}}} - \tfrac{C_{\mathrm{TP}} - C_{\mathrm{TN}}}{C_{\mathrm{FP}} - C_{\mathrm{TN}}}\big)\mathrm{fn} + \mathrm{fp} = \tfrac{C_{\mathrm{FN}} - C_{\mathrm{TP}}}{C_{\mathrm{FP}} - C_{\mathrm{TN}}}\mathrm{fn} + \mathrm{fp} = c \times \mathrm{fn} + \mathrm{fp},$$

which is a strictly increasing function of total cost (given a fixed sample).

In practice, confusion matrices are often compressed into cost-independent univariate evaluation measures. These include, for instance, true positive rate (tpr, also known as recall) $\frac{\mathrm{tp}}{\mathrm{ap}}$, true negative rate (tnr) $\frac{\mathrm{tn}}{\mathrm{an}}$, and precision $\frac{\mathrm{tp}}{\mathrm{pp}}$. On their own, however, these measures are one-sided, as perfect tpr (or tnr) can be obtained trivially by classifying every observation positive (or negative). While perfect precision cannot be achieved trivially, generally the case is that the less observations are classified positive, the higher the value of precision. In order to consider these trade-offs, many commonly applied evaluation measures are defined as combinations of the ones discussed above. Two such measures, F_β-score and informedness, are considered in the following section.

3 Cost-consistency of F_β-score and informedness

F_β-score (or sometimes just F-score or F-measure) has its origins in information retrieval but has since become highly popular in machine learning as well and is often recommended as the go-to evaluation measure in imbalanced classification (typically, with $\beta \in \{\frac{1}{2}, 1, 2\}$). F_β-score is defined as a β-weighted harmonic mean of precision and recall:

$$F_\beta = (1+\beta^2)\frac{\text{precision} \times \text{recall}}{(\beta^2 \times \text{precision}) + \text{recall}} = \frac{(1+\beta^2)\text{tp}}{(1+\beta^2)\text{tp} + \beta^2\text{fn} + \text{fp}}.$$

However, studies have questioned F_β-score's position as one of the conventional evaluation measures in (imbalanced) classification due to, for instance, its inability to consider true negatives [4, 5, 6]. Another notable issue with F_β-score is that, although it is defined as a harmonic mean, it can actually be reduced into a weighted arithmetic mean of recall and precision, where the relative weight of recall and precision depend on the number of predicted positives, pp [4, 6]. The problem is that pp depends on the classifier itself, i.e., the thing that is supposed to be evaluated [4, 6]. The arithmetic mean reformulation [4, 6] is

$$F_\beta = p_\beta \times \text{recall} + (1-p_\beta) \times \text{precision, where } \begin{cases} p_\beta = \frac{\beta^2\text{ap}}{\beta^2\text{ap}+\text{pp}} \\ 1-p_\beta = \frac{\text{pp}}{\beta^2\text{ap}+\text{pp}} \end{cases},$$

and the relative weight of recall compared to precision is $\frac{p_\beta}{1-p_\beta} = \beta^2\frac{\text{ap}}{\text{pp}}$.

Assuming a simple cost structure as discussed in Section 2, (scaled) total cost is given as $c \times \text{fn} + \text{fp}$. That is, for a change Δfn = -1, a cost-invariant change in a confusion matrix is $\Delta\text{fp} = c$, and, consequently, $\Delta\text{tp} = 1$ and $\Delta\text{tn} = -c$. We call an evaluation measure cost-consistent if the value of the measure increases if and only if total cost decreases. A necessary condition for cost-consistency is that, for any cost-invariant change in a confusion matrix, the value of the evaluation measure must also remain unchanged.

Informedness (see, e.g., [5]), which is defined as tpr + tnr - 1, is cost-consistent in a certain situation. A cost-invariant change induces a change of the order of $\frac{1}{\text{ap}}$ in tpr and a change of the order of $-\frac{c}{\text{an}}$ in tnr. These changes cancel each other out if $c = \frac{an}{ap}$. That is, if the cost ratio matches the class imbalance ratio $\frac{an}{ap}$, informedness is cost-consistent[1] (as, in this case, the evaluation measure also necessarily increases if total cost decreases).

On the other hand, there does not exist any scenario (i.e., a cost ratio, a class imbalance ratio, and/or a value of β) in which F_β-score would be cost-consistent. Given a cost-invariant change in a confusion matrix for a unit change in tp, that is, Δtp = -Δfn = 1 and Δtn = -Δfp = $-c$, we can solve the value of β that corresponds to ΔF_β-score = 0 using the arithmetic mean reformulation of F_β

[1] Note that there exists also a measure called balanced accuracy to which the same result applies. Balanced accuracy is defined as $\frac{1}{2}(\text{tpr} + \text{tnr})$. It is essentially the same measure as informedness – just scaled linearly to a range from 0 to 1.

[4, 6]. The changes in recall $\Delta\mathrm{R}$, precision $\Delta\mathrm{P}$, and weight of recall Δp_β are

$$\Delta\mathrm{R} = \frac{\mathrm{tp}+1}{\mathrm{ap}} - \frac{\mathrm{tp}}{\mathrm{ap}} = \frac{1}{\mathrm{ap}}, \qquad \Delta\mathrm{P} = \frac{\mathrm{tp}+1}{\mathrm{pp}+c+1} - \frac{\mathrm{tp}}{\mathrm{pp}} = \frac{\mathrm{pp}-(c+1)\mathrm{tp}}{\mathrm{pp}(\mathrm{pp}+c+1)},$$

$$\Delta p_\beta = \frac{\beta^2\mathrm{ap}}{\beta^2\mathrm{ap}+\mathrm{pp}+c+1} - \frac{\beta^2\mathrm{ap}}{\beta^2\mathrm{ap}+\mathrm{pp}} = \frac{-\beta^2\mathrm{ap}(c+1)}{(\beta^2\mathrm{ap}+\mathrm{pp})(\beta^2\mathrm{ap}+\mathrm{pp}+c+1)},$$

and the change in the weight of precision is simply $\Delta(1-p_\beta) = -\Delta p_\beta$. The cost-invariant change in F_β-score is

$$\begin{aligned}\Delta F_\beta =&\Delta\big(p_\beta\mathrm{R}\big) + \Delta\big((1-p_\beta)\mathrm{P}\big)\\ =&\Delta p_\beta\mathrm{R} + p_\beta\Delta\mathrm{R} + \Delta p_\beta\Delta\mathrm{R} + \Delta(1-p_\beta)\mathrm{P} + (1-p_\beta)\Delta\mathrm{P} + \Delta(1-p_\beta)\Delta\mathrm{P}\\ =&\Delta p_\beta\big((\mathrm{R}+\Delta\mathrm{R}) - (\mathrm{P}+\Delta\mathrm{P})\big) + p_\beta(\Delta\mathrm{R}-\Delta\mathrm{P}) + \Delta\mathrm{P}\\ =&\Delta p_\beta\big((\mathrm{R}+\Delta\mathrm{R}) - (\mathrm{P}+\Delta\mathrm{P})\big) + p_\beta\big(\Delta\mathrm{R} - (1-\tfrac{1}{p_\beta})\Delta\mathrm{P}\big).\end{aligned}$$

From the equations of $\Delta\mathrm{R}$, $\Delta\mathrm{P}$, and Δp_β, we know that

$$\mathrm{R}+\Delta\mathrm{R} = \frac{\mathrm{tp}+1}{\mathrm{ap}}, \qquad \mathrm{P}+\Delta\mathrm{P} = \frac{\mathrm{tp}+1}{\mathrm{pp}+c+1},$$

$$(1-\tfrac{1}{p_\beta})\Delta\mathrm{P} = \left(1 - \frac{\beta^2\mathrm{ap}+\mathrm{pp}}{\beta^2\mathrm{ap}}\right)\frac{\mathrm{pp}-(c+1)\mathrm{tp}}{\mathrm{pp}(\mathrm{pp}+c+1)} = -\frac{\mathrm{pp}-(c+1)\mathrm{tp}}{\beta^2\mathrm{ap}(\mathrm{pp}+c+1)},$$

and combining these, ΔF_β can be written as

$$\begin{aligned}\Delta F_\beta =&\frac{-\beta^2\mathrm{ap}(c+1)}{(\beta^2\mathrm{ap}+\mathrm{pp})(\beta^2\mathrm{ap}+\mathrm{pp}+c+1)}\left(\frac{\mathrm{tp}+1}{\mathrm{ap}} - \frac{\mathrm{tp}+1}{\mathrm{pp}+c+1}\right)+\\ &\frac{\beta^2\mathrm{ap}}{\beta^2\mathrm{ap}+\mathrm{pp}}\left(\frac{1}{\mathrm{ap}} + \frac{\mathrm{pp}-(c+1)\mathrm{tp}}{\beta^2\mathrm{ap}(\mathrm{pp}+c+1)}\right).\end{aligned}$$

Setting $\Delta F_\beta = 0$ and solving for β^2 gives

$$\beta^2 = \begin{cases} -1 \\ \frac{(c+1)\mathrm{tp}-\mathrm{pp}}{\mathrm{ap}} \end{cases}.$$

As β is a (positive) real number, and as pp = tp + fp, the only solution is

$$\beta = \sqrt{\frac{(c+1)\mathrm{tp}-\mathrm{pp}}{\mathrm{ap}}} = \sqrt{\frac{c\mathrm{tp}+\mathrm{tp}-\mathrm{tp}-\mathrm{fp}}{\mathrm{ap}}} = \sqrt{\frac{c\mathrm{tp}-\mathrm{fp}}{\mathrm{ap}}}.$$

Now, for a total cost $C^* = c\mathrm{fn}+\mathrm{fp}$, it holds that $\mathrm{fp} = C^* - c\mathrm{fn} = C^* - c(\mathrm{ap}-\mathrm{tp})$. Based on this, the solution of β can be written as

$$\beta = \sqrt{\frac{c\mathrm{tp} - C^* + c\mathrm{ap} - c\mathrm{tp}}{\mathrm{ap}}} = \sqrt{c - \frac{C^*}{\mathrm{ap}}}.$$

That is, the solution of β depends on total cost itself, and thus, irrespective of cost and class imbalance ratios, there does not exist any β for which F_β-score would be cost-consistent (given the assumed simple cost-structure).

Cost-invariant curves of F_β-score and informedness are illustrated in Figures 1 and 2. The illustrations were generated by decreasing tp (from tpr = 1) step by step and solving the rest of the cost-invariant confusion matrix each step for a given total cost, with ap = 100, an = 2,000, and $c = \frac{an}{ap} = 20$. Total cost of a curve (ranging from 0 to 2,000 = $c \times$ ap = an) is presented on the right-hand side. Figure 1 shows cost-invariant F_β-score curves[2] with two commonly used values, $\beta = 1$ and $\beta = 2$, and Figure 2 shows cost-invariant informedness curves with $c = \frac{an}{ap}$ and with an increase of 10% to the cost ratio.

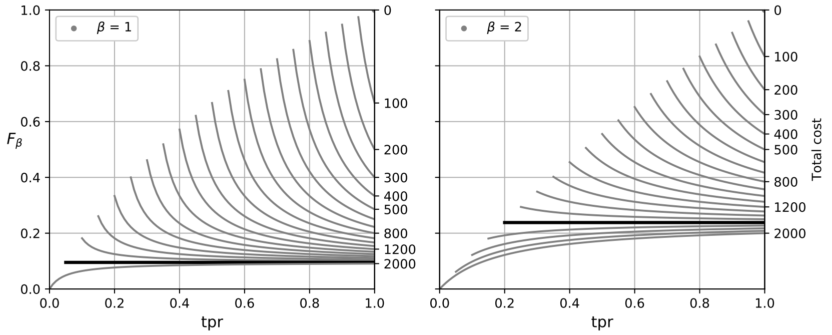

Fig. 1: Cost-invariant curves of F_β-score with $\beta = 1$ and $\beta = 2$. Total costs corresponding to constant F_β-scores, $C^* = (c - \beta^2)$ap, are bolded.

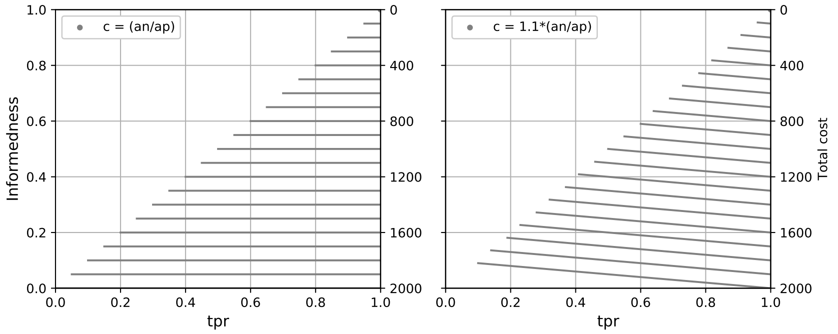

Fig. 2: Cost-invariant curves of informedness with two different cost ratios. Informedness is cost-consistent when $c = \frac{\text{an}}{\text{ap}}$.

[2] Note that, although precision is not defined if pp = 0, F_β-score is set to zero if pp = 0.

In Figure 1, the total costs corresponding to constant F_β-scores are highlighted with bolded lines. If total cost is less than that line, F_β-score has a bias towards too low tpr (and a bias towards too high tpr, if total cost is higher). On the other hand, as shown in Figure 2, even if the true cost ratio slightly differs from the class imbalance ratio, a similar bias in informedness is not as severe. However, the bias in informedness naturally increases the further away the true cost ratio is from the class imbalance ratio.

4 Discussion and conclusion

Total cost would be an appropriate evaluation measure for imbalanced classification, but exact costs are rarely available and estimating them is difficult. However, it would still make sense to apply an evaluation measure that is cost-consistent at least in a certain scenario, for example, informedness (or balanced accuracy) – even if that would implicitly assume a simple cost-structure. Another option would be to estimate the cost function and to use total cost as an evaluation measure. In practice, however, choosing a cost function may seem arbitrary. Yet, not setting a cost function and, instead, using heuristic evaluation measures can be seen as an arbitrary choice as well.

In real-life classification tasks, the aim of evaluation is simply to find the best classifier for the given application, and, consequently, the applied evaluation measure(s) can be anything that the practitioners see apt for the task – though, it would still be important that the practitioners understand how the measures behave, as using suboptimal evaluation measures may lead to suboptimal decisions. On the other hand, e.g., in method comparisons carried out by researchers, choosing an unconventional evaluation measure can be seen to compromise the integrity of obtained conclusions. Thus, research aiming to establish appropriate measures in imbalanced classification is of high importance.

References

[1] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[2] B. Zadrozny and C. Elkan. Learning and making decisions when costs and probabilities are both unknown. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 204–213, 2001.

[3] C. Elkan. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, pages 973–978, 2001.

[4] D. Hand and P. Christen. A note on using the F-measure for evaluating record linkage algorithms. *Statistics and Computing*, 28:539–547, 2018.

[5] D. M. W. Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2020.

[6] P. Christen, D. Hand, and N. Kirielle. A review of the F-measure: Its history, properties, criticism, and alternatives. *ACM Comput. Surv.*, 56(3), 2023.