RECLAIM: Reverse Engineering Classification Metrics

Flavio Giobergia
Department of Control and Computer Engineering
Politecnico di Torino
Turin, Italy
flavio.giobergia@polito.it

Elena Baralis
Department of Control and Computer Engineering
Politecnico di Torino
Turin, Italy
elena.baralis@polito.it

***Abstract*—Being able to compare machine learning models in terms of performance is a fundamental part of improving the state of the art in a field. However, there is a risk of getting locked into only using a few – possibly not ideal – performance metrics, only for comparability with earlier works. In this work, we explore the possibility of reconstructing new classification metrics starting from what little information may be available in existing works. We propose three approaches to reconstruct confusion matrices and, as a consequence, other classification metrics. We empirically verify the quality of the reconstructions, drawing conclusions on the usefulness that various classification metrics have for the reconstruction task.**

***Index Terms*—model evaluation, performance reconstruction, classification metrics**

I. Introduction

The field of machine learning has continually expanded in recent years. With this expansion, a growth in the number of publications has followed. These publications typically advance the state of the art by introducing new techniques or modifications to existing ones. The performance of such techniques is typically measured on existing benchmark datasets. It often happens that such works only publish partial information in terms of performance achieved. In the case of classification problems, for example, only the accuracy of a model may be reported, whereas information such as the precision or recall may be omitted. This makes it impossible to assess the performance of the published model in terms of metrics that differ from the ones presented in the original manuscript. As a consequence, novel works that wish to compare against state-of-the-art techniques must run the comparisons in terms of fixed metrics, even when such metrics may not be the most adequate, only because of compatibility reasons.

Although these problems could be overcome by making the implementations and datasets freely available, this option cannot always be pursued: either because the authors cannot be reached, or because they do not wish to share the implementation details of their works.

To work around this problem, we analyze a classification scenario and explore how, and to what extent, we may be able to reconstruct new metrics not known in advance starting from the information that is actually available. We introduce RECLAIM (Reverse Engineering CLAssIfication Metrics) a methodology that can either reconstruct new metrics exactly, or that produces upper and lower boundaries for them – depending on the amount of available information. We experimentally show the quality of the reconstructions obtained. We make the source code for RECLAIM openly available, to allow other researchers to easily use it[1].

II. Related works

The idea of reconstructing unavailable information from what little *is* available has been often approached in the fields of statistics and machine learning. In past works, researchers focused on the reconstruction of plausible samples from an unknown distribution, given some summary statistics (e.g. mean, standard deviation) and additional context-specific constraints (e.g. the domain of possible values). For example, SPRITE [1] generates random samples with the expected mean, then redistributes "mass" across the samples to obtain the desired standard deviation. Instead, CORVIDS [2] uses Diophantine equations to reconstruct all possible response patterns that can generate some provided summary statistics. These approaches can be useful when the data itself has not been made available by the authors of the works under study.

The GRIM test [3] has been created for similar situations. However, rather than reconstructing the raw samples, the GRIM test can be used to check that there can exist a dataset with the required characteristics, in terms of sample size and mean. This test can be used to verify the reliability of results published in scientific publications and has been used by the authors to identify various such situations: this led to the identification of various reporting errors. As we will show, RECLAIM may also be used to verify whether a reported result is plausible or not – thus extending the usefulness of the GRIM test to various classification metrics.

Other reconstruction approaches instead attempt to rebuild samples used for the training of a model, starting from the model itself or from its predictions. For example, in [4], the authors show how training samples can be extracted by querying a language model. Similarly, the membership inference problem [5] attempts to confirm or deny whether a given sample has been used in the training of a model. These techniques are once again focused on the reconstruction of

[1] https://github.com/fgiobergia/RECLAIM

data points, rather than metrics, but still address a problem where part a proposed solution (in this case, the training data) is not available.

In this work we instead focus on the reconstruction of additional evaluation metrics given some information on the data (number of samples and labels distribution) and on the classifier (e.g. one of a number of classification metrics obtained by the model). This specific problem, to the best of our knowledge, has not been previously approached in literature.

III. Methodology

For a dataset X and corresponding binary labels y, we assume known the following information: its total population C and the portion of positive labels N_P. As a consequence, the number of negative labels $N_N = C - N_P$ is known. We assume that some classification model c has been tested on X to predict y. The main assumption about the model is that it is not known (e.g. the case if the model has been built by a third party). The confusion matrix for such a model (i.e. the distribution of values across *true positives (tp)*, *true negatives (tn)*, *false positives (fp)* and *false negatives (fn)*) is also assumed to be unknown. For convenience, we can represent a confusion matrix as a vector $v \in \mathbb{N}^4$. Throughout the paper, we will refer to v or to the four quantities interchangeably. However, we assume that some metrics $m_1(v)$, $m_2(v)$, ... (e.g. precision, recall, accuracy) are known about the model. Most such metrics can be expressed as a function of the confusion matrix [6]. Our goal is to use the known information on the model to rebuild the confusion matrix (or an estimate of its boundaries), so as to be able to reconstruct additional evaluation metrics $\bar{m}_1(v), \bar{m}_2(v), ...$ for c that were not previously known. For this work, we restrict the reconstructing/reconstructed metrics to accuracy, precision, recall (or sensitivity) and F_β score. However, additional metrics can be introduced so long as they can be represented as a linear function of the confusion matrix (e.g. specificity, error rate).

We can express each known constraint as a scalar product of the confusion matrix and a vector of weights. For example, $C = [1\ 1\ 1\ 1]v$ and $N_P = [1\ 0\ 0\ 1]v$. Similarly, we can rewrite metrics such as the precision as[2] $(P-1)tp + Pfp = 0$.

Thus, $0 = [P-1\ 0\ P\ 0]v$. Any of the mentioned metrics can be likewise transformed. Table I provides the respective coefficients and intercepts for some common such metrics. Given n metrics, we can build a matrix $W \in \mathbb{R}^{n\times 4}$ of constraints, whose rows are the coefficients, and an intercept $b \in \mathbb{R}^n$. We can thus rewrite the system of constraints as $Wv = b$.

In this work we refer to a situation where C and N_P are known, along with other constraints in the form of metrics. The proposed methodology works even when C or N_P are replaced with other metrics. However, the proposed results may not reflect the original dataset in terms of size and distribution of labels.

[2]For $tp + fp \neq 0$

We explore three different approaches to this reconstruction problem:

- When enough information is known about the metrics, we can solve a system of equations and produce an exact confusion matrix
- When only one metric is known, we frame an integer programming (IP) problem to obtain boundaries for the confusion matrix
- When the metric known is the accuracy, we extract a closed form solution and prove some useful properties of the reconstructed confusion matrix boundaries

A. Fully constrained problem

If W is square ($n = 4$) and invertible ($det(W) \neq 0$), finding the confusion matrix is trivial: $v = W^{-1}b$. If W and b are known with full precision, the reconstructed v is expected to be $\in \mathbb{N}^4$. When rounding is applied to the available metrics, which in turn affect W and b, we instead find that $v \in \mathbb{R}^4$ is an approximation of the true confusion matrix. Given the assumption on $det(W)$, v is guaranteed to exist. However, if the reconstructing metrics are inconsistent, the values in v may be unacceptable (for example, if any of the values is negative). This property may be used to determine the reliability of published results.

If there are more constraints available than degrees of freedom ($n > 4$), W^{-1} cannot be computed. The Moore-Penrose pseudoinverse [7] can instead be used to obtain the best fitting solution (in terms of least squares). We have empirically observed that adding additional constraints helps reconstruct confusion matrices that are closer to the true ones, when rounding is applied to the metrics.

B. Confusion matrix boundaries reconstruction

It may be the case that not enough constraints are available to reconstruct the confusion matrix. In those situations, we can still infer useful information on the confusion matrix in terms of upper and lower boundaries. We will focus on the scenario where 3 constraints are known (C, N_P and an additional one), with a brief discussion for cases where a lower number of constraints available.

In the general case, having 3 constraints (i.e. 1 degree of freedom) implies that infinite solutions exist. Although the Moore-Penrose pseudoinverse of W could be computed for situations where $n < 4$, the result would be a single solution (the one with lowest norm) that may not be close to the correct one. Instead, we propose an approach to define upper and lower boundaries for the correct solution.

Given 1 degree of freedom, we can fix a single quantity (e.g. the true positives tp) and identify all other quantities in the confusion matrix. Since all quantities in v are positive integers, we can frame the problem as an Integer Programming (IP) one. In particular, we can identify the minimum and maximum values for one of the dimensions of v. The choice of which dimension should be chosen is not trivial. In some cases, any one will be acceptable. In other cases (e.g. when N_P and recall are both known, since $tp = RN_P$), the choice should be such

TABLE I
FORMULATION OF VARIOUS CONSTRAINTS IN TERMS OF A LINEAR COMBINATION OF THE CONFUSION MATRIX.

Metric name	Linear combination of confusion matrix	Coefficients	Intercept
Count (C)	$tp+tn+fp+fn=C$	[1 1 1 1]	C
# positive samples (N_P)	$tp+fn=N_P$	[1 0 0 1]	N_P
Accuracy (A)	$(A-1)tp+(A-1)tn+Afp+Afn=0$	$[A{-}1\ A{-}1\ A\ A]$	0
Precision (P)	$(P-1)tp+Pfp=0$	$[P{-}1\ 0\ P\ 0]$	0
Recall (R)	$(R-1)tp+Rfn=0$	$[R{-}1\ 0\ 0\ R]$	0
F_β score (F_β)	$(F_\beta-1)(1+\beta^2)tp+F_\beta fp+F_\beta\beta^2 fn=0$	$[(F_\beta{-}1)(1{+}\beta^2)\ 0\ F_\beta\ F_\beta\beta^2]$	0

that a degree of freedom is actually removed. In the case of N_P and recall, for example, any choice other than tp can be selected. In the general case, we can select the first dimension whose introduction does not make the matrix of constraints singular. We identify the position of such dimension as d. Based on the range of values that can be assumed by v_d, the range for the other dimensions of v can be defined.

In particular we can frame the following IP problem to identify the lower bound for v_d:

$$\begin{aligned} \min_{v} \quad & v_d \\ \text{s.t.} \quad & \sum_j W_{ij} v_j = b_i \qquad (i=0,1,2), \\ & v_j \geq 0 \qquad\qquad (j=0,1,2,3), \\ & v_j \text{ integer} \qquad (j=0,1,2,3). \end{aligned} \tag{1}$$

An additional constraint on $tp \geq 1$ must be added in the situations where the 0 solution cannot not accepted. For example, if a precision $P > 0$ is provided, the constraint $(P-1)tp+Pfp=0$ would be followed for $tp=0 \wedge fp=0$. However, this constraint clearly does not result in a valid precision. By enforcing a constraint such as $tp \geq 1$, we avoid this situation.

Similarly, we can frame a maximization problem based on Equation 1 to identify the upper bound for v_d. When an upper and a lower bound for v_d are found, the boundaries for the rest of the confusion matrix follow, since there are no other degrees of freedom left.

As already argued, the previous considerations apply to the case with a single degree of freedom. When fewer than 3 constraints are available, the number of degrees of freedom increases. While similarly posed IP problems could be framed, the boundaries could grow so large as to not provide any meaningful information on the original performance. As such, the 1- and 2-constraints scenarios are not covered in this work.

The proposed IP problem is acceptable when the metrics are known with full numerical precision (i.e. no rounding has been applied). However, the full-precision results are seldom published in papers. When rounding is applied, a full precision value m is mapped to the nearest "allowed" value, $\hat{m}$: it follows that there exists an interval of points around $\hat{m}$ whose rounding will always produce $\hat{m}$. Conventionally, when reporting k significant figures, the interval of points around $\hat{m}$ that will be converted to $\hat{m}$ will be[3] $[\hat{m}-\epsilon, \hat{m}+\epsilon]$, where $\epsilon = 5 \cdot 10^{-k-1}$. For example, a rounded accuracy of 0.8913 (4 significant figures) may only have been produced by an "actual" accuracy in the range $[0.89125, 0.89135]$ ($\epsilon = 5 \cdot 10^{-5}$). The equality constraints from Equation 1 can thus be converted into inequality constraints:

$$\begin{aligned} \min_{v} \quad & v_d \\ \text{s.t.} \quad & \sum_j W_{ij} v_j \geq b_i - \epsilon_i \qquad (i=0,1,2), \\ & \sum_j W_{ij} v_j \leq b_i + \epsilon_i \qquad (i=0,1,2), \\ & v_j \geq 0 \qquad\qquad (j=0,1,2,3), \\ & v_j \text{ integer} \qquad (j=0,1,2,3). \end{aligned} \tag{2}$$

Where ϵ_i can be computed for all constraint based on their rounding. It follows that a more significant rounding will result in a wider set of possible solutions being found. In the experimental section we explore how this rounding affects the obtained results.

Although we have discussed the boundaries obtained for the confusion matrix, we have not considered how these affect the boundaries of the reconstructed metrics, which is of as much interest. For this kind of study, we studied the situation where accuracy is known, along with C and N_P.

C. Confusion matrix boundaries from accuracy

The scenario where the only available metric is the accuracy is arguably one of the most interesting ones. The accuracy metric is a commonly used one. However, for unbalanced problems, the performance of the classifier on minority classes (which are often those of more interest) is not well-represented by accuracy. For these reasons, we explore the accuracy-based situation and provide a closed form representation of all boundaries of the confusion matrix.

When the accuracy is known, we can set a constraint on the sum of true positives and true negatives (i.e. all samples that have been classified correctly). We can once again focus on identifying an upper and lower bound for the number of true positives.

For convenience, we will refer to the quantity $A' = tp+fn$ as the accuracy. This is the raw count of correctly classified samples and is related to the commonly known concept of accuracy as $A' = AC$.

For a fixed value of accuracy, we can identify the upper bound of tp (tp_{max}) by simply studying the case where as many positive samples have been labelled correctly. This

[3]Although the upper bound should not be included, we include it to produce a non-strict inequality, necessary when framing an IP problem

TABLE II
SUMMARY OF THE POSSIBLE UPPER AND LOWER BOUNDARIES ON tp

	$A' > N_P$	$A' \leq N_P$
$A' > C - N_P$	$tp_{min} = A' - C + N_P$ $tp_{max} = N_P$	$tp_{min} = A' - C + N_P$ $tp_{max} = A'$
$A' \leq C - N_P$	$tp_{min} = 0$ $tp_{max} = N_P$	$tp_{min} = 0$ $tp_{max} = A'$

results in two cases: one where $A' > N_P$, in which case $tp_{max} = N_P$ (all positive samples have been correctly predicted), the other where $A' \leq N_P$, in which case $tp_{max} = A'$ (all correctly predicted samples are positive).

Since we are studying the binary case, the lower bound for tp (tp_{min}) will occur when tn reaches its upper bound (tn_{max}) – this is due to the constraint imposed by the fixed accuracy. We can thus define two cases, that for $A' > N_N$ – where $tn_{max} = N_N$ and that for $A' \leq N_N$, here $tn_{max} = A'$. In both cases, $tp_{min} = A' - tn_{max}$.

We have thus identified four different situations that can occur. These situations define specific boundaries for the true positives and, as a consequence, for the confusion matrix. Table II summarizes these four scenarios. When a value for $tp \in [tp_{min}, tp_{max}]$ ia fixed, the rest of the corresponding confusion matrix can be inferred given the other constraints in place.

As shown in Table II, there are four possible scenarios that can be identified, based on the value of accuracy and on the distribution of class labels, identified by N_P and $N_N = C - N_P$. We explore the cases in which each scenario can occur.

For convenience, we will identify the four scenarios as follows:

i $A' > N_P \wedge A' > N_N$
ii $A' \leq N_P \wedge A' > N_N$
iii $A' > N_P \wedge A' \leq N_N$
iv $A' \leq N_P \wedge A' \leq N_N$

The scenarios under study all depend on how A' relates to N_P and N_N (i.e. whether it is larger or smaller than those values). We can thus identify two cases: one where $N_P < N_N$ (the positive class is the minority one) and the alternative scenario, $N_P \geq N_N$ (the positive class is the majority one).

In the first case, where $N_P < N_N < C$, we can find A' in any of three non-overlapping situations: $A' \in [0, N_P]$, $A' \in (N_P, N_N]$ and $A' \in (N_N, C]$. It can be easily verified that the three situations relate, respectively, to scenarios iv, iii and i.

Similarly, the second case is for $N_N \leq N_P < C$. Through a similar reasoning, we find that $A' \in [0, N_N]$, $A' \in (N_N, N_P]$ and $A' \in (N_P, C]$ relate to scenarios iv, ii and i respectively.

We see that, if A' is assumed uniformly distributed in $[0, C]$, the probability of each scenario is a function of the balance of the classes. It should be noted, however, that a naive classifier K that always predicts as output the majority class label will have an accuracy $A'_K = max(N_P, N_N)$. This means that any classifier that outperforms K in terms of accuracy will belong to scenario i.

For completeness, we show that some commonly adopted metrics (specifically recall, precision, F_β score) are bounded as well. To this end, we show that those functions are monotonic w.r.t. tp: as such, tp_{min} and tp_{max} will identify a lower and an upper bound.

1) Boundaries of recall: The recall is defined as $\frac{tp}{N_P}$. Since N_P is a positive constant, the recall is a linear function of tp and, as such, it is monotonic increasing. The boundaries for the recall are therefore:

$$R \in \left[\frac{tp_{min}}{N_P}, \frac{tp_{max}}{N_P}\right] \tag{3}$$

2) Boundaries of precision: Using the available constraints, we can write $fp = C + tp - A' - N_P$. As such, the precision is defined as the following homographic function:

$$P = \frac{tp}{2\,tp + C - A' - N_P} \tag{4}$$

We can study the monotonicity of $P(tp)$ by studying $\frac{\partial P}{\partial tp}$:

$$\frac{\partial P}{\partial tp} = \frac{C - A' - N_P}{(2\,tp + C - A' - N_P)^2} \tag{5}$$

The precision is monotonic increasing (non-negative first derivative) when $A' \leq C - N_P$ and is monotonic decreasing otherwise. As such, $P \in [P_{min}, P_{max}]$, where:

$$P_{min} = \begin{cases} \dfrac{tp_{min}}{2\,tp_{min} + C - A' - N_P} & \text{if } A' \leq C - N_P \\ \dfrac{tp_{max}}{2\,tp_{max} + C - A' - N_P} & \text{otherwise} \end{cases} \tag{6}$$

$$P_{max} = \begin{cases} \dfrac{tp_{max}}{2\,tp_{max} + C - A' - N_P} & \text{if } A' \leq C - N_P \\ \dfrac{tp_{min}}{2\,tp_{min} + C - A' - N_P} & \text{otherwise} \end{cases} \tag{7}$$

3) Boundaries of F_β score: Since $fp = C + tp - A' - N_P$ and $fn = N_P - tp$, we can express the F_β score as a function of tp:

$$F_\beta = \frac{(1+\beta^2)tp}{2\,tp + (\beta^2 - 1)N_P + C - A'} \tag{8}$$

As with the precision (Equation 4), the F_β score is a homographic function, with first derivative:

$$\frac{\partial F_\beta}{\partial tp} = \frac{(\beta^2+1)((\beta^2-1)N_P + C - A')}{(2\,tp + (\beta^2-1)N_P + C - A')^2} \tag{9}$$

We once again notice that this function is monotone. It is increasing for $(\beta^2-1)N_P + C - A' \geq 0$, decreasing otherwise. In the common case where $\beta = 1$, we notice that the function will always be increasing, since $A' \leq C$ by definition. In that case, we can identify the boundaries as:

$$F_1 \in \left[\frac{2\,tp_{min}}{2\,tp_{min} + C - A'}, \frac{2\,tp_{max}}{2\,tp_{max} + C - A'}\right] \tag{10}$$

For this specific case with known accuracy, we have shown that the boundaries on the confusion matrix also define the boundaries on other classification metrics: the highest and lowest values that can be obtained for each metric are the

ones that are computed on the upper and lower bounds of the confusion matrix. Although we do not formally prove this for other combinations of reconstructing metrics, we empirically observed throughout the experimental phase that this is also the case for all other situations.

IV. Experimental results

In this section we report the experimental results obtained when reconstructing various classification metrics, given some constraints. We only focus the experimental part on the results obtained by solving the IP problem, since (1) the results obtained for the fully-constrained problem are exact or, when not exact, their quality is a function of the rounding applied and (2) for the accuracy-specific scenario, the boundaries on the confusion matrix and on various commonly adopted metrics are already provided in closed form.

All results reported in this section will be in the form of widths of ranges. With the methodologies proposed in this work we can reconstruct, for any metric, a range of values $[m_{min}, m_{max}]$ such that the true (unknown) metric m is in $[m_{min}, m_{max}]$. The width of the range of possible values is $m_{max} - m_{min}$ and represents the size of the span of values that could be assumed by m. An accurate reconstruction is given by a small value of $m_{max} - m_{min}$ (a perfect reconstruction is given by $v_{max} - v_{min} = 0$).

A. Experimental setup

We have shown how the reconstruction of the confusion matrix depends on characteristics of the dataset itself (dataset size and balancedness of the problem), as well as characteristics of the classification models (in terms of performance achieved). To control for these various aspects, we focus the experimental section on synthetic datasets, for which various aspects can be easily controlled. Each synthetic dataset is generated to have C points in its test set, N_P of which belong to the positive class. Points belonging to the positive and negative classes are drawn from two normal distributions in a D-dimensional space, unless otherwise stated. We study the effect on the boundaries identified as C and N_P change (Subsections IV-B and IV-C respectively). Since no relationship has been identified with the dimensionality of the dataset, D is fixed to 16. For all experiments, the datasets will be divided into a training set and a test set, using an 80/20 split.

To control the classifier's performance, we intervene on the capacity of the classification model used. In particular, we use a decision tree classifier for all experiments, and we vary the depth m_d that the tree can reach. However, drawing samples from two normal distributions requires very simple decision boundaries to be identified, thus making it impossible to properly study the reconstruction behavior at higher capacities. For that reason, we introduced a different dataset for this study. More specifically, because of the properties of decision trees, we created a 2D dataset where each point $(x, y) = (z+\epsilon_1, z+\epsilon_2)$ is generated by drawing a scalar z from $U(0, 1)$ and ϵ_1, ϵ_2 from $U(0, \epsilon)$. The class label is assigned based on whether $x > y$ (i.e. whether $\epsilon_1 > \epsilon_2$. For this dataset, the perfect decision boundary is given by the identity function. However, since decision trees can only produce splits that are orthogonal to the dimensions of the dataset, the only way to approximate the identity function is to iteratively add splits. As such, increasing the depth that can be reached by the tree is guaranteed to produce better performance (up to the point where a satisfactory approximation of the identity function is reached). Figure 1 shows how the performance of the decision tree improves as with its capacity, until saturation (at a depth of ~ 15).

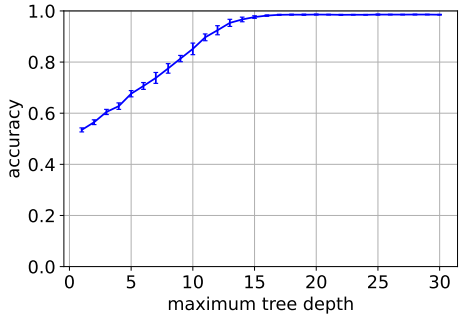

Fig. 1. Accuracy as the maximum depth of the tree increases, along with its capacity. The error bars represent $\pm$ 1 standard deviation of the accuracy computed on 10 separate datasets. Each dataset contains 25,000 points and two balanced classes. Other metrics can be easily shown to follow the same trend as the accuracy

In Subsection IV-D we present the results as the capacity of the model varies.

Finally, in Subsection IV-E we study the case where two real datasets are used, to observe how the conclusions drawn for the synthetic datasets generalize to real-world cases.

All experiments, unless otherwise stated, are performed by applying a rounding to 4 significant figures to the reconstructing metrics, to simulate a realistic situation.

B. Dataset size

We study the quality of the reconstructed metrics as the dataset size varies from 500 to 250,000 samples (i.e. test sets of 100 to 50,000 samples). We maintain a balanced dataset (50% positive and 50% negative samples) and a decision tree with maximum depth of 6. We solve the IP problem passing C and N_P as constraints, as well as one of the four metrics under study. Figure 2 shows the results.

As a general trend, we observe that the range of values reconstructed are hardly affected by the size of the dataset, as they are almost constant as the dataset size increases. This behavior is to be expected, considering that the dataset size only represents a constraint on the overall number of points contained in the confusion matrix, and not on their displacement. Nonetheless, we can draw other interesting conclusions from this experiment.

Most notably, the metric that is least reliable in defining the quality of the ranges is the precision. While this metric reconstructs, on average, better ranges w.r.t. the recall, its error bars clearly indicate a high variability in performance.

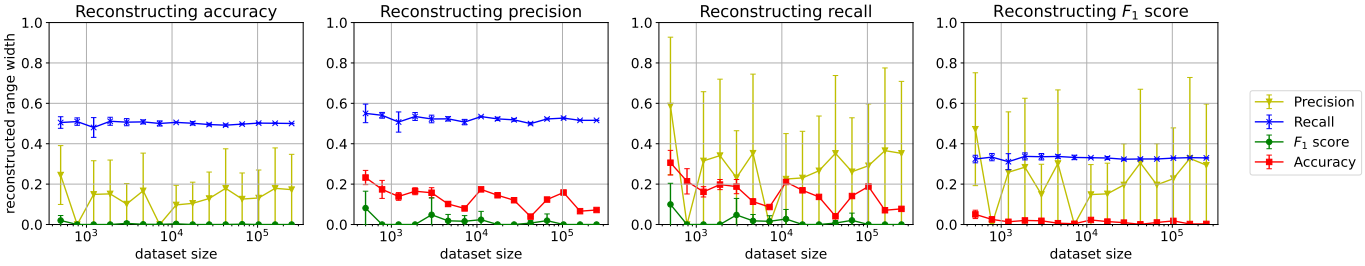

Fig. 2. Reconstructed ranges for various metrics (one for each figure), given C, N_P and either of the other metrics (specified by the lines), as the dataset size – and consequently C – varies. Each range is calculated as the difference between the upper and lower bounds obtained for the reconstructed metric, and is obtained as the mean over 10 separate datasets. The error bars represent $\pm$ 1 standard deviation across the 10 datasets.

This means that the reconstructions obtained starting from the precision should be evaluated case-by-case, as they could be particularly accurate, or not useful at all, regardless of the size of the dataset. This behavior occurs because the precision metric only bounds the true and false positives, introducing no constraint on the values of true and false negatives. As such, no constraint can be placed on true and false negatives other than that given by C. Since all other metrics also depend on the false negatives (recall, F_1 score) and true negatives (accuracy), this lack of information for the precision is reflected in a high uncertainty in the boundaries reconstructed.

The recall instead yields more consistently poor reconstructions of accuracy, precision and F_1 score. This means that, as a general rule, the recall itself is not a useful metric for reconstruction. On the other hand, both accuracy and, to a greater extent, the F_1 score allow for the consistent reconstruction of narrow boundaries.

C. Dataset balancedness

As already established, the balancedness of the dataset is a key factor in extrapolating the confusion matrix of a binary problem: in fact, it helps characterize the sum of true positives and false negatives. We study the quality of the reconstruction of the boundaries as the fraction of positive samples varies from 1% to 99%. The dataset size is fixed to 25,000 samples ($C = 5,000$) and the capacity of the tree (maximum depth) is set to 6. Figure 3 shows the results obtained for this analysis.

In this case, there is a clear trend that correlates the fraction of positive samples with the quality of the reconstruction (i.e. we obtain better reconstruction when the positive class is the majority one). The one exception to this trend is once again given by the precision: due to the same reasons discussed in Subsection IV-B, the high variability in reconstructed boundaries makes it difficult to make *a priori* assumptions on the quality of the results.

The linear trend of the recall is justified by the fact that the recall introduces a constraint on the true positives and false negatives, whereas true negatives and false positives (whose sum is the number of negative samples N_N) are only constrained by the relationship $tn+fp = C-N_P$. Thus, either quantity can vary from 0 to N_N, as long as the sum of the two is N_N. Because of this, the reconstruction of the ranges will depend on the size of N_N (which is linearly decreased in this study).

As with the dataset size, we once again observe that the accuracy and the F_1 score allow for a high quality reconstruction of all other metrics.

D. Model capacity

We intervene on the maximum depth that can be reached by a decision tree to limit the capacity of the model and, as a consequence, to control the performance of the classifier.

The results obtained for the reconstruction are shown in Figure 4. The behavior observed for precision and recall is similar to the one discussed in Subsection IV-B although in this case, when the model operates at full capacity, the precision can be used to build excellent reconstructions when the model is at full capacity.

Both accuracy and F_1 score show worse performance for models with lower capacity (i.e. when metrics have lower values), but improve as the capacity increases. When the model reaches its full capacity, the reconstructed range is ≈ 0 (i.e. almost perfect reconstructions).

E. Real-world datasets

In the previous experiments, we generated the datasets used to satisfy some properties of interest. Here, we present the results obtained on two real datasets. In particular, we study the performance on Adult [8] and a binary subset of ImageNet. Adult is a commonly adopted tabular dataset with 48,842 samples and a binary target label. The dataset is unbalanced, with 7,841 positive samples ($\approx$ 16%). ImageNet [9] is a famous dataset of images, with more than 1 million images belonging to 1,000 classes. For this study we extracted a subset of the entire dataset, with 2,600 images of cats[4] and 2,600 images of dogs[5]. For Adult, we trained a random forest classifier with 100 estimators, whereas for "Binary ImageNet" we used a headless ResNet50 model [10], with a fully-connected output layer and a sigmoid activation function. The output layer was fine-tuned for the specific binary classification task.

Table III presents the results obtained for both datasets. To assess the effect that rounding has on the results, we

[4]classes n02123045 (tabby cat) and n02124075 (egyptian cat)

[5]classes n02110341 (dalmatian, coach dog, carriage dog) and n02097474 (Tibetan terrier, chrysanthemum dog)

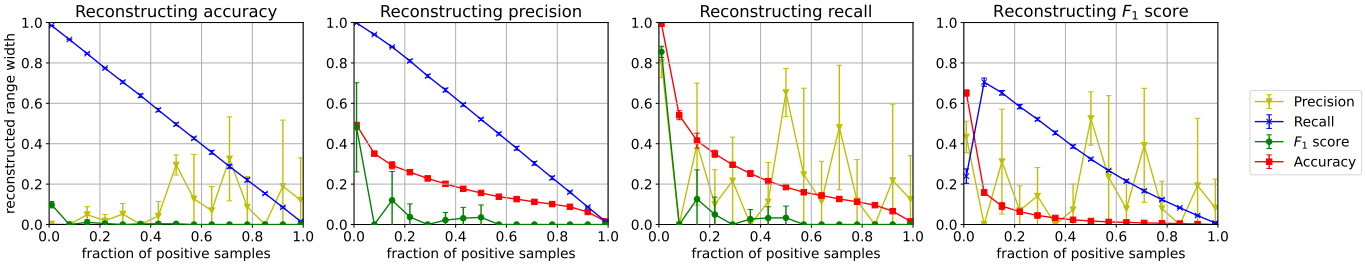

Fig. 3. Reconstructed ranges for various metrics (one for each figure), given C, N_P and either of the other metrics (specified by the lines), as the fraction of positive samples – and consequently N_P – varies. Each range is calculated as the difference between the upper and lower bounds obtained for the reconstructed metric, and is obtained as the mean over 10 separate datasets. The error bars represent $\pm$ 1 standard deviation across the 10 datasets.

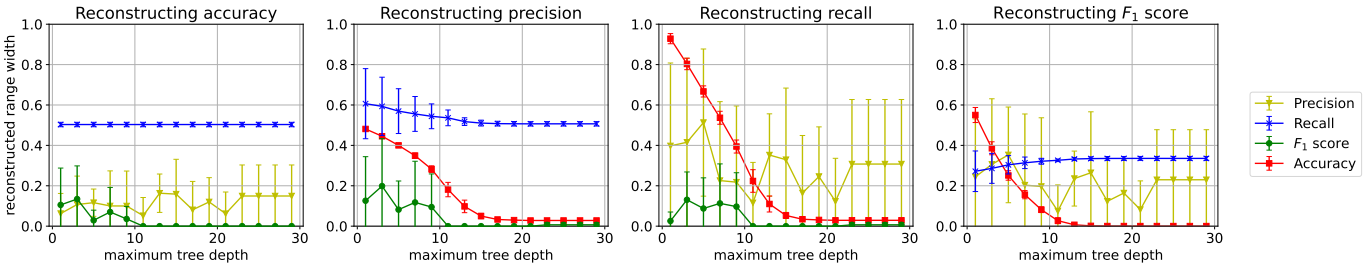

Fig. 4. Reconstructed ranges for various metrics (one for each figure), given C, N_P and either of the other metrics (specified by the lines), as the capacity of the model – and consequently the various performance metrics – varies. Each range is calculated as the difference between the upper and lower bounds obtained for the reconstructed metric, and is obtained as the mean over 10 separate datasets. The error bars represent $\pm$ 1 standard deviation across the 10 datasets.

reconstructed all metrics using both a full-precision and a rounded version of each "reconstructing" metric.

For both datasets, the results are in line with what we observed for the synthetic results. For Adult, the positive label is the minority one: the results obtained reflect the ones shown in Figure 3, for a small N_P.An aspect of interest is the behavior of the rounded reconstruction: in most cases, we obtain reconstructed ranges that are similar to the ones at full precision. However, some exceptions occur where the quality is greatly affected. We empirically observe that all such scenarios occur for cases with large variance. The mean reconstructed range widens when the metrics are rounded, but the variance of those ranges decreases. We discuss a possible reason for this in Appendix A.

The results on Binary ImageNet also reflect the ones observed on the synthetic data, for a balanced classification problem. The same effect already observed for Adult when rounding is applied is also found for this dataset.

V. Conclusions

In this paper we presented RECLAIM, a methodology to reconstruct previously unavailable performance metrics, starting from other known ones (e.g. made available by the authors of a publication), by reconstructing the underlying confusion matrix. When enough constraints are available, we can solve the problem exactly. When it is not, we can reconstruct upper and lower boundaries for the confusion matrix and, as a consequence, for other classification metrics. We empirically observed that some metrics (accuracy, F_1 score) carry more information and help produce more reliable reconstructions. With RECLAIM, we aim to provide help to researchers in comparing their performance against other state of the art models that have not been made available.

For the future works, we mainly aim to narrow down the boundaries of the reconstructed metrics, by making some assumptions on the model and the data distribution. Additionally, we are going to extend the methodology to other types of problem (e.g. multi-class classification and regression). Finally, we intend to apply RECLAIM for the detection of inconsistent results in publications, with the goal of helping improve the reliability of scientific writings.

VI. Acknowledgements

This work has been partially supported by the SmartData@PoliTO center on Big Data and Data Science.

References

[1] J. A. Heathers, J. Anaya, T. van der Zee, and N. J. Brown, "Recovering data from summary statistics: Sample Parameter Reconstruction via Iterative TEchniques (SPRITE)," PeerJ Preprints, preprint, May 2018. [Online]. Available: https://peerj.com/preprints/26968v1

[2] S. Wilner, K. Wood, and D. J. Simons, "Complete recovery of values in diophantine systems (corvids)," *PsyArXiv preprint*, 2018.

[3] N. J. L. Brown and J. A. J. Heathers, "The GRIM Test: A Simple Technique Detects Numerous Anomalies in the Reporting of Results in Psychology," *Social Psychological and Personality Science*, vol. 8, no. 4, pp. 363–369, May 2017. [Online]. Available: http://journals.sagepub.com/doi/10.1177/1948550616673876

[4] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson *et al.*, "Extracting training data from large language models," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 2633–2650.

TABLE III

RECONSTRUCTED RANGES FOR BINARY CLASSIFIERS TRAINED ON TWO SEPARATE DATASETS. EACH COLUMN REPRESENTS THE WIDTH OF THE RECONSTRUCTED RANGE FOR EACH METRIC. EACH ROW REPRESENTS ONE ADDITIONAL CONSTRAINT OTHER THAN C AND N_P USED FOR THE RECONSTRUCTION. THE "ROUNDED" VERSION OF EACH METRIC SHOWS THE RECONSTRUCTION WHEN THE METRIC IS PROVIDED AS A ROUNDED VALUE (4 SIGNIFICANT FIGURES). LOWER VALUES REPRESENT A BETTER RECONSTRUCTION (WHEN 0, THE METRIC IS RECONSTRUCTED PERFECTLY). CONFIDENCE INTERVALS OBTAINED BY RUNNING THE EXPERIMENT 10 TIMES ON DIFFERENT TRAIN/TEST SPLITS.

		Reconstructed range			
	Reconstructing metric	**Accuracy**	**Precision**	**Recall**	**F_1 score**
Adult	**Accuracy**	0.0 ± 0.0	0.3806 ± 0.0062	0.6146 ± 0.0162	0.2088 ± 0.012
	Accuracy (rounded)	0.0 ± 0.0	0.3806 ± 0.0062	0.6146 ± 0.0162	0.2088 ± 0.012
	Precision	0.0303 ± 0.0381	0.0 ± 0.0	0.3104 ± 0.3931	0.225 ± 0.2999
	Precision (rounded)	0.0903 ± 0.0066	0.0 ± 0.0	0.9118 ± 0.0505	0.6896 ± 0.0773
	Recall	0.8408 ± 0.0039	0.8949 ± 0.0038	0.0 ± 0.0	0.5857 ± 0.0059
	Recall (rounded)	0.8408 ± 0.0039	0.8949 ± 0.0038	0.0 ± 0.0	0.5857 ± 0.0059
	$\mathbf{F_1}$ score	0.0138 ± 0.0213	0.0589 ± 0.0901	0.0867 ± 0.1333	0.0 ± 0.0
	$\mathbf{F_1}$ score (rounded)	0.0757 ± 0.0048	0.4439 ± 0.0569	0.4797 ± 0.0219	0.0001 ± 0.0
Binary ImageNet	**Accuracy**	0.0 ± 0.0	0.0093 ± 0.0034	0.0093 ± 0.0035	0.0 ± 0.0
	Accuracy (rounded)	0.0 ± 0.0	0.0093 ± 0.0034	0.0093 ± 0.0035	0.0 ± 0.0
	Precision	0.162 ± 0.176	0.0 ± 0.0	0.3237 ± 0.3533	0.2587 ± 0.3175
	Precision (rounded)	0.2737 ± 0.1482	0.0 ± 0.0	0.5492 ± 0.2975	0.4329 ± 0.2793
	Recall	0.4982 ± 0.0122	0.4994 ± 0.012	0.0 ± 0.0	0.3315 ± 0.011
	Recall (rounded)	0.4982 ± 0.0122	0.4994 ± 0.012	0.0 ± 0.0	0.3315 ± 0.011
	$\mathbf{F_1}$ score	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	$\mathbf{F_1}$ score (rounded)	0.0 ± 0.0	0.0074 ± 0.003	0.0074 ± 0.0031	0.0 ± 0.0

[5] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18.

[6] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International journal of data mining & knowledge management process*, vol. 5, no. 2, p. 1, 2015.

[7] R. Penrose, "A generalized inverse for matrices," in *Mathematical proceedings of the Cambridge philosophical society*, vol. 51, no. 3. Cambridge University Press, 1955, pp. 406–413.

[8] R. Kohavi *et al.*, "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid." in *Kdd*, vol. 96, 1996, pp. 202–207.

[9] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

APPENDIX A

In this appendix, we discuss a possible reason for the observation that some reconstructed ranges that present large variance (e.g. those obtained from precision) show an increase in mean reconstructed range, but a decrease in variance of the reconstructed ranges, when rounding is applied.

To this end, we introduce some additional notation, although in a simplified and non-rigorous manner. We call $f(m) : \mathbb{R} \to \mathbb{R}$ the function that produces the width of the range given a reconstructing metric (along with C and N_P, which we consider fixed). When testing different runs, we sample a neighborhood of the underlying "true" value m^*: each sample $\hat{m}$ is an estimate of m^*. If f is reasonably "smooth" (i.e. it has low entropy) in an interval of m^*, the resulting sampling of that interval will have a low variance. On the other hand, a higher entropy f will result in samples that have a high variance (e.g. Figure 5).

The introduction of rounding means that, when evaluating $f(\hat{m})$, we are no longer considering a single point $\hat{m}$, but rather an interval of points around $\hat{m}$ (i.e. all points that, when rounded, will produce $\hat{m}$). Among these, the worst case (i.e. the largest reconstructed range) is selected, since we are allowing for all valid solutions to be found. Thus, we extract the maximum value in a neighborhood of $\hat{m}$. Applying a maximum to an interval of values acts as a low pass filter that smooths f. We show this behavior for a toy function f in Figure 5. There, we observe how the various $\hat{m}$ sampled during each run vary widely with the "full-precision" f (hence the high variance), whereas they vary significantly less when sampling the "rounding" f. We indeed observe this decrease in variance for the rounded versions. The other obvious consequence is that the mean estimated ranges will be larger for the rounded version.

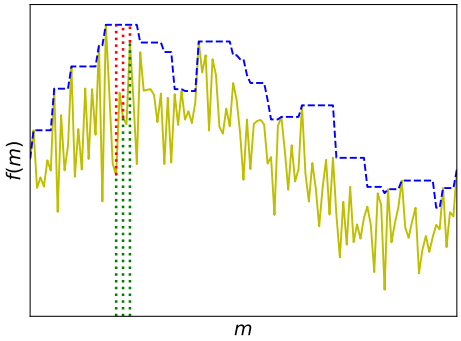

Fig. 5. Representation of a function $f(m)$ that maps an input metric m to the reconstructed range for a different metric. In yellow is the function when m is given with full precision. The dashed blue curve is the function when m is provided as a rounded value (in other words, the worst case of a neighborhood of m is selected). The vertical dotted lines represent three samples $\hat{m}$ of the input metric (e.g. obtained on 3 separate runs). The mean estimated range is lower for the full-precision f, whereas the variance will be lower for the "rounding" f