The F-Measure Paradox

Tim vor der Brück

Fernfachhochschule Schweiz (FFHS)
Brig, Switzerland
Email: tim.vorderbrueck@ffhs.ch

***Abstract*—Paradoxes have raised a lot of interest in mathematics and computer science. What fascinates people about them is that such a paradox contains a self-contradictory statement that dissents with usual believes and expectations. The range of discovered paradoxes is long. One of the most famous is probably the proposition of Russell that states that no set can exist that contains all sets that do not contain itself as a subset. The paradox arises in the proof, where it is shown that such a set must contain itself if and only if it does not contain itself. In this paper, we derive a paradox about the F-measure, one of the most important metrics in machine learning. The contribution of this paper is two-fold. On the one hand, we investigate typical properties of the F-Measure, on the other hand, we show that they are contradictory and therefore constitute a paradox, to several properties of the harmonic mean, where the F-Measure is a special case of.**

Keywords–F-Measure; paradox; precision; recall; NaN.

I. INTRODUCTION

The word *paradox* originates from Greece and is composed of the word para (beyond) and doxa (opinion). A paradox contains a self-contradictory statement and dissents with people's believes and expectations [1]. It often does not have a direct practical use case but it gives theoretical insights and helps to understand certain problems better. Especially in the area of mathematics, there is a large amount of identified paradoxes. A quite well-known paradox is the proposition of Russel.

Russel's Paradox: This proposition [2] claims that no set can exist that contains all sets that do not contain themselves and nothing more. The proof is done by contradiction. Let us assume such a set would exist. Then exactly one of the following propositions must be true about this set:

- This set contains itself. This is not possible since this set only contains sets that do not contain themselves.
- This set does not contain itself. Then per definition, this set must contain itself, which is a contradiction.

Since both cases lead to a contradiction, such a set cannot exist.

Banach-Tarski Paradox: Another well-known paradox from mathematics is the so-called Banach-Tarski-Paradox [3] that claims that a sphere can be decomposed and put together afterward in such a way that one has obtained two spheres of the same volume as the original sphere. Thus, one of the spheres was seemingly created out of nothing. This paradox is based on the principle that some concepts of mathematics cannot be transferred into reality.

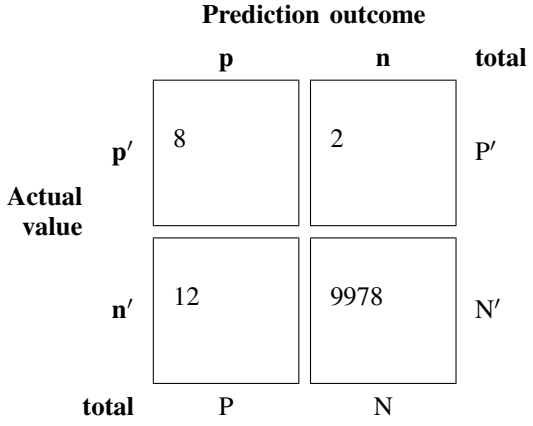

Figure 1. Example of confusion matrix for an imbalanced class distribution.

Stein's Paradox: Normally, the expected value is best approximated by the average value, since the average value is actually its best unbiased estimator. Stein's paradox [4] states that, if several expected values of the same type are to be determined (like batting statistics for a collection of baseball players), the isolated averages are no longer the best choice. Instead, all the estimates should be determined jointly by shifting the individual estimates in direction of the overall cross-estimate average.

Accuracy Paradox: Related to Data Science is the so-called accuracy paradox [5][6]. It states that when comparing two classification methods, the one with the lower accuracy can have in fact higher predictory capability. This phenomenon usually occurs in the case of highly imbalanced class distributions. Consider for example a very infrequent event like a rare disease that only shows up for around 0.1% of the cases. Let us assume, we have a method that can detect 40% of the events correctly and its precision is 80%. So, its confusion matrix could look like the one in Figure 1, where the columns denote the predicted and the rows the actual values. The obtained accuracy of this method would then amount to 9986/10000=0.9986 while predicting always the majority class (event not occurring), which has in fact no predictive power, would achieve an accuracy of 0.999.

In this paper, we will first derive several general statements about the harmonic mean of two variables. Afterward, we will proof that these statements are indeed incorrect for the F_1-score, which is a special case of the harmonic mean, in partic-

 ISBN: 978-1-61208-812-9

ular, it is the harmonic mean of precision and recall. Finally, we will analyze the reasons for this paradox and investigate the consequences for mathematical proofs in general.

The remainder of this paper is organized as follows. In Section II, we define the general harmonic mean and show several of its principal properties. Section III gives an overview of the F_1-score, which is the harmonic mean of precision and recall. In the next section (Section IV), the paradox of the F_1-score is described. The findings and the cause of this paradox are discussed in the next Section V. Finally, the paper concludes with Section VI, which summarizes the obtained results.

II. Harmonic Mean

The harmonic mean $H(a, b)$ of two values a and b is the Hoelder-mean with coefficient -1 and is formally given by [7]:

$$H(a,b) = \left(\frac{a^{-1} + b^{-1}}{2}\right)^{-1} = \frac{2}{\frac{1}{a} + \frac{1}{b}} \tag{1}$$

with $a, b \in \mathbb{R}$ (or $\mathbb{C}$). An alternative and simpler formulation is:

$$H(a,b) = \frac{2ab}{a+b} \tag{2}$$

In contrast to the arithmetic mean, the harmonic mean is a rather pessimistic mean that is drawn in direction to the minimum of both arguments. Note that it can only be applied to argument values of identical signs [7]. To see this, consider the following example:

$$H(-2,4) = \frac{2(-2)\cdot 4}{-2+4} = \frac{-16}{2} = -8 \notin [-2,4]$$

Since -8 is not located between -2 and 4, it cannot possibly constitute any mean of those values.

Consider now the following two propositions (1+2):

$$\begin{aligned} &1.\ a = 0 \Rightarrow H(a,b) = 0 \\ &2.\ H(a,b) = 0 \Rightarrow a = 0 \end{aligned} \tag{3}$$

Note that without limitation of generality a=0 can be replaced by b=0 due to the symmetry of the harmonic mean.

It is not difficult to show that the first statement is true and the second false.

Proof: Let us first have a look at proposition 1. The following two cases can be discerned: $b \neq 0$ and $b = 0$. First, we consider the case $b \neq 0$. Plugging in $a = 0$ in formula 2 results in:

$$H(a,b) = \frac{2 \cdot 0 \cdot b}{0+b} = \frac{0}{b} = 0 \tag{4}$$

Now consider $a = 0, b = 0$. Plugging both values into $H(a, b)$ results in an expression $\frac{0}{0}$, which is not defined. Let us, however, look at the behavior of $H(a, b)$ for a and b approaching zero using formula 1. Since the sign of a and b must coincide, we get:

$$\lim_{a,b\to 0} H(a,b) = \frac{2}{\frac{1}{a} + \frac{1}{b}} = \frac{2}{\infty} = 0 \tag{5}$$

Therefore, it is a reasonable approach to define $H(0, 0)$ as 0, to which we henceforth abide.

Proposition 2 is straight-forward to show by the following counterexample: $H(2, 0) = 0$ but $2 \neq 0$. ■

One can also draw some conclusions, under which conditions the harmonic mean H and one of its input arguments have to coincide. In particular, assuming a is not diminishing, then from the fact that a and H coincide one can infer that b must also assume their common value. The opposite, however, is false.

Formally, the first proposition (proposition 3) is true and the second (proposition 4) is false:

$$\begin{aligned} &3.\ a \neq 0 \wedge a = H(a,b) \Rightarrow b = H(a,b) \\ &4.\ a \neq 0 \wedge b = H(a,b) \Rightarrow a = H(a,b) \end{aligned} \tag{6}$$

Proof of Proposition 3: From the definition of the harmonic mean, it follows that: $H(a,b) = \frac{2ab}{a+b}$
Since $H(a, b)$ equals a, we can plugin a on the left-hand side:
$a = \frac{2ab}{a+b}$
Since $a \neq 0$, both sides can be divided by a
$1 = \frac{2b}{a+b}$
Afterward, we multiply both sides by $a + b$:
$a + b = 2b$
By subtracting b from both sides one finally obtains:
$a = b$ ■

The opposite direction (proposition 4) can be shown by contraction, let a=1,b=0=H(a,b), then herewith it follows that $a \neq H(a, b)$.

III. F_1-Score

The F_1-Score is the harmonic mean of precision and recall, where precision is the percentage of predicted positive events that are indeed positive, while recall is the percentage of positive events that are actually correctly detected by the algorithm [8]. All three measures originated from the area of information retrieval but quickly spread into other areas of machine learning too. Let TP be the true positives, i.e., the number of positive events that were correctly classified by the algorithm, FP the number of negative events that were actually classified as positive, and FN the number of positive events that were misclassified as negative. Then precision (prec), recall (rec), and F-measure are formally defined as follows:

$$\begin{aligned} prec &= \frac{TP}{TP+FP} \\ rec &= \frac{TP}{TP+FN} \\ F_1(prec, rec) &= H(prec, rec) \\ &= \frac{2prec \cdot rec}{prec + rec} \end{aligned} \tag{7}$$

Note that recall or precision can potentially be undefined. Consider, for example, that the positive class never shows up in the evaluation data. In this case, TP and FN assume both zero, which results in an undefined recall value. Similarly, if the positive class is never predicted, the precision is left undefined. Analogously to the definition of floating point numbers, we use the expression *NaN* to denote an undefined value, which

stands for *Not a Number*. We also define arithmetic on *NaN* in the following way by following the Bochvar extension [9]. Let $a \in \mathbb{R} \cup \{NaN\}$ be arbitrarily chosen, then:

$$\begin{aligned} a \cdot NaN &= NaN \\ a + NaN &= NaN \\ a - NaN &= NaN \\ \frac{a}{NaN} &= NaN \end{aligned} \tag{8}$$

As one can easily perceive, if at least one of the operator arguments assumes NaN, then also the result is NaN. Therefore, NaN is also called an absorbing element. Regarding the algebraic structure, $\mathbb{R} \cup \{NaN\}$ is a semi-group for both summation and multiplication with 0 (1 respectively) as its neutral element. $(\mathbb{R} \cup \{NaN\}, +)$ and $(\mathbb{R} \cup \{NaN\} \backslash \{0\}, \cdot)$ are no groups, since there is no inverse element of NaN.

Consider the example precision=NaN, and recall=0, then the F_1-score becomes

$$\begin{aligned} & F_1(NaN, 0) \\ =& \frac{2NaN \cdot 0}{NaN + 0} = \frac{NaN}{NaN} = NaN \end{aligned} \tag{9}$$

Note that sometimes, the F_1-score is also defined directly based on TP, FP, and FN as follows [10]:

$$F_1(TP, FP, FN) = \frac{2TP}{2TP + FP + FN} \tag{10}$$

which leads to other behaviors regarding definedness. However, in this paper, we stick to the usual definition based on precision and recall.

IV. The F-Measure Paradox

Recall the four propositions from Section II.

$$\begin{aligned} &1.\ a = 0 \Rightarrow H(a,b) = 0 \\ &2.\ H(a,b) = 0 \Rightarrow a = 0 \\ &3.\ a \neq 0, a = H(a,b) \Rightarrow b = H(a,b) \\ &4.\ a \neq 0, b = H(a,b) \Rightarrow a = H(a,b) \end{aligned} \tag{11}$$

If we set a=prec(ision) and b=rec(all), those four propositions become:

$$\begin{aligned} &1.\ prec = 0 \Rightarrow F_1(prec, rec) = 0 \\ &2.\ F_1(prec, rec) = 0 \Rightarrow prec = 0 \\ &3.\ prec \neq 0, prec = F_1(prec, rec) \Rightarrow rec = F_1(prec, rec) \\ &4.\ prec \neq 0, rec = F_1(prec, rec) \Rightarrow prec = F_1(prec, rec) \end{aligned} \tag{12}$$

From Section II, one would expect that Proposition 1 and 3 are true and Proposition 2 and 4 are false. But, surprisingly, it is just the opposite. In fact, propositions 1 and 3 are false and propositions 2 and 4 are true.

Proof: For proposition 1, we give a counterexample. Consider the confusion matrix in Figure 2. For this matrix, the precision assumes 0 and the recall NaN. Therefore, the F_1-Score is given as $\frac{2 \cdot 0 \cdot NaN}{2 + NaN} = NaN \neq 0$, which concludes the proof by counterexample.

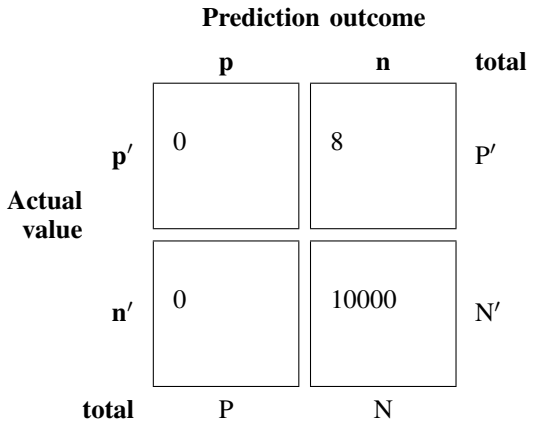

Figure 2. Example confusion matrix as counterexample for proposition 1

Proposition 2: Consider the second proposition and let us assume that the F_1-Score is zero. Hence, either precision or recall is zero. In case, the precision is zero, our proof is finished. So let us, therefore, assume instead that the recall is zero. Since the F_1-score is defined (not NaN), both recall and precision must be defined too. Furthermore, we have:

$$\begin{aligned} 0 = rec &= \frac{TP}{TP + FN} \\ &\Rightarrow TP = 0 \\ &\Rightarrow \frac{TP}{TP + FP} = 0 \\ &(\text{Precision is not } NaN, \text{ therefore } TP + FP \neq 0) \\ &\Rightarrow prec = 0 \end{aligned} \tag{13}$$

■

Proposition 3: Again, we give a counterexample, we can use the same confusion matrix as for proposition 1. With this we get $prec = NaN = F_1(prec, rec)$ and $rec = 0$.

Proposition 4:

Proof: We discern the following three cases:
Case 1: $rec = F_1(prec, rec) = NaN$

$$\begin{aligned} rec &= F_1(prec, rec) = NaN \\ \Rightarrow TP &= 0 \\ \Rightarrow FP + TP &= 0 \\ (\text{since } prec &\neq 0) \\ \Rightarrow prec &= NaN = rec = F_1(prec, rec) \end{aligned} \tag{14}$$

Case 2: $rec = F_1(prec, rec) = 0$
Due to proposition 2, it follows that $prec = 0 = F_1(prec, rec)$. Since the precision cannot diminish, this case actually turns out to be impossible.
Case 3: $rec = F_1(prec, rec) \neq 0$ and $rec = F_1(prec, rec) \neq NaN$.

The precision cannot assume NaN, since otherwise the F_1 score would be NaN, too.

Furthermore, from the definition of the harmonic mean, it is known that:

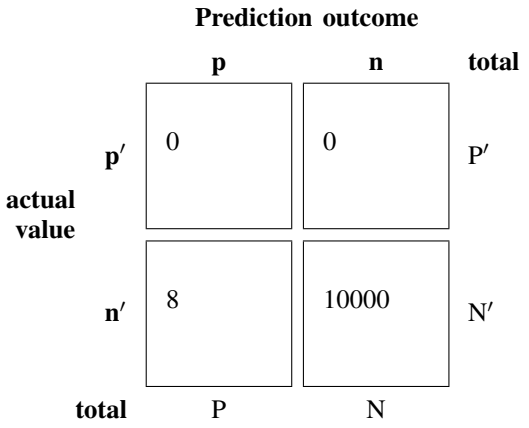

Figure 3. Example confusion matrix as an counterexample for proposition 1 (case recall)

$rec = F_1(prec, rec) = \frac{2prec \cdot rec}{prec + rec}$

Since $rec \neq 0$, we can divide both sides of the equation by the recall and obtain:
$1 = \frac{2prec}{prec + rec}$
We then multiply both sides of the equation by $prec + rec$:
$prec + rec = 2prec$
Finally, we subtract from both sides of the equation the precision and get:
$rec = prec$ which constitutes just the conducted claim. ■

Finally, let us investigate if the same paradox holds also for F-measure and recall instead of precision. Analogously to the precision case, we first present a counterexample (see Figure 3).

This time, the recall is 0 and the precision NaN, which leads to a NaN F_1-Score.

Proof: Let us now prove the second proposition for F-measure and recall. Again, since the F_1-score is zero, neither of precision and recall can be NaN. If the recall is zero, our proof is finished. Therefore, let us instead assume the precision is zero.

$$\begin{aligned} 0 = Precision &= \frac{TP}{TP + FP} \\ &\Rightarrow TP = 0 \\ &\Rightarrow \frac{TP}{TP + FN} = 0 \\ &\Rightarrow Recall = 0 \end{aligned} \tag{15}$$

■

The two remaining properties 3 and 4 can be proven analogously.

V. Discussion

Purely formally seen, the computation rules for NaN values are mathematically consistent and correct and also reflect the standard procedure for computer-based F_1-Score implementations if they make use of ordinary floating-point computation logic. It remains, however, to investigate, if these rules are also reasonable in the given context. The answer is partly yes and partly no. Consider first the case the recall is undefined (NaN), which means that the positive class never shows up in the evaluation data set. If the algorithm predicts only for a single data item the positive class, then the precision immediately turns to zero. In this case, an NaN F_1-Score seems to be the best choice.

However, if the precision is undefined (NaN), the matters look a bit different. In this case, the tested machine learning method would never predict the positive class. If the positive class shows up quite a few times in the evaluation data, such a method would clearly perform very poorly and an F_1-score of zero would seem adequate.

So far, we investigated only the F_1-score, although in the title we mentioned the F-measure in general. This generalized F-measure is given by:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}} \tag{16}$$

For $\beta = 2$ we get for instance the F_2-score, which is defined as:

$$F_2 = 5 \cdot \frac{\text{precision} \cdot \text{recall}}{(4 \cdot \text{precision}) + \text{recall}} \tag{17}$$

The F-measure for $\beta \geq 2$ penalizes a poor recall stronger than a bad precision, since in some situations, like cancer detection, missing any items of the positive class can be fatal in practice. However, the use of different weighting factors does not influence in any way the properties derived here. Thus, our findings also hold for the F-measure in general.

Finally, this paradox also reveals a shortcoming of most mathematical proofs. Undefined values are not rare in practice. They can be caused by missing values or incomputability as investigated here. Albeit, in proofs, they are usually completely ignored. The paradox investigated here shows that such undefined values can easily flip statements completely around.

While the findings as stated here are mainly theoretical, they can have some practical implications as well. If the different behaviors of harmonic mean and F-measure as described here were ignored, then in certain anomalous situations, incorrect conclusions might be drawn from the data.

VI. Conclusion

We presented two basic statements about the harmonic mean, where the first is true and the second false. However, for the F_1-score as the harmonic mean of precision and recall, the truth value of both statements is completely turned around. This paradox is caused by the fact that the possibility that input values can be undefined is not taken into account in the original propositions for the harmonic mean. Hence, with this paradox, we also revealed an important shortcoming of mathematical proofs in general.

References

[1] S. J. Farlow, Paradoxes in Mathematics. Chicago, Illinois: Dover Books, 2014.

[2] A. Whitehead and B. Russell, The Principles of Mathematics, 2nd ed. New York, New York: W. W. Norton & Company, 1996.

 ISBN: 978-1-61208-812-9

[3] S. Banach and A. Tarski, “Sur la décomposition des ensembles de points en parties respectivement congruentes (on the decomposition of sets of points into respectively congruent parts),” Fundamenta Mathematicae, vol. 6, 1924, pp. 244–277.

[4] B. Efron and C. Morris, “Stein’s paradox in statistics,” Scientific American, vol. 236, no. 5, 1977, pp. 119–127.

[5] B. Abma, “Evaluation of requirements management tools with support for traceability-based change impact analysis,” Master’s Thesis, University of Twente, 2009.

[6] J. S. Akosa, “Predictive accuracy : A misleading performance measure for highly imbalanced data,” in Proceedings of the SAS Global Forum, 2017, p. 2–5.

[7] D. B. MacNeil, Fundamentals of Modern Mathematics: A Practical Review. Dover Publications, 2013.

[8] C. D. Manning and H. Schütze, An Introduction to Information Retrieval. Cambridge University Press, 2009.

[9] L. Běhounek and M. Daňková, “Extending aggregation functions for undefined inputs,” in Proceedings of the International Symposium on Aggregation and Structures, 2018, pp. 15–16.

[10] D. Chicco and G. Jurman, “The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” BMC Genomics, vol. 21, no. 6, 2020, pp. 1–13.

 ISBN: 978-1-61208-812-9