The *Tile*: A 2D Map of Ranking Scores for Two-Class Classification

Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck
Montefiore Institute, University of Liège, Liège, Belgium
{S.Pierard,Anais.Halin,Anthony.Cioppa,Adrien.Deliege,M.VanDroogenbroeck}@uliege.be

Abstract

In the computer vision and machine learning communities, as well as in many other research domains, rigorous evaluation of any new method, including classifiers, is essential. One key component of the evaluation process is the ability to compare and rank methods. However, ranking classifiers and accurately comparing their performances, especially when taking application-specific preferences into account, remains challenging. For instance, commonly used evaluation tools like Receiver Operating Characteristic (ROC) and Precision/Recall (PR) spaces display performances based on two scores. Hence, they are inherently limited in their ability to compare classifiers across a broader range of scores and lack the capability to establish a clear ranking among classifiers. In this paper, we present a novel versatile tool, named the Tile*, that organizes an infinity of ranking scores in a single 2D map for two-class classifiers, including common evaluation scores such as the accuracy, the true positive rate, the positive predictive value, Jaccard's coefficient, and all F_β scores. Furthermore, we study the properties of the underlying ranking scores, such as the influence of the priors or the correspondences with the ROC space, and depict how to characterize any other score by comparing them to the Tile. Overall, we demonstrate that the Tile is a powerful tool that effectively captures all the rankings in a single visualization and allows interpreting them.* [1]

1. Introduction

Two-class classification is a fundamental task, encountered in numerous real-world scenarios. For instance, it plays a vital role in medical diagnostics, such as blood tests, MRI scans, and other imaging techniques to determine whether a patient has a disease or is healthy. In security systems, alarms must activate only when intrusions are detected. Similarly, in quality control, identifying defects in manufactured items is crucial to ensure faulty products do not reach the market. To address these challenges, selecting the right classifier is essential. However, this requires ranking classifiers in the context of application-specific preferences. For example, in medical testing, minimizing false negatives is critical since failing to diagnose a patient could have life-threatening consequences. In security systems, the focus is often on maximizing true negatives, accepting occasional false alarms as a trade-off for ensuring safety. Meanwhile, in quality control, false positives can be costly, as they may trigger unnecessary halts in production. Each application thus has unique requirements regarding the types of errors a classifier can tolerate.

A wide range of scores penalizing different types of errors are available in the literature. However, selecting

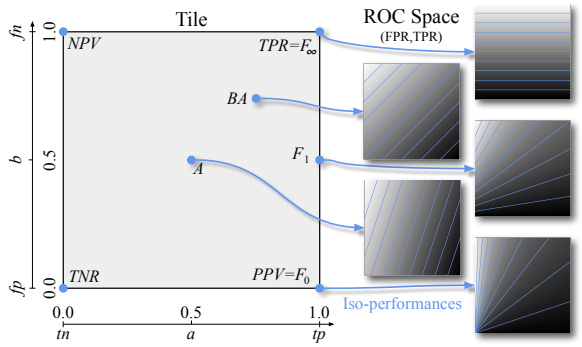


Figure 1. **Introducing the Tile.** We introduce a new visual tool, called the Tile, representing an infinite family of ranking scores to evaluate the performances of two-class classifiers at a glance. In this figure, we highlight the correspondences between specific ranking scores on the Tile and their corresponding set of iso-performance lines in the ROC space. Notably, the variation of iso-performance lines along the right border of the Tile demonstrates the limitations of the ROC space for ranking performance. This visualization illustrates how the Tile simplifies the task of ranking classifiers and enhances the interpretation of performance scores across various evaluation spaces, such as the ROC space.

[1] This paper is the second of a trilogy. In a nutshell, paper A [28] presents an axiomatic framework and an infinite family of scores for ranking classifiers. In this paper (paper B [29]), we particularize this framework to two-class classification and introduce the *Tile*, a visual tool that organizes these scores in a single 2D map. Finally, paper C [21] provides a guide to using the Tile according to four practical scenarios. For that, we present different Tile flavors that are applied to a real application.

the appropriate score to rank classifiers taking application-specific preferences into account can be challenging. Additionally, the common practice of ranking based on a single score, as done in most benchmarks, can lead to a suboptimal classifier choice. Moreover, so-called evaluation spaces combining two scores, such as Receiver Operating Characteristic (ROC) and Precision/Recall (PR) spaces, do not allow ranking classifiers. This raises a recurring question: how can classifiers be effectively ranked to best align with the specific needs of each application?

In this paper, we introduce a novel visual tool for two-class classification, called the *Tile*, which organizes an infinity of performance orderings derived from ranking scores into a two-dimensional map. Our Tile is parametrized by two parameters reflecting application-specific preferences: the first controls the trade-off between true positives and true negatives, while the second balances false positives and false negatives. This parametrization allows mapping common performance scores from the literature, such as the accuracy or F_1, as illustrated on the left side of Fig. 1. Next, we analyze the correspondences between the Tile and standard evaluation spaces such as ROC and PR, with a particular emphasis on iso-performance lines. As shown in Fig. 1, iso-performance lines can be drawn on the ROC space for each score of the Tile. Even though rankings of classifiers on the ROC space is dependent on the choice of a score, our Tile enables a straightforward, unified interpretation of the ranking of two-class classifiers at a glance. Finally, we demonstrate that our Tile's organization allows to easily rank classifiers and study some ranking properties such as the orderings induced by all ranking scores, the characterization of any score, and the robustness of ranking scores.

Contribution. We summarize our contributions as follows. **(i)** We introduce a novel visual tool, called the Tile, that organizes an infinity of ranking scores on a two-dimensional map. **(ii)** We analyze the correspondences between the Tile and common evaluation spaces such as ROC and PR, with a particular focus on iso-performance lines. **(iii)** We show that our Tile's organization allows to easily compare ranking scores, rank classifiers, and study ranking properties.

2. Preliminaries and Related Work

We first present the necessary preliminaries, including the mathematical framework, definitions of the relevant scores, their underlying structure, and a review of related work to set the context. By coherence, we adopt the mathematical framework, terminology, and notations introduced in paper A [28]. Note that all acronyms and mathematical symbols used in this paper are defined where they appear[2].

[2] A list of them is provided in the supplementary material.

2.1. Mathematical Framework

The framework is based on the probability theory. A score X is a real-valued function defined on a subset $\mathrm{dom}(X)$ of the performance space $\mathbb{P}_{(\Omega,\Sigma)}$. The latter is the set of all possible probability measures on the measurable space (Ω, Σ), where the sample space (*a.k.a.* universe) is Ω, and the event space (*a.k.a.* σ-algebra) is Σ. A performance P is an element of $\mathbb{P}_{(\Omega,\Sigma)}$, thus, a probability measure.

Moreover, the framework is also based on the order theory. The symbols $\lesssim_X$ and $\gtrsim_X$ are used to denote, respectively, the ordering and dual ordering induced by the score X. When both performances $P_1, P_2 \in \mathrm{dom}(X)$, P_1 is *worse than*, *equivalent to*, or *better than* P_2 when $X(P_1) < X(P_2)$, $X(P_1) = X(P_2)$, or $X(P_1) > X(P_2)$, respectively. When $P_1 \notin \mathrm{dom}(X)$ or $P_2 \notin \mathrm{dom}(X)$, they are *equivalent* if $P_1 = P_2$, and *incomparable* otherwise.

Two-class crisp classification. We particularize this probabilistic framework to the case in which the sample space contains two satisfying and two unsatisfying elements, thus $|\Omega| = 4$ and $\Sigma = 2^\Omega$. Using the binary random variable $S : \Omega \to \{0, 1\}$ to denote the satisfaction, we have $|\{\omega \in \Omega : S(\omega) = 0\}| = |S^{-1}(\{0\})| = 2$, and $|\{\omega \in \Omega : S(\omega) = 1\}| = |S^{-1}(\{1\})| = 2$. Variable S determines how the elements of Ω are interpreted. The most popular interpretation is undoubtedly the popular *two-class crisp classification*. For this choice, we take $\Omega = \{tn, fp, fn, tp\}$. The samples tn, fp, fn and tp are interpreted as, respectively, a *true negative*, a *false positive* (type I error, false alarm), a *false negative* (type II error, miss), and a *true positive* (hit). In that case, $S(tn) = S(tp) = 1$ and $S(fp) = S(fn) = 0$.

Classes and predictions. We can introduce the set of classes $\mathbb{C} = \{c_-, c_+\}$, as well as the random variables $Y : \Omega \to \mathbb{C}$ and $\hat{Y} : \Omega \to \mathbb{C}$ such that $Y(tn) = Y(fp) = c_-$, $Y(fn) = Y(tp) = c_+$, $\hat{Y}(tn) = \hat{Y}(fn) = c_-$, and $\hat{Y}(fp) = \hat{Y}(tp) = c_+$. Indeed, Y and $\hat{Y}$ can be interpreted as the ground-truth and predicted classes. The *no-skill performances* are those such that $P(Y, \hat{Y}) = P(Y)P(\hat{Y})$. The satisfaction is the indicator $S = \mathbf{1}_{Y=\hat{Y}}$.

2.2. Scores

In the literature, numerous performance scores[3] (also called metrics, measures, indicators, criteria, factors, and indices [10, 37]) have been introduced for two-class crisp classification. In fact, the list of scores is almost endless, as can be seen from numerous reviews [2, 10, 11, 13, 30, 36]. Choosing one score over another depends on the application field (medical, machine learning, statistics, *etc.*).

Unconditional probabilistic scores. Unconditional probabilistic scores, denoted by X_E^U, are parameterized by an

[3] In this paper we choose the term *score* to avoid any possible confusion with the mathematical meaning of the terms *metric*, *measure* and *indicator*.

event $E \in \Sigma$ such that $\emptyset \subsetneq E \subsetneq \Omega$:

$$X_E^U : \mathbb{P}_{(\Omega,\Sigma)} \to [0,1] : P \mapsto P(E)\,. \tag{1}$$

There exist only 14 unconditional probabilistic scores. For singleton events, there is $PTN = X^U_{\{tn\}}$ (called *rejection rate*), $PFP = X^U_{\{fp\}}$, $PFN = X^U_{\{fn\}}$, and $PTP = X^U_{\{tp\}}$ (called *detection rate*). There are also the *priors* for the negative and positive classes, given by $\pi_- = X^U_{\{tn,fp\}}$ and $\pi_+ = X^U_{\{fn,tp\}}$ respectively, as well as the *negative* and *positive prediction rates*, given by $\tau_- = X^U_{\{tn,fn\}}$ and $\tau_+ = X^U_{\{fp,tp\}}$ respectively. The *prevalence* is a synonym used for π_+. The *accuracy* (*a.k.a. matching coefficient*) corresponds to the expected value of S and is given by $A = X^U_{\{tn,tp\}}$. Its complement [10] is the *error rate* or *misclassification rate* .

Conditional probabilistic scores. The conditional probabilistic scores $X^C_{E_1|E_2}$ are parameterized by two events, $E_1, E_2 \in \Sigma$ such that $\emptyset \subsetneq E_1 \subsetneq E_2 \subseteq \Omega$:

$$X^C_{E_1|E_2} : \mathrm{dom}(X^C_{E_1|E_2}) \to [0,1] : P \mapsto P(E_1|E_2)\,, \tag{2}$$

with $\mathrm{dom}(X^C_{E_1|E_2}) = \{P \in \mathbb{P}_{(\Omega,\Sigma)} : P(E_2) \neq 0\}$. There exist only 50 conditional probabilistic scores (including the 14 unconditional ones, as $X^U_E = X^C_{E|\Omega}$). The probabilities of making a correct decision for negative and positive inputs are given by the *True Negative Rate* $TNR = X^C_{\{tn\}|\{tn,fp\}}$ (*a.k.a.* specificity, selectivity, inverse recall) and the *True Positive Rate* $TPR = X^C_{\{tp\}|\{fn,tp\}}$ (*a.k.a. sensitivity*, *recall*). Their complements are the *False Positive Rate* FPR and the *False Negative Rate* FNR respectively. The probabilities of negative and positive predictions being correct are given by the *Negative Predictive Value* $NPV = X^C_{\{tn\}|\{tn,fn\}}$ (*a.k.a.* inverse precision) and the *Positive Predictive Value* $PPV = X^C_{\{tp\}|\{fp,tp\}}$ (*a.k.a.* precision). Their complements are the *False Omission Rate* FOR , and the *False Discovery Rate* FDR respectively. *Jaccard's coefficient* is $J_- = X^C_{\{tn\}|\{tn,fp,fn\}}$ for the negative class and $J_+ = X^C_{\{tp\}|\{fp,fn,tp\}}$ for the positive class. The latter is also called *Tanimoto coefficient*, *similarity*, *intersection over union*, *critical success index* [23], as well as *G-measure* in [14].

Non-probabilistic scores. There is also an infinite number of scores that have no probabilistic significance.

We start with transformations of probabilistic scores. *Bennett's* S [38] is related to the accuracy by $S = 2A - 1$. The *F-one* score , also called *Dice-Sørensen coefficient*, is related to Jaccard by $F_1 = {}^{2J_+}/_{J_++1}$. The *standardized negative and predictive values* are transformations of the negative and predictive values given by $SNPV = \frac{TNR}{TNR+FNR} = \frac{NPV\pi_+}{NPV(\pi_+-\pi_-)+\pi_-}$ and $SPPV = \frac{TPR}{FPR+TPR} = \frac{PPV\pi_-}{PPV(\pi_--\pi_+)+\pi_+}$ [22]. The *likelihood ratios* [6, 17, 18, 30] are also transformations of these scores. The *Negative Likelihood Ratio* is $NLR = \frac{FNR}{TNR} = \frac{1-SNPV}{SNPV}$. The *Positive Likelihood Ratio* [1] is $PLR = \frac{TPR}{FPR} = \frac{SPPV}{1-SPPV}$. *Cohen's* κ *statistic* [12, 24] is a transformation of the accuracy: $\kappa = \frac{A-A\circ\text{no–skill}}{1-A\circ\text{no–skill}}$ where $\circ$ is the function composition operator and no–skill is the operation that transforms a performance P into P' such that $P'(Y,\hat{Y}) = P(Y)P(\hat{Y})$. It is also known as *Heidke Skill Score* [10, 40]. It is also common to combine probabilistic scores, either linearly or by averaging them. For example, the *Bias Index* BI has been defined as $\tau_+ - \pi_+$ in [9].

Many authors combine TNR and TPR. Their weighted arithmetic mean is the *weighted accuracy* WA . When the weights are 0.5, we obtain the *balanced accuracy* BA , and when the weights correspond to the priors, we get the *accuracy*. Instead of taking an arithmetic mean, it has been proposed to take the geometric mean [3, 20], which leads to the *G-measure* [10] (not to be confused with the G-measure of [14]). *Youden's index* [16, 42] or *Youden's* J_Y *statistic* is $J_Y = TNR + TPR - 1 = 2BA - 1$. It is also called *informedness* and *Peirce Skill Score* [10, 40]. The *determinant* $|C|$ [41] of the (normalized) confusion matrix is $|C| = \pi_-\pi_+J_Y$.

Some authors prefer to combine PPV with TPR. Their weighted harmonic mean is the *F-score* F_β . Others prefer to combine NPV with PPV. The *markedness* [31] is defined as $NPV + PPV - 1$ and is also known as the *Clayton Skill Score* [10, 40]. It has also been proposed to combine the 4 probabilistic scores TNR, TPR, NPV, and PPV. Their arithmetic mean is called *Average Conditional Probability* ACP [8], and their harmonic mean is P_4 [35]. Finally, a plethora of other scores can be found in the literature concerning similarity scores [4, 7, 11, 19, 26, 39]. Many of them are peculiar cases of the ranking scores that are discussed in detail in this paper.

2.3. Structuring the Scores

In [13], 18 scores have been experimentally structured in the form of histograms of linear and rank correlations, based on 30 datasets. Similarly, Choi *et al.* [11] made a hierarchical clustering of 76 binary similarity and distance scores based on a random binary dataset. In [26], 7 properties have been arbitrarily defined, and 11 scores have been classified into 5 classes based on whether the properties are verified. In [37], taking the object-oriented software development standpoint, the structure takes the form of a *Unified Modeling Language* (UML) diagram to represent the concepts of measure, metric, and indicator, as well as the relationships between the three concepts. [10] proposes to distinguish between only two types of scores: the measures and the metrics. Based on a sample of 44 scores and related quantities (22 measures and 22 metrics), they propose to divide measures into 4 levels (base, 1st, 2nd and 3rd), and metrics into 3 levels

(base, 1st and 2nd). The 4 basic measures are the elements of the confusion matrix (also known as the contingency matrix). Based on this, the authors draw what they describe as "the periodic table of elements in binary classification performance". It is a map of the 44 scores, organized around the confusion matrix, the vertical dimension being related to the ground-truth class and the horizontal dimension to the predicted class. In this work, we avoid mixing scores and only consider those suitable for performance ordering and performance-based ranking of classifiers.

2.4. Reminders on the Axioms of Ranking

We also leverage the axiomatic definitions around the notion of performance introduced in paper A [28]. In a nutshell, the 1st axiom states that if several performances have been ranked, then removing or adding a performance should not affect the relative order of the previously present performances. We will reuse the second and third axioms later in this paper, rephrased as follows.

Axiom 2 (reminder)**.** *If the degree of satisfaction that can be obtained with a 1st classifier is for sure less or equal than the degree of satisfaction that can be obtained with a 2nd one, then the former classifier is not better than the latter.*

Axiom 3 (reminder)**.** *Given a set of classifiers, and a arbitrary set of possible operations to perturb (*e.g.*, add noise to their output) and/or combine them, then, on the basis of their performances only, it must be impossible to determine a sequence of these operations that would lead with certainty to a classifier either better than the best of the initial ones or worse than the worst of the initial ones.*

Moreover, we show for the first time that it is possible to establish a continuous, two-dimensional map of an infinity of scores, and that the map includes many scores that are widespread in the literature.

3. Ranking Scores for Two-Class Classification

3.1. Particularization of Ranking Scores

To build the Tile, we consider the family of ranking scores in the special case of the two-class crisp classification task:

$$R_I(P) = \frac{\sum_{\omega \in \{tn,tp\}} I(\omega) P(\{\omega\})}{\sum_{\omega \in \{tn,fp,fn,tp\}} I(\omega) P(\{\omega\})} \tag{3}$$

where $I : \Omega \to \mathbb{R}_{\geq 0}$ is a non-negative random variable, called *importance*, such that $I(tn) + I(tp) \neq 0$ and $I(fp) + I(fn) \neq 0$. These scores allow for the comparison of all two-class classification performances in $\mathrm{dom}(R_I)$, even when the classes are imbalanced ($\pi_- \neq \pi_+$).

Example 1 (Probabilistic ranking scores)**.** *The family of ranking scores includes the following* 9 *probabilistic scores:* NPV, $X^C_{\{tn,tp\}|\{tn,fn,tp\}}$, TPR, J_-, A, J_+, TNR, $X^C_{\{tn,tp\}|\{tn,fp,tp\}}$, *and* PPV.

Example 2 (F-scores)**.** *For all $\beta \geq 0$, F_β is a ranking score.*

Example 3 (PABDC)**.** *The ranking scores for which the importance values $I(\omega)$ are rational numbers correspond to the class of* Presence/Absence Based Dissimilarity Coefficients *(PABDC) satisfying the first* 9 *properties listed in [5] (see Prop. 1 in that paper).*

3.2. Properties

We start by examining the effect of the target/prior shift operation [34], denoted by $\mathrm{shift}_{\pi \to \pi'}$. It transforms a performance P into P' such that $P'(E) = P(E)\frac{\pi'_-}{\pi_-}$ for all $E \in 2^{\{tn,fp\}}$ and $P'(E) = P(E)\frac{\pi'_+}{\pi_+}$ for all $E \in 2^{\{fn,tp\}}$.

Property 1. *The composition of a ranking score with a target/prior shift operation is a ranking score. We have $R_I \circ \mathrm{shift}_{\pi \to \pi'} = R_{I'}$ with $I'(\omega) = I(\omega)\frac{\pi'_-}{\pi_-}$ for all $\omega \in \{tn, fp\}$ and $I'(\omega) = I(\omega)\frac{\pi'_+}{\pi_+}$ for all $\omega \in \{fn, tp\}$.*

As we have particularized the axiomatic framework to the particular case of two-class crisp classification, the following property holds.

Property 2. *Multiplying both $I(tn)$ and $I(tp)$, or both $I(fp)$ and $I(fn)$, by the same strictly positive constant leads to another ranking score that is monotonously increasing with the original one. The induced performance ordering is thus unaffected by such a transformation.*

Thanks to the last property, we can get rid of the redundancy between the performance orderings induced by the ranking scores by focusing on the *canonical ranking scores*.

Definition 1. *Canonical ranking scores are given by*

$$R_{I_{a,b}} = \frac{(1-a)PTN + aPTP}{(1-a)PTN + (1-b)PFP + bPFN + aPTP},$$

where $a, b \in [0, 1]$ and $I_{a,b}$ is the importance given by $I_{a,b}(tn) = 1 - a$, $I_{a,b}(fp) = 1 - b$, $I_{a,b}(fn) = b$, and $I_{a,b}(tp) = a$.

Consequently, the NPV, PPV, TNR, TPR, A, and F_β scores belong to the canonical ranking scores.

To avoid potential issues, it should be noted that averaging ranking scores cannot be done without precautions. In fact, there are only a few special cases in which a score obtained in this way can be used for ranking. The same precaution also applies for canonical ranking scores. For example, because of this issue, the orderings induced by ACP, P_4, and $VUT = \int_0^1 \int_0^1 R_{I_{a,b}}\, da\, db$ are incompatible with the axioms of ranking. If a compromise has to be found between several canonical ranking scores, we recommend averaging the importances rather than the scores.

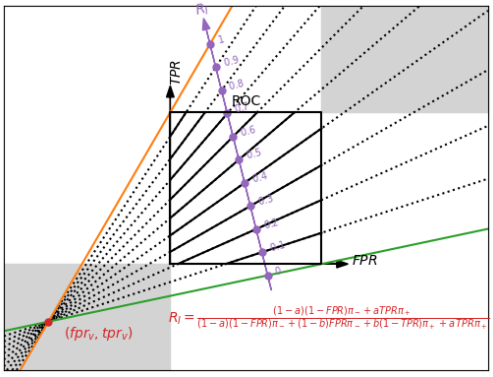
3.3. Correspondences with the ROC Space

We now study the correspondences between the canonical ranking scores and the *Receiver Operating Characteristic* (ROC) space (*i.e.*, $FPR \times TPR$) for fixed priors.

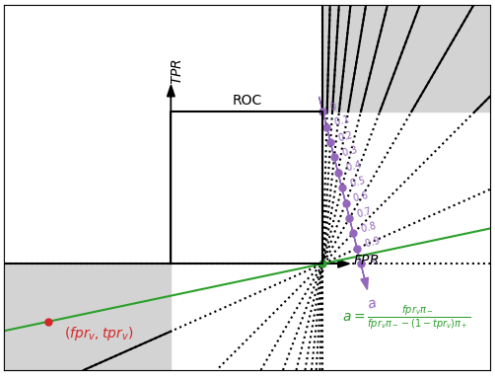
Three pencils of lines are depicted in Fig. 2. The first pencil (see Fig. 2a) can be used to read the value of $R_{I_{a,b}}$ in any point of ROC (on any line of slope $-\pi_-/\pi_+$, in purple in the figure), when the location of the red point is known. The red point corresponds to the locus of iso-performance lines for a given ranking score. It corresponds to the intersection of the green and orange lines, that is, the iso-performance lines for which the score is equal to 0 and 1, respectively. The equations of these lines can be obtained by looking at other pencils, as following. The second pencil (see Fig. 2b) can be used to find the green line based on the value of the parameter a (the green line is vertical when $a = 0$ and horizontal when $a = 1$). Finally, the third pencil (see Fig. 2c) can be used to find the orange line based on the value of the parameter b (the orange line is vertical when $b = 0$ and horizontal when $b = 1$). This procedure generalizes the geometric constructions provided by Flach in [14] for A, F_1, and PPV.

Contour plots. Figure 2a can be considered as a *contour plot* depicting the preorder induced by a score. The depicted curves (lines in our case) are called *iso-performance lines* in [32] and *isometrics* in [14]. Applying any monotonic function to the score leaves these curves unchanged. Hence, a line corresponds to a set of *equivalent* performances. All performances that are on the top-left side of the line are *better than* them, and all performances that are on the bottom-right side of the line are *worse than* them. Note that this geometric analysis is not peculiar to the canonical ranking scores. It is valid for all ranking scores, as for any ranking score there exists a canonical ranking score such that the orderings induced by them are equal.

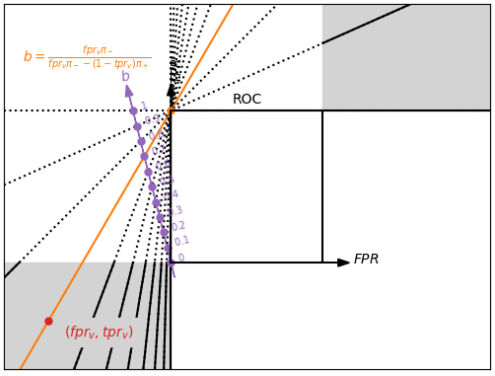
Interpretation with respect to Axiom 2. It is also interesting to note that, for any I, the red point is in one of the gray areas (extending to infinity). When $a < b$, the red point is on the right and above ROC, and when $a > b$, it is on the left and below ROC. When $a = b$, all lines in Fig. 2a (including the green and orange ones) are parallel (*e.g.*, for TNR, A, and TPR) and the red point is a point at infinity. This is related to Axiom 2. Indeed, in the case of fixed priors, the performance orderings induced by any score that has a pencil of iso-performance lines is a performance ordering that can be induced by a ranking score. This is because the family of ranking scores covers all cases where the vertex of the pencil is in one of the gray areas. Putting the vertex outside these areas would be illogical as there would either be better performances than the one located in the upper-left corner of ROC, or worse than the one located in the lower-right corner of ROC, which is prohibited by Axiom 2.



(a) The locus of equivalent performances (*i.e.*, those for which the ranking score takes a given value) are lines (restricted to ROC). These lines form a pencil whose vertex (fpr_v, tpr_v) (in red) is located outside the ROC space, either in the bottom-left or upper-right areas (in gray). The value taken by the score varies linearly along any line of slope $-\pi_-/\pi_+$ (in purple).



(b) The locus of the vertices (fpr_v, tpr_v) (red point) for all ranking scores with a given value of a are lines (restricted to the gray areas). These lines form a pencil whose vertex is located at the bottom-right corner of ROC. The value of a varies linearly along any line of slope $-\pi_-/\pi_+$ (in purple).



(c) The locus of vertices (fpr_v, tpr_v) (red point) for all ranking scores with a given value of b are lines (restricted to the gray areas). These lines form a pencil whose vertex is located at the top-left corner of ROC. The value of b varies linearly along any line of slope $-\pi_-/\pi_+$ (in purple).

Figure 2. The geometry of the ranking scores $R_{I_{a,b}}$ in the ROC plane (FPR, TPR). Example given for the class priors $\pi_+ = 1 - \pi_- = 0.2$ and the importance given by $(a, b) = (0.95, 0.7)$.

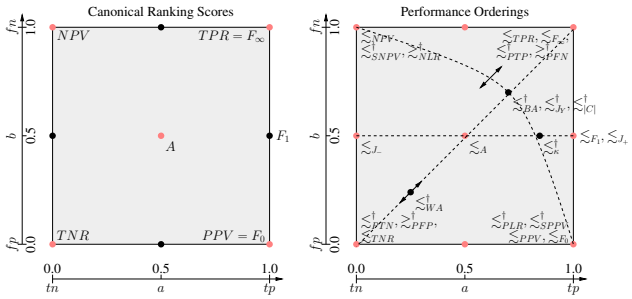


Figure 3. Placement of the canonical ranking scores (left) and of some performance orderings (right) on the Tile. The symbol † indicates the orderings that are specific for given priors. For the orderings whose locations are prior-dependent, we arbitrarily chose a negative prior of 0.7. Double arrows $\leftrightarrow$ indicate the direction in which $\lesssim_{WA}$ moves when the weights are tuned and how the curve on which $\lesssim_{BA}$ and $\lesssim_{\kappa}$ moves when the priors are tuned. The colored points correspond to probabilistic scores.

Interpretation with respect to Axiom 3. As the ROC space is, for fixed priors, a linear projection of $\mathbb{P}_{(\Omega,\Sigma)}$, any convex in the ROC space corresponds to a convex in $\mathbb{P}_{(\Omega,\Sigma)}$. Therefore, a convex combination of performances cannot be better than the best of the combined performances and cannot be worse than the worst of the combined performances.

On the diversity. The performance orderings induced by the scores $R_{I_{a,b}}$ are all different. For two different ranking scores, one has different pencil vertices (red points), leading to different iso-performance lines, sets of equivalent performances, and performance orderings. Hence, there is no redundancy in the canonical ranking scores.

4. The Tile for Two-Class Classification

The Tile, which is depicted in Fig. 3, is defined as follows.

Definition 2. *The Tile is the mapping* $[0,1]^2 \to \mathbb{X}_{(\Omega,\Sigma)} : (a,b) \mapsto R_{I_{a,b}}$, *where* $\mathbb{X}_{(\Omega,\Sigma)}$ *denotes the set of all scores.*

Although attempts to organize a selection of two-class classification scores in the 2D plane are common, to our knowledge, this is the first time that it is done mathematically, quantitatively, and automatically, without the intervention of a human expert. The Tile is not limited to the spatial organization of scores, however.

4.1. Canonical Ranking Scores on the Tile

The left-hand side of Fig. 3 shows the layout of the Tile with some canonical ranking scores on it. Two opposite corners correspond to NPV and PPV, and the other corners to TNR and TPR. The accuracy A is in the center. These are the only 5 canonical ranking scores that are also probabilistic scores. F_β scores are between TPR and PPV ($\beta = \sqrt{b/1-b}$), F_1 being in the middle of the right side.

Interpolations. By construction, the canonical scores can be interpolated, in the Tile, as follows:

- horizontally: with the f-mean such that $f : x \mapsto x^{-1}$;
- vertically: with the f-mean such that $f : x \mapsto (1-x)^{-1}$.

The fact that F_1 (the harmonic mean between TPR and PPV) appears at the mid-position between TPR and PPV is a consequence of this property.

Indistinguishable samples. The scores that can be calculated when the two unsatisfying outcomes (fp and fn) are grouped together (*i.e.*, $\Omega = \{tn, tp, incorrect\}$) are those for which $I(fp) = I(fn)$. They are located on the median horizontal, passing through A. Likewise, the scores that can be calculated when two satisfying outcomes (tn and tp) are grouped together (*i.e.*, $\Omega = \{fp, fn, correct\}$) are those for which $I(tn) = I(tp)$. They are located on the median vertical, passing through A.

Operations on performances. Some remarkable geometric properties of the Tile are related to the operations that can be applied on performances. They are given hereafter.

- Changing either the predicted or ground-truth class amounts to, respectively, flipping the Tile around the raising diagonal (TNR-TPR axis) or falling diagonal (NPV-PPV axis), and complementing the scores to 1.
- Swapping the predicted and ground-truth classes is equivalent to vertical mirroring.
- Swapping the positive and negative classes is equivalent to applying a central symmetry.
- The target/prior shift operation [34] moves the performance orderings on the Tile. We found it very useful, when priors are fixed, to represent the displacement that would have occurred if we had started with uniform priors and applied the target/prior shift.

4.2. Performance Orderings on the Tile

We have put some performance orderings on the right-hand side of Fig. 3 on the Tile. For a position (a,b), we have mentioned the orderings induced by the canonical ranking score $R_{I_{a,b}}$, and all the other orderings that are equal to it, either on $\mathbb{P}_{(\Omega,\Sigma)}$ or on a subset of it (fixed priors).

With all performances. The ordering $\lesssim_{R_I}$ corresponds to location $(a,b) = (\frac{I(tp)}{I(tn)+I(tp)}, \frac{I(fp)}{I(fn)+I(fp)})$. Those induced by the 9 probabilistic ranking scores given in Example 1 are depicted by colored points on Fig. 3. While it was not possible to place J_- and J_+ on the Tile of canonical ranking scores, the corresponding orderings $\lesssim_{J_-}$ and $\lesssim_{J_+}$ are on the Tile of performance orderings. The well-known fact that F_1 and J_+ lead to the same ordering [14] can be easily visualized on the Tile since $\lesssim_{F_1}$ and $\lesssim_{J_+}$ are at the same place. All orderings induced by the similarity coefficients of the T_θ and S_θ families, defined in [19], are equal to $\lesssim_{F_1}$ and $\lesssim_A$, respectively. The orderings induced by the family of similarity coefficients defined in [4] are the ones on the

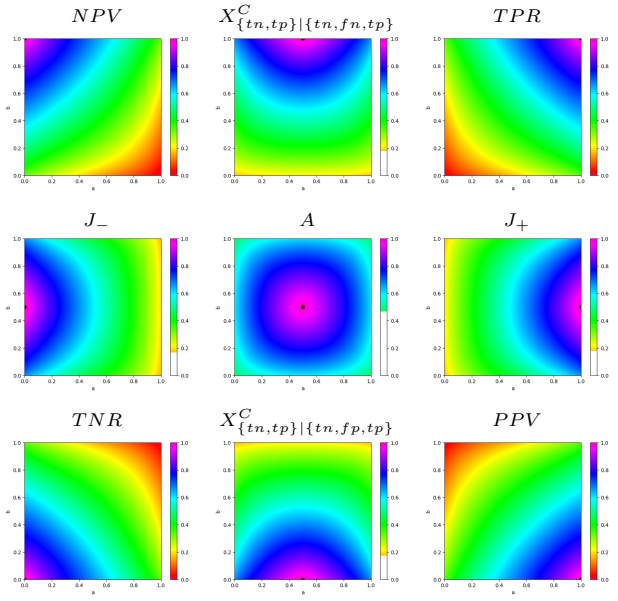
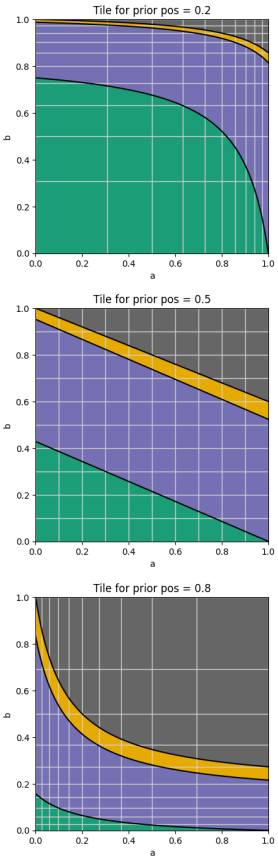


Figure 4. Tiles showing the rank correlations (Kendall τ) between 9 probabilistic scores (those that belong to the ranking scores, as given in Example 1), and all ranking scores, for a uniform distribution of performances. The correlation values have been estimated based on 10,000 performances drawn at random.

line segment between $\lesssim_A$ and $\lesssim_{F_1}$.

With fixed priors. It is also possible to place, on the Tile, the orderings that are equal to $\lesssim_{R_{I_{a,b}}}$ for a given subset of performances only. An important practical case is when the priors are fixed. In this case, the orderings induced by the unconditional probabilistic scores PTN, PFP (dual order), PFN (dual order), and PTP can be placed on the Tile. Also, the standardized negative predictive value $SNPV$, the negative likelihood ratio NLR (dual order), the standardized positive predictive value $SPPV$, the positive likelihood ratio PLR, the weighted accuracy $WA = \lambda_- TNR + \lambda_+ TPR$, the balanced accuracy BA, the score $|C|$, Youden's index J_Y, and Cohen's κ can be placed on the Tile. Some of them have a fixed position, while for others, the position depends on the priors: $\lesssim_{WA}$, $\lesssim_{BA}$ and $\lesssim_{J_Y}$ sweep the ascending diagonal, while $\lesssim_\kappa$ sweeps the median horizontal. The performance ordering $\lesssim_{WA}$ for WA is at $(a,b) = (\frac{\lambda_+\pi_-}{\lambda_+\pi_- + \lambda_-\pi_+}, \frac{\lambda_+\pi_-}{\lambda_+\pi_- + \lambda_-\pi_+})$, $\lesssim_{BA}$ is at $(a,b) = (\pi_-, \pi_-)$, and $\lesssim_\kappa$ is at $(a,b) = (\frac{\pi_-^2}{\pi_-^2+\pi_+^2}, \frac{1}{2})$.

Characterizing scores. The Tile can be used to characterize any score, showing the rank correlations between that score and all canonical ranking scores, for a given performance distribution. For example, Fig. 4 shows the results obtained with the 9 probabilistic scores that belong to the ranking scores (see Example 1), for a uniform distribution of performances (*i.e.*, a symmetric Dirichlet distribution with all concentration parameters set to one). For this distribution, these rank correlations are either null (NPV with PPV, TNR with TPR) or positive. Such an analysis can be easily performed with any score; the Tile turns out to be a very practical visualization tool to gain intuition about the behavior of the plethora of scores that exist.



	performances			
	P_- ●	P_1 ●	P_2 ●	P_+ ●
PTN	0.80	0.56	0.40	0.00
PFP	0.00	0.24	0.40	0.80
PFN	0.20	0.06	0.04	0.00
PTP	0.00	0.14	0.16	0.20

	performances			
	P_- ●	P_1 ●	P_2 ●	P_+ ●
PTN	0.50	0.35	0.25	0.00
PFP	0.00	0.15	0.25	0.50
PFN	0.50	0.15	0.10	0.00
PTP	0.00	0.35	0.40	0.50

	performances			
	P_- ●	P_1 ●	P_2 ●	P_+ ●
PTN	0.20	0.14	0.10	0.00
PFP	0.00	0.06	0.10	0.20
PFN	0.80	0.24	0.16	0.00
PTP	0.00	0.56	0.64	0.80

Figure 5. Toy examples showing on the Tile which of the 4 performances P_-, P_1, P_2, and P_+ is the best. The 3 examples differ in the class priors (either 0.2, 0.5, or 0.8 for the positive class). In all examples, P_- (●) is the performance of classifiers predicting always the negative class, P_1 (●) is such that $TNR(P_1) = 0.7$ and $TPR(P_1) = 0.7$, P_2 (●) is such that $TNR(P_2) = 0.5$ and $TPR(P_2) = 0.8$, and P_+ (●) is the performance of classifiers predicting always the positive class.

Visualizing the robustness. The importance values $I(\omega)$ are design choices for competitions. Several recent papers [25, 27] have alerted the scientific community about the necessary robustness: the performance-based rankings should not vary much when parameters are slightly perturbed. Figure 4 makes clear that, for a uniform distribution of performances, the performance orderings do not vary much when the parameters a and b are slightly perturbed.

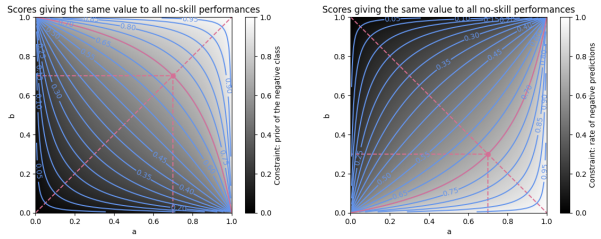


Figure 6. In the Tile, the ranking scores that put all no-skill performances on an equal footing are along a curve γ_π between NPV (upper-left corner) and PPV (lower-right corner) when the class priors $P(Y)$ are fixed, and along a curve γ_τ between TNR (lower-left corner) and TPR (upper-right corner) when the rates of predictions $P(\hat{Y})$ are fixed. The pink curves correspond to the constraints $\pi_- = 0.7$ (on the left) and $\tau_- = 0.7$ (on the right).

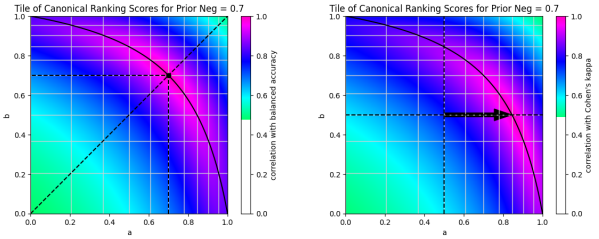


Figure 7. Tiles showing the Kendall rank correlation coefficient τ for BA (left) and κ (right), for a uniform distribution of performances P such that $\pi_- = 0.7$. The correlation values have been estimated based on $10,000$ performances drawn at random.

4.3. Rankings on the Tile

For a given set of classifiers to rank, one can use the Tile to show which classifier is ranked first, second, third, *etc*., according to $\lesssim_{R_{I_{a,b}}}$, in (a, b). This is shown in Fig. 5 with a toy example. When $\pi_- = \pi_+$, the regions where the different classifiers are ranked first are convex polygons. When $\pi_- \neq \pi_+$, the borders between these regions are curved.

4.4. More About No-Skill Performances

How can we rank no-skill performances ex aequo? The ranking scores allow ranking all no-skill performances (*i.e*., those for which the groundtruth and predicted classes are independent) ex aequo when some constraints on the compared performances are added. The more common constraint is undoubtedly that the priors are fixed. In this case, the canonical ranking scores that put the no-skill performances on the same footing are located on the curve $\gamma_\pi : \pi_+^2\, a\, b = \pi_-^2\, (1-a)\, (1-b)$. Another interesting constraint is that the rates of predictions are fixed. By symmetry, the canonical ranking scores that put the no-skill performances on the same footing are located on the curve $\gamma_\tau : \tau_+^2\, a\, (1-b) = \tau_-^2\, (1-a)\, b$. The γ_π and γ_τ curves are depicted in Fig. 6.

A new look at the balanced accuracy BA and Cohen's kappa κ. Figure 7 shows the rank correlations for both BA (on the left) and κ (on the right). We can see that BA is perfectly correlated with the ranking scores at the intersection between the curve γ_π and the rising diagonal, which is at (π_-, π_-), and that κ is perfectly correlated with the ranking scores at the intersection between the curve γ_π and the median horizontal, which is at $(\frac{\pi_-^2}{\pi_-^2+\pi_+^2}, \frac{1}{2})$.

Correction for chance. The idea of correcting a score for what can be achieved by chance is common in the literature. Scott [33] and Fleiss [15] proposed a correction for accuracy. Cohen [12] proposed another correction for it, that differs in what is considered to be achievable by chance. We noticed that Scott's π and Fleiss's κ do not satisfy the axioms of ranking, even in the case of fixed priors. Cohen's correction, on the other hand, allows nothing more than what we can do with ranking scores, with fixed priors: correcting $R_{I_{a,b}}$ in the same way as Cohen [12] did with A, that is $\frac{R_{I_{a,b}} - R_{I_{a,b}} \circ \text{no-skill}}{1 - R_{I_{a,b}} \circ \text{no-skill}}$ leads to a score that is perfectly rank-correlated with $R_{I_{a',b'}}$, where $a' = \frac{\pi_-^2\,(1-b)}{\pi_-^2\,(1-b)+\pi_+^2\,b}$ and $b' = b$. Geometrically, an entire horizontal line of the Tile is crushed into a single point (the intersection it has with the γ_π curve). There is thus an enormous loss of diversity when applying Cohen's correction to our scores.

5. Conclusion

In this paper, we presented the *Tile*, which organizes ranking scores for two-class classification task, according to the random variable called *importance* capable of considering application-specific preferences. The scores organized on this Tile are called the *canonical ranking scores* and include well-known scores, such as the accuracy or F_β scores. These canonical ranking scores lead to performance orderings that are different in each point. The Tile is a visual tool that can be used in different ways. It can be used to establish correspondences of the Tile with the ROC space and to show how to read the values taken by the canonical ranking scores in any point of the ROC space. We also showed how to use the Tile to (1) study the behavior of any score, by depicting rank correlations between a score and all ranking scores, (2) rank classifiers, using a toy example, (3) study the influence of priors on the ranking scores, (4) study properties of no-skill performances, and (5) clarify what the balanced accuracy and Cohen's kappa are. In summary, the Tile offers a comprehensive visual framework for ranking two-class classifiers, empowering researchers and practitioners to make informed, application-specific decisions, and ultimately driving advancements in the field of machine learning.

Acknowledgments. The work by S. Piérard and A. Halin was supported by the Walloon Region (Service Public de Wallonie Recherche, Belgium) under grant n°2010235 (ARIAC by DIGITALWALLONIA4.AI). A. Deliège is funded by the F.R.S.-FNRS under project grant T.0065.22. A. Cioppa is funded by the F.R.S.-FNRS.

References

[1] Douglas G. Altman and Martin Bland. Diagnostic tests 2: Predictive values. *Br. Medical J.*, 309(6947):102–102, 1994. 3

[2] Davide Ballabio, Francesca Grisoni, and Roberto Todeschini. Multivariate comparison of classification performance measures. *Chemom. Intell. Lab. Syst.*, 174:33–44, 2018. 2

[3] Ricardo Barandela, Josep Salvador Sánchez, Vicente García, and Erick Rangel. Strategies for learning in class imbalance problems. *Pattern Recognit.*, 36(3):849–851, 2003. 3

[4] Ildar Z. Batyrshin, Nailya Kubysheva, Valery Solovyev, and Luis A. Villa-Vargas. Visualization of similarity measures for binary data and 2x2 tables. *Comput. Y Sist.*, 20(3):345–353, 2016. 3, 6

[5] Forrest B. Baulieu. A classification of presence/absence based dissimilarity coefficients. *J. Classif.*, 6(1):233–246, 1989. 4

[6] Christopher D. Brown and Herbert T. Davis. Receiver operating characteristics curves and related decision measures: A tutorial. *Chemom. Intell. Lab. Syst.*, 80(1):24–38, 2006. 3

[7] Michael Brusco, J. Dennis Cradit, and Douglas Steinley. A comparison of 71 binary similarity coefficients: The effect of base rates. *PLOS ONE*, 16(4):e0247751, 2021. 3

[8] Moisès Burset and Roderic Guigó. Evaluation of gene structure prediction programs. *Genomics*, 34(3):353–367, 1996. 3, 14

[9] Ted Byrt, Janet Bishop, and John B. Carlin. Bias, prevalence and kappa. *J. Clin. Epidemiology*, 46(5):423–429, 1993. 3, 14

[10] Gurol Canbek, Seref Sagiroglu, Tugba Taskaya Temizel, and Nazife Baykal. Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. In *Int. Conf. Comput. Sci. Eng. (UBMK)*, pages 821–826, Antalya, Turkey, 2017. 2, 3

[11] Seung-Seok Choi, Sung-Hyuk Cha, and Charles C. and Tappert. A survey of binary similarity and distance measures. *J. Syst. Cybern. Informatics*, 8(1):43–48, 2010. 2, 3

[12] Jacob Cohen. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, 20(1):37–46, 1960. 3, 8

[13] Cèsar Ferri, José Hernández-Orallo, and Ramona Modroiu. An experimental comparison of performance measures for classification. *Pattern Recognit. Lett.*, 30(1):27–38, 2009. 2, 3

[14] Peter Flach. The geometry of ROC space: Understanding machine learning metrics through ROC isometrics. In *Int. Conf. Mach. Learn. (ICML)*, pages 194–201, Washington, DC, USA, 2003. 3, 5, 6, 14

[15] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychol. Bull.*, 76(5):378–382, 1971. 8

[16] Ronen Fluss, David Faraggi, and Benjamin Reiser. Estimation of the Youden index and its associated cutoff point. *Biom. J.*, 47(4):458–472, 2005. 3

[17] Ian A. Gardner and Matthias Greiner. Receiver-operating characteristic curves and likelihood ratios: improvements over traditional methods for the evaluation and application of veterinary clinical pathology tests. *Veterinary Clin. Pathol.*, 35(1):8–17, 2006. 3

[18] Afina S. Glas, Jeroen Lijmer, Martin H. Prins, Gouke Bonsel, and Patrick M. M. Bossuyt. The diagnostic odds ratio: a single indicator of test performance. *J. Clin. Epidemiology*, 56(11):1129–1135, 2003. 3

[19] John C. Gower and Pierre Legendre. Metric and euclidean properties of dissimilarity coefficients. *J. Classif.*, 3(1):5–48, 1986. 3, 6

[20] Xinjian Guo, Yilong Yin, Cailing Dong, Gongping Yang, and Guangtong Zhou. On the class imbalance problem. In *Int. Conf. Nat. Comput.*, pages 192–201, Jinan, China, 2008. 3

[21] Anaïs Halin, Sébastien Piérard, Anthony Cioppa, and Marc Van Droogenbroeck. A hitchhiker's guide to understanding performances of two-class classifiers. *arXiv*, abs/2412.04377, 2024. 1

[22] Thomas F. Heston. Standardizing predictive values in diagnostic imaging research. *J. Magn. Reson. Imaging*, 33(2): 505–505, 2011. 3, 14

[23] Robin J. Hogan, Christopher A. T. Ferro, Ian T. Jolliffe, and David B. Stephenson. Equitability revisited: Why the "equitable threat score" is not equitable. *Weather. Forecast.*, 25 (2):710–726, 2010. 3, 14

[24] Uzay Kaymak, Arie Ben-David, and Rob Potharst. The AUK: A simple alternative to the AUC. *Eng. Appl. Artif. Intell.*, 25(5):1082–1089, 2012. 3

[25] Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P. Bradley, Aaron Carass, Carolin Feldmann, Alejandro F. Frangi, Peter M. Full, Bram van Ginneken, Allan Hanbury, Katrin Honauer, Michal Kozubek, Bennett A. Landman, Keno März, Oskar Maier, Klaus Maier-Hein, Bjoern H. Menze, Henning Müller, Peter F. Neher, Wiro Niessen, Nasir Rajpoot, Gregory C. Sharp, Korsuk Sirinukunwattana, Stefanie Speidel, Christian Stock, Danail Stoyanov, Abdel Aziz Taha, Fons van der Sommen, Ching-Wei Wang, Marc-André Weber, Guoyan Zheng, Pierre Jannin, and Annette Kopp-Schneider. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat. Commun.*, 9(1):1–13, 2018. 7

[26] Ivan Ramirez Mejia and Ildar Batyrshin. Towards a classification of binary similarity measures. In *Adv. Soft Comput.*, pages 325–335. Springer Int. Publ., 2018. 3

[27] Tran Thien Dat Nguyen, Hamid Rezatofighi, Ba-Ngu Vo, Ba-Tuong Vo, Silvio Savarese, and Ian Reid. How trustworthy are performance evaluations for basic vision tasks? *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(7):8538–8552, 2023. 7

[28] Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck. Foundations

of the theory of performance-based ranking. *arXiv*, abs/2412.04227, 2024. 1, 2, 4, 11, 12, 13, 23

[29] Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck. The Tile: A 2D map of ranking scores for two-class classification. *arXiv*, abs/2412.04309, 2024. 1

[30] David Martin Ward Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. *J. Mach. Learn. Technol.*, 2(1):37–63, 2011. 2, 3

[31] David M. W. Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv*, abs/2010.16061, 2020. 3

[32] Foster Provost and Tom Fawcett. Robust classification for imprecise environments. *Mach. Learn.*, 42(3):203–231, 2001. 5

[33] William A. Scott. Reliability of content analysis: The case of nominal scale coding. *Public Opin. Q.*, 19(3):321–325, 1955. 8

[34] Tomas Sipka, Milan Sulc, and Jiri Matas. The hitchhiker's guide to prior-shift adaptation. In *IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, pages 2031–2039, Waikoloa, HI, USA, 2022. 4, 6, 13

[35] Mikołaj Sitarz. Extending F1 metric, probabilistic approach. *Adv. Artif. Intell. Mach. Learn.*, 3(2):1025–1038, 2023. 3, 14

[36] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Inf. Process. & Manag.*, 45(4):427–437, 2009. 2

[37] Putnam P. Texel. Measure, metric, and indicator: An object-oriented approach for consistent terminology. *Proc. IEEE Southeastcon*, pages 1–5, 2013. 2, 3

[38] Matthijs J. Warrens. The effect of combining categories on Bennett, Alpert and Goldstein's S. *Stat. Methodol.*, 9(3): 341–352, 2012. 3

[39] Matthijs J. Warrens. A comparison of multi-way similarity coefficients for binary sequences. *Int. J. Res. Rev. Appl. Sci.*, 16(1):64–75, 2013. 3

[40] Daniel S. Wilks. *Statistical methods in the atmospheric sciences*. Elsevier, fourth edition, 2020. 3, 14

[41] Matthias Wimmer, Bernd Radig, and Michael Beetz. A person and context specific approach for skin color classification. In *IEEE Int. Conf. Pattern Recognit. (ICPR)*, pages 39–42, Hong Kong, China, 2006. 3

[42] William John Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950. 3, 14

A. Supplementary Material

Contents

A.1. List of symbols . . . 12
A.1.1 Mathematical Symbols . . . 12
A.1.2 Symbols related to our mathematical framework of paper A [28] . . . 12
A.1.3 Symbols used for operations on performances . . . 13
A.1.4 Symbols used in the performance ordering and performance-based ranking theory . . . 13
A.1.5 Symbols used for the particular case of two-class crisp classifications . . . 13
A.2. Supplementary material about Sec. 4.1 . . . 15
A.2.1 Operations on performances . . . 15
A.2.2 When the priors are fixed: superimposing a grid on the Tile . . . 17
A.3. Supplementary material about Sec. 4.2 . . . 18
A.3.1 Placing performance orderings induced by probabilistic scores on the Tile . . . 18
A.3.2 Placing performance orderings induced by the scores F_β on the Tile . . . 19
A.3.3 Placing performance orderings induced by the score κ on the Tile . . . 19
A.3.4 Placing performance orderings induced by the score WA on the Tile . . . 22
A.3.5 Placing performance orderings induced by some other non-probabilistic scores on the Tile . . . 23
A.4. Supplementary material about Sec. 4.3 . . . 26
A.4.1 Algorithmic contributions . . . 26
A.5. Supplementary material about Sec. 4.4 . . . 27
A.5.1 What is achievable by no-skilled classifiers? . . . 27
A.6. Averaging all canonical ranking scores: the *Volume Under Tile* . . . 28
A.6.1 Definition . . . 28
A.6.2 Closed-form expression . . . 28
A.6.3 Discussion: can we use it to rank? . . . 29

A.1. List of symbols

A.1.1 Mathematical Symbols

- $\mathbf{1}_U$: the 0-1 indicator function of subset U
- $\mathbb{R}$: the real numbers
- $\mathcal{R}$: a relation
- conv: the set of convex combinations
- $\vee$: the *inclusive disjunction* (*i.e.*, logical or)
- $\wedge$: the *conjunction* (*i.e.*, logical and)
- $\circ$: the composition of functions, *i.e.* $(g \circ f)(x) = g(f(x))$
- $\mathbf{E}$: the mathematical expectation

A.1.2 Symbols related to our mathematical framework of paper A [28]

We organize these symbols according to the 6 pillars depicted in the graphical abstract of paper A [28].

Symbols related to the 1st pillar
- Ω: the sample space (universe)
- ω: a sample (*i.e.*, an element of Ω)
- Σ: the event space (a σ-algebra on Ω, *e.g.* 2^Ω)
- E: an event (*i.e.*, an element of Σ)
- (Ω, Σ): the measurable space
- $\mathbb{P}_{(\Omega,\Sigma)}$: all performances on (Ω, Σ)
- Π: a set of performances ($\Pi \subseteq \mathbb{P}_{(\Omega,\Sigma)}$)
- P: a performance (*i.e.*, an element of $\mathbb{P}_{(\Omega,\Sigma)}$)

Symbols related to the 2nd pillar
- $\lesssim$: binary relation *worse or equivalent* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $\gtrsim$: binary relation *better or equivalent* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $\sim$: binary relation *equivalent* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $>$: binary relation *better* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $<$: binary relation *worse* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $\lesssim\!\!\!\!\!\!\diagup\;\;\diagdown\!\!\!\!\!\!\gtrsim$: binary relation *incomparable* on $\mathbb{P}_{(\Omega,\Sigma)}$

Symbols related to the 3rd pillar
- S: the random variable *Satisfaction*

Symbols related to the 4th pillar
- $\mathbb{E}$: the set of entities to rank
- ϵ: an entity, *i.e.* an element of $\mathbb{E}$
- eval: the performance *evaluation* function
- Φ: some performances that are for sure achievable

Symbols related to the 5th pillar
- $\mathbb{X}_{(\Omega,\Sigma)}$: all scores on (Ω, Σ)
- X: a score
- $\mathrm{dom}(X)$: the domain of the score X
- X_E^U: the *unconditional probabilistic score* parameterized by the event E
- $X_{E_1|E_2}^C$: the *conditional probabilistic score* parameterized by the events E_1 and E_2

Symbols related to the 6th pillar
- I: the random variable *Importance*

A.1.3 Symbols used for operations on performances

- filter_I: the *filtering* operation, parameterized by a random variable I, as defined in paper A [28]
- no–skill: the operation that transforms a performance P into P' such that $P'(Y, \hat{Y}) = P(Y)P(\hat{Y})$
- $\mathrm{shift}_{\pi \to \pi'}$: the *prior/target shift* operation [34]
- $\mathrm{change}_{\hat{Y}}$: the operation that changes the predicted class $\hat{Y}$
- change_{Y}: the operation that changes the ground-truth class Y
- $\mathrm{swap}_{Y \leftrightarrow \hat{Y}}$: the operation that swaps the predicted ($\hat{Y}$) and ground-truth (Y) classes
- $\mathrm{swap}_{c_- \leftrightarrow c_+}$: the operation that swaps the classes c_- and c_+.

A.1.4 Symbols used in the performance ordering and performance-based ranking theory

- $\mathrm{rank}_{\mathbb{E}}$: the *ranking* function, w.r.t. the set of entities $\mathbb{E}$
- $\lesssim_X$: the ordering induced by the score X
- $\gtrsim_X$: the dual (inverted) ordering induced by the score X
- R_I: the *ranking score* parameterized by the importance I
- $R_{I_{a,b}}$: the *canonical ranking score* parameterized by the parameters a and b
- a: the parameter specifying the relative importance given to the incorrect outcomes ($S = 0$), it corresponds to the horizontal axis of the Tile
- b: the parameter specifying the relative importance given to the correct outcomes ($S = 1$), it corresponds to the vertical axis of the Tile
- τ: the rank correlation coefficient of Kendall

A.1.5 Symbols used for the particular case of two-class crisp classifications

Particularization of the mathematical framework

- tn: the sample *true negative*
- fp: the sample *false positive*, *a.k.a.* type I error
- fn: the sample *false negative*, *a.k.a.* type II error
- tp: the sample *true positive*

Extensions to the mathematical framework

- ROC: the *Receiver Operating Characteristic* space, *i.e.* $FPR \times TPR$
- PR: the *Precision-Recall* space, *i.e.* $TPR \times PPV$
- Y: the random variable for the ground truth
- $\hat{Y}$: the random variable for the prediction
- $\mathbb{C}$: the set of classes
- c: a class (*i.e.*, an element of $\mathbb{C}$)
- c_-: the negative class
- c_+: the positive class
- γ_π: the locus, on the Tile, of all the canonical ranking scores that put the no-skill performances on the same footing, for fixed class priors
- γ_τ: the locus, on the Tile, of all the canonical ranking scores that put the no-skill performances on the same footing, for fixed rates of predictions

Some unconditional probabilistic scores

- PTN: the probability of the elementary event *true negative*, *a.k.a. rejection rate*
- PFP: the probability of the elementary event *false positive*
- PFN: the probability of the elementary event *false negative*
- PTP: the probability of the elementary event *true positive*, *a.k.a. detection rate*
- π_-: the *prior of the negative class*
- π_+: the *prior of the positive class*, *a.k.a. prevalence*
- τ_-: the *rate of negative predictions*
- τ_+: the *rate of positive predictions*

- A: the *accuracy*, *a.k.a. matching coefficient*

Some conditional probabilistic scores
- TNR: the *True Negative Rate*, *a.k.a. specificity*, *selectivity*, *inverse recall*
- FPR: the *False Positive Rate*
- TPR: the *True Positive Rate*, *a.k.a. sensitivity*, *recall*
- FNR: the *False Negative Rate*
- NPV: the *Negative Predictive Value*, *a.k.a. inverse precision*
- FOR: the *False Omission Rate*
- NPV: the *Positive Predictive Value*, *a.k.a. precision*
- FDR: the *False Discovery Rate*
- J_-: Jaccard index for the negative class
- J_+: Jaccard index for the positive class, *a.k.a. Tanimoto coefficient*, *similarity*, *intersection over union*, *critical success index* [23], *G-measure* [14]

Some other scores
- S: Bennett, Alpert and Goldstein's S
- F_β: the F-scores
- F_1: the F-one score, *a.k.a. Dice-Sørensen coefficient*
- $SNPV$: the score *Standardized Negative Predictive Value* [22]
- $SPPV$: the score *Standardized Positive Predictive Value* [22]
- NLR: the score *Negative Likelihood Ratio*
- PLR: the score *Positive Likelihood Ratio*
- κ: Cohen's kappa statistic, *a.k.a. Heidke Skill Score* [40]
- π: Scott's pi statistic
- κ: Fleiss's kappa statistic
- BI: the *Bias Index*, as defined in [9]
- WA: the *Weighted Accuracy*
- BA: the *Balanced Accuracy*
- J_Y: Youden's index [42], *a.k.a.* Youden's J statistic, *informedness*, *Peirce Skill Score* [40]
- $|C|$: the determinant of the (normalized) confusion matrix or contingency matrix
- ACP: the *Average Conditional Probability*, *i.e.* the arithmetic mean of the Tile's four corners [8]
- P_4: the harmonic mean of the Tile's four corners [35]
- VUT: the score *Volume Under Tile*, *i.e.* the arithmetic mean of all canonical scores (see Appendix A.6)

Table 1. Summary of the effects on the Tile of 5 operations on performances.

operation on performances		the ordering that was at $(a_{\text{origin}}, b_{\text{origin}})$ is moved at $(a_{\text{adapted}}, b_{\text{adapted}})$		note
$\text{change}_{\hat{Y}}$	(see Lemma 1)	$a_{\text{adapted}} = b_{\text{origin}}$	$b_{\text{adapted}} = a_{\text{origin}}$	the preorder is inverted (dual)
change_{Y}	(see Lemma 2)	$a_{\text{adapted}} = 1 - b_{\text{origin}}$	$b_{\text{adapted}} = 1 - a_{\text{origin}}$	the preorder is inverted (dual)
$\text{swap}_{Y \leftrightarrow \hat{Y}}$	(see Lemma 3)	$a_{\text{adapted}} = a_{\text{origin}}$	$b_{\text{adapted}} = 1 - b_{\text{origin}}$	the preorder is unchanged
$\text{swap}_{c_- \leftrightarrow c_+}$	(see Lemma 4)	$a_{\text{adapted}} = 1 - a_{\text{origin}}$	$b_{\text{adapted}} = 1 - b_{\text{origin}}$	the preorder is unchanged
$\text{shift}_{\pi \to \pi'}$	(see Lemma 5)	$a_{\text{adapted}} = f^{-1}(a_{\text{origin}})$	$b_{\text{adapted}} = f^{-1}(b_{\text{origin}})$	the preorder is unchanged

A.2. Supplementary material about Sec. 4.1

A.2.1 Operations on performances

We present here the proofs for the geometric properties of the Tile that are related to some operations that can be applied to performances. A summary is provided in Tab. 1.

Lemma 1. *Let* $\text{change}_{\hat{Y}} : \mathbb{P}_{(\Omega,\Sigma)} \to \mathbb{P}_{(\Omega,\Sigma)}$ *be the operation that changes the predicted class* $\hat{Y}$*. We have* $R_{I_{a,b}} \circ \text{change}_{\hat{Y}} = 1 - R_{I_{b,a}}$.

Proof. Let P be a two-class classification performance and $P' = \text{change}_{\hat{Y}}(P)$. If $P' \in \text{dom}(R_{I_{a,b}})$, then $P \in \text{dom}(R_{I_{b,a}})$ and

$$\begin{aligned}
R_{I_{a,b}}(P') &= \frac{(1-a)P'(\{tn\}) + aP'(\{tp\})}{(1-a)P'(\{tn\}) + (1-b)P'(\{fp\}) + bP'(\{fn\}) + aP'(\{tp\})} \\
&= \frac{(1-a)P(\{fp\}) + aP(\{fn\})}{(1-a)P(\{fp\}) + (1-b)P(\{tn\}) + bP(\{tp\}) + aP(\{fn\})} \\
&= 1 - \frac{(1-b)P(\{tn\}) + bP(\{tp\})}{(1-b)P(\{tn\}) + (1-a)P(\{fp\}) + aP(\{fn\}) + bP(\{tp\})} \\
&= 1 - R_{I_{b,a}}(P)
\end{aligned}$$

□

Lemma 2. *Let* $\text{change}_{Y} : \mathbb{P}_{(\Omega,\Sigma)} \to \mathbb{P}_{(\Omega,\Sigma)}$ *be the operation that changes the ground-truth class* Y*. We have* $R_{I_{a,b}} \circ \text{change}_{Y} = 1 - R_{I_{1-b,1-a}}$.

Proof. Let P be a two-class classification performance and $P' = \text{change}_{Y}(P)$. If $P' \in \text{dom}(R_{I_{a,b}})$, then $P \in \text{dom}(R_{I_{1-b,1-a}})$ and

$$\begin{aligned}
R_{I_{a,b}}(P') &= \frac{(1-a)P'(\{tn\}) + aP'(\{tp\})}{(1-a)P'(\{tn\}) + (1-b)P'(\{fp\}) + bP'(\{fn\}) + aP'(\{tp\})} \\
&= \frac{(1-a)P(\{fn\}) + aP(\{fp\})}{(1-a)P(\{fn\}) + (1-b)P(\{tp\}) + bP(\{tn\}) + aP(\{fp\})} \\
&= 1 - \frac{bP(\{tn\}) + (1-b)P(\{tp\})}{bP(\{tn\}) + aP(\{fp\}) + (1-a)P(\{fn\}) + (1-b)P(\{tp\})} \\
&= 1 - R_{I_{1-b,1-a}}(P)
\end{aligned}$$

□

Lemma 3. *Let* $\text{swap}_{Y \leftrightarrow \hat{Y}} : \mathbb{P}_{(\Omega,\Sigma)} \to \mathbb{P}_{(\Omega,\Sigma)}$ *be the operation that swaps the predicted* ($\hat{Y}$) *and ground-truth* (Y) *classes. We have* $R_{I_{a,b}} \circ \text{swap}_{Y \leftrightarrow \hat{Y}} = R_{I_{a,1-b}}$.

Proof. Let P be a two-class classification performance and $P' = \mathrm{swap}_{Y \leftrightarrow \hat{Y}}(P)$. If $P' \in \mathrm{dom}(R_{I_{a,b}})$, then $P \in \mathrm{dom}(R_{I_{a,1-b}})$ and

$$\begin{aligned}
R_{I_{a,b}}(P') &= \frac{(1-a)P'(\{tn\}) + aP'(\{tp\})}{(1-a)P'(\{tn\}) + (1-b)P'(\{fp\}) + bP'(\{fn\}) + aP'(\{tp\})} \\
&= \frac{(1-a)P(\{tn\}) + aP(\{tp\})}{(1-a)P(\{tn\}) + (1-b)P(\{fn\}) + bP(\{fp\}) + aP(\{tp\})} \\
&= \frac{(1-a)P(\{tn\}) + aP(\{tp\})}{(1-a)P(\{tn\}) + bP(\{fp\}) + (1-b)P(\{fn\}) + aP(\{tp\})} \\
&= R_{I_{a,1-b}}(P)
\end{aligned}$$

□

Lemma 4. *Let* $\mathrm{swap}_{c_- \leftrightarrow c_+} : \mathbb{P}_{(\Omega,\Sigma)} \to \mathbb{P}_{(\Omega,\Sigma)}$ *be the operation that swaps the classes* c_- *and* c_+*. We have* $R_{I_{a,b}} \circ \mathrm{swap}_{c_- \leftrightarrow c_+} = R_{I_{1-a,1-b}}$.

Proof. Let P be a two-class classification performance and $P' = \mathrm{swap}_{c_- \leftrightarrow c_+}(P)$. If $P' \in \mathrm{dom}(R_{I_{a,b}})$, then $P \in \mathrm{dom}(R_{I_{1-a,1-b}})$ and

$$\begin{aligned}
R_{I_{a,b}}(P') &= \frac{(1-a)P'(\{tn\}) + aP'(\{tp\})}{(1-a)P'(\{tn\}) + (1-b)P'(\{fp\}) + bP'(\{fn\}) + aP'(\{tp\})} \\
&= \frac{(1-a)P(\{tp\}) + aP(\{tn\})}{(1-a)P(\{tp\}) + (1-b)P(\{fn\}) + bP(\{fp\}) + aP(\{tn\})} \\
&= \frac{aP(\{tn\}) + (1-a)P(\{tp\})}{aP(\{tn\}) + bP(\{fp\}) + (1-b)P(\{fn\}) + (1-a)P(\{tp\})} \\
&= R_{I_{1-a,1-b}}(P)
\end{aligned}$$

□

Lemma 5. *Let* $\mathrm{shift}_{\pi \to \pi'}$ *be the operation that applies a prior/target shift on the distribution* $P(Y)$*, transforming the priors* (π_-, π_+) *into* (π'_-, π'_+)*. The ordering induced by* $R_{I_{a,b}} \circ \mathrm{shift}_{\pi \to \pi'}$ *is the same as the ordering induced by* $R_{I_{f(a),f(b)}}$*, where* f *is the function*

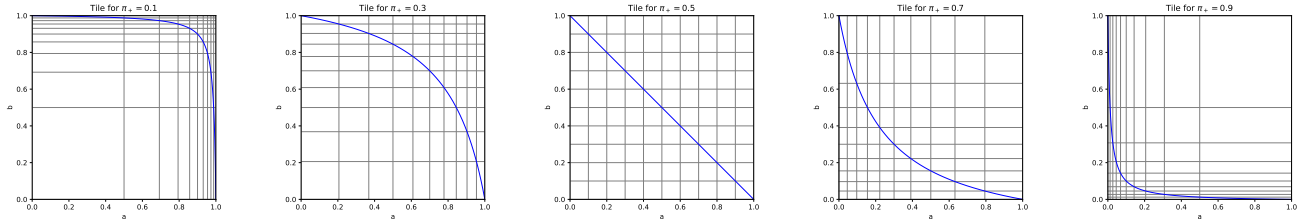
$$f : x \mapsto f(x) = \frac{x\frac{\pi'_+}{\pi_+}}{(1-x)\frac{\pi'_-}{\pi_-} + x\frac{\pi'_+}{\pi_+}}\,.$$

Proof. Let P be a two-class classification performance and $P' = \mathrm{shift}_{\pi \to \pi'}(P)$. We have:

$$\begin{aligned}
R_{I_{a,b}}(P') &= \frac{(1-a)P'(\{tn\}) + aP'(\{tp\})}{(1-a)P'(\{tn\}) + (1-b)P'(\{fp\}) + bP'(\{fn\}) + aP'(\{tp\})} \\
&= \frac{(1-a)P(\{tn\})\frac{\pi'_-}{\pi_-} + aP(\{tp\})\frac{\pi'_+}{\pi_+}}{(1-a)P(\{tn\})\frac{\pi'_-}{\pi_-} + (1-b)P(\{fp\})\frac{\pi'_-}{\pi_-} + bP(\{fn\})\frac{\pi'_+}{\pi_+} + aP(\{tp\})\frac{\pi'_+}{\pi_+}} \\
&= \frac{I'(tn)P(\{tn\}) + I'(tp)P(\{tp\})}{I'(tn)P(\{tn\}) + I'(fp)P(\{fp\}) + I'(fn)P(\{fn\}) + I'(tp)P(\{tp\})} \\
&= R_{I'}(P)
\end{aligned}$$

with

$$\begin{cases}
I'(tn) = g(a)\,(1 - f(a)) \\
I'(fp) = g(b)\,(1 - f(b)) \\
I'(fn) = g(b)f(b) \\
I'(tp) = g(a)f(a)
\end{cases}$$



(a) Shift from $(\pi_-, \pi_+) = (0.5, 0.5)$ to $(0.9, 0.1)$.

(b) Shift from $(\pi_-, \pi_+) = (0.5, 0.5)$ to $(0.7, 0.3)$.

(c) Reference.

(d) Shift from $(\pi_-, \pi_+) = (0.5, 0.5)$ to $(0.3, 0.7)$.

(e) Shift from $(\pi_-, \pi_+) = (0.5, 0.5)$ to $(0.1, 0.9)$.

Figure 8. Visualization of how the Tile deforms with a shift for the distribution $P(Y)$. The central figure shows an initial Tile, in which we have arbitrarily drawn a regular grid. The other figures show the result of Lemma 5, representing the way the Tile deforms. We've also drawn the descending diagonal on our reference Tile. Its deformation gives birth to the family of curves denoted by γ_π in our paper.

and

$$g : x \mapsto g(x) = (1 - x)\frac{\pi'_-}{\pi_-} + x\frac{\pi'_+}{\pi_+}\,.$$

The random variable I' can be rewritten as $I' = (\mathbf{1}_{S=0}g(b) + \mathbf{1}_{S=1}g(a))I_{f(a),f(b)}$. Using Property 2, we conclude that

$$\begin{aligned}
&\frac{\partial R_{I'}}{\partial R_{I_{f(a),f(b)}}} > 0 \\
\Rightarrow &\frac{\partial \left(R_{I_{a,b}} \circ \text{shift}_{\pi \to \pi'}\right)}{\partial R_{I_{f(a),f(b)}}} > 0 \\
\Rightarrow &\lesssim_{R_{I_{a,b}} \circ \text{shift}_{\pi \to \pi'}} = \lesssim_{R_{I_{f(a),f(b)}}} \,.
\end{aligned}$$

□

A.2.2 When the priors are fixed: superimposing a grid on the Tile

We found it very useful, when priors are fixed, to represent the displacement that would have occurred if we had started with uniform priors and applied the target/prior shift. This can be done thanks to Lemma 5. In practice, we do this by superimposing a grid on the Tile, as done in Fig. 5 and Fig. 7. We provide more information on the subject in Fig. 8. Vertical lines are drawn at

$$a = f^{-1}(x) = \frac{x\frac{0.5}{\pi_+}}{(1-x)\frac{0.5}{\pi_-} + x\frac{0.5}{\pi_+}} \text{ for } x = 0.0, 0.1, 0.2, \ldots, 0.9, 1.0\,, \tag{4}$$

and horizontal lines are drawn at

$$b = f^{-1}(y) = \frac{y\frac{0.5}{\pi_+}}{(1-y)\frac{0.5}{\pi_-} + y\frac{0.5}{\pi_+}} \text{ for } y = 0.0, 0.1, 0.2, \ldots, 0.9, 1.0\,. \tag{5}$$

Table 2. The nine ranking scores for two-class classifications that belong to the family of probabilistic scores.

Ranking score R_I	Probabilistic writing	$I(tn)$	$I(fp)$	$I(fn)$	$I(tp)$	Canonical?	location of $\lesssim_{R_I}$ on the Tile
Negative Predictive Value NPV	$P(S=1 \mid \hat{Y}=c_-)$	1	0	1	0	yes	$(a,b)=(0,1)$
$X^C_{\{tn,tp\}\mid\{tn,fn,tp\}}$	$P(S=1 \mid Y=c_+ \vee \hat{Y}=c_-)$	1	0	1	1	no	$(a,b)=(\frac{1}{2},1)$
True Positive Rate TPR	$P(S=1 \mid Y=c_+)$	0	0	1	1	yes	$(a,b)=(1,1)$
Jaccard's index for the negative class J_-	$P(S=1 \mid Y=c_- \vee \hat{Y}=c_-)$	1	1	1	0	no	$(a,b)=(0,\frac{1}{2})$
Accuracy A	$P(S=1)$	$1/2$	$1/2$	$1/2$	$1/2$	yes	$(a,b)=(\frac{1}{2},\frac{1}{2})$
Jaccard's index for the positive class J_+	$P(S=1 \mid Y=c_+ \vee \hat{Y}=c_+)$	0	1	1	1	no	$(a,b)=(1,\frac{1}{2})$
True Negative Rate TNR	$P(S=1 \mid Y=c_-)$	1	1	0	0	yes	$(a,b)=(0,0)$
$X^C_{\{tn,tp\}\mid\{tn,fp,tp\}}$	$P(S=1 \mid Y=c_- \vee \hat{Y}=c_+)$	1	1	0	1	no	$(a,b)=(\frac{1}{2},0)$
Positive Predictive Value PPV	$P(S=1 \mid \hat{Y}=c_+)$	0	1	0	1	yes	$(a,b)=(1,0)$

A.3. Supplementary material about Sec. 4.2

A.3.1 Placing performance orderings induced by probabilistic scores on the Tile

The following lemma explains that some probabilistic scores are ranking scores, and specifies the importance values for them. Based on this, placing the performance orderings induced by probabilistic scores is straightforward, as the performance ordering induced by the ranking score R_I is located at $(a,b)=(\frac{I(tp)}{I(tn)+I(tp)}, \frac{I(fn)}{I(fp)+I(fn)})$.

Lemma 6. *Let S be a binary-valued satisfaction. A probabilistic score $X^C_{E_1|E_2}$, with $\emptyset \subsetneq E_1 \subsetneq E_2 \subseteq \Omega$, is a ranking score R_I if $I = k\mathbf{1}_{E_2}$, $k>0$, and $E_1 = E_2 \cap E_S$, with $E_S = \{\omega \in \Omega : S(\omega)=1\}$.*

Proof. Let us first check the equality of the domains. We have

$$\mathrm{dom}(R_I) = \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : \mathbf{E}_P[I] \neq 0\right\}$$

and

$$\mathrm{dom}(X^C_{E_1|E_2}) = \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : P(E_2) \neq 0\right\}.$$

Using the "fundamental bridge", we have

$$P(E_2) \neq 0 \Leftrightarrow \mathbf{E}_P[\mathbf{1}_{E_2}] \neq 0 \Leftrightarrow \mathbf{E}_P[I] \neq 0\,,$$

and thus

$$\mathrm{dom}(R_I) = \mathrm{dom}(X^C_{E_1|E_2})\,.$$

We now check the equality of the values taken by both scores. We have, for all $P \in \mathrm{dom}(X^C_{E_1|E_2})$,

$$X^C_{E_1|E_2}(P) = P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} = \frac{P(E_1)}{P(E_2)}\,.$$

Using again the "fundamental bridge",

$$X^C_{E_1|E_2}(P) = \frac{\mathbf{E}_P[\mathbf{1}_{E_1}]}{\mathbf{E}_P[\mathbf{1}_{E_2}]} = \frac{\mathbf{E}_P[\mathbf{1}_{E_2 \cap E_S}]}{\mathbf{E}_P[\mathbf{1}_{E_2}]} = \frac{\mathbf{E}_P[\mathbf{1}_{E_2} S]}{\mathbf{E}_P[\mathbf{1}_{E_2}]} = \frac{\mathbf{E}_P[\frac{I}{k} S]}{\mathbf{E}_P[\frac{I}{k}]} = \frac{\mathbf{E}_P[IS]}{\mathbf{E}_P[I]} = R_I(P)\,.$$

□

Thanks to Lemma 6, we can place on the Tile the probabilistic scores that are ranking scores. In the particular case of two-class classification, there exist 9 pairs (E_1, E_2) such that $\emptyset \subsetneq E_1 \subsetneq E_2 \subseteq \Omega$ and $E_1 = E_2 \cap E_S$. There are thus 9 ranking scores that are also probabilistic scores. Among them, 5 are canonical and can be placed directly on the Tile. For the 4 others, the induced performance ordering can be placed on the Tile. A summary is provided in Table 2.

A.3.2 Placing performance orderings induced by the scores F_β on the Tile

Lemma 7 (F-scores)**.** *In two-class classification, all F-scores are canonical ranking scores: $F_\beta = R_I$ with $I(tn) = 0$, $I(fp) = \frac{1}{1+\beta^2}$, $I(fn) = \frac{\beta^2}{1+\beta^2}$, and $I(tp) = 1$.*

Proof. For the sake of concision, let us pose $TN = P(\{tn\})$, $FP = P(\{fp\})$, $FN = P(\{fn\})$, and $TP = P(\{tp\})$.

We first check that F_β and R_I have the same domain.

$$\begin{aligned}
\mathrm{dom}(F_\beta) &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : (1+\beta^2)TP + \beta^2 FN + FP \neq 0\right\} \\
&= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : (1+\beta^2)\mathbf{E}_P[I] \neq 0\right\} \\
&= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : \mathbf{E}_P[I] \neq 0\right\} \\
&= \mathrm{dom}(R_I)
\end{aligned}$$

Then, we check the equality of the values taken by both scores. By definition of F_β, we have, for all $P \in \mathrm{dom}(F_\beta)$,

$$\begin{aligned}
F_\beta(P) &= \frac{(1+\beta^2)TP}{(1+\beta^2)TP + \beta^2 FN + FP} \\
&= \frac{0TN + 1TP}{0TN + \frac{1}{1+\beta^2}FP + \frac{\beta^2}{1+\beta^2}FN + 1TP} \\
&= \frac{I(tn)TN + I(tp)TP}{I(tn)TN + I(fp)FP + I(fn)FN + I(tp)TP} \\
&= R_I(P)\,.
\end{aligned}$$

In conclusion, $F_\beta = R_I$. As Moreover, as $I(tn) + I(tp) = I(fp) + I(fn)$, R_I is canonical. □

A.3.3 Placing performance orderings induced by the score κ on the Tile

Lemma 8 (Cohen's κ statistic)**.** *Cohen's κ statistic increases linearly with a canonical ranking score when the priors are fixed. In two-class classification, let the priors of the negative and positive classes be fixed and denoted, respectively, by π_- and π_+. In this case, $(\pi_-^2 + \pi_+^2)\kappa + 2\pi_-\pi_+ = R_I$ with $I(tn) = \frac{\pi_+^2}{\pi_-^2+\pi_+^2}$, $I(fp) = \frac{1}{2}$, $I(fn) = \frac{1}{2}$, and $I(tp) = \frac{\pi_-^2}{\pi_-^2+\pi_+^2}$.*

Proof. Let us begin by introducing the set of performances with fixed and strictly positive priors:

$$\mathbb{P}^* = \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : P(\{fn, tp\}) = \pi_+\right\} \qquad \text{with } \pi_+ \in (0,1)\ .$$

For the sake of concision, let us pose $TN = P(\{tn\})$, $FP = P(\{fp\})$, $FN = P(\{fn\})$, $TP = P(\{tp\})$, $\tau_- = P(\{fn, tn\})$, and $\tau_+ = P(\{fp, tp\})$.

Let us first observe that $\mathrm{dom}(\kappa) \cap \mathbb{P}^* = \mathrm{dom}(R_I) \cap \mathbb{P}^*$. By definition of R_I,

$$\mathrm{dom}(R_I) = \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : \mathbf{E}_P[I] \neq 0\right\}\ ,$$

and by definition of κ,

$$\mathrm{dom}(\kappa) = \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : \pi_-\tau_- + \pi_+\tau_+ \neq 1\right\}\ .$$

We discuss three cases.

1. When $\pi_+ = 0$, which implies that $FN = 0 \wedge TP = 0$, we have

$$\begin{aligned}
\mathrm{dom}(\kappa) \cap \mathbb{P}^* &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : \pi_-\tau_- + \pi_+\tau_+ \neq 1\right\} \cap \mathbb{P}^* \\
&= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : \tau_- \neq 1\right\} \cap \mathbb{P}^* \\
&= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : FN + TN \neq 1\right\} \cap \mathbb{P}^* \\
&= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : TN \neq 1\right\} \cap \mathbb{P}^* \\
&= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : \mathbf{E}_P[I] \neq 0\right\} \cap \mathbb{P}^* \\
&= \mathrm{dom}(R_I) \cap \mathbb{P}^*\,.
\end{aligned}$$

2. When $\pi_+ \in]0,1[$, we have

$$\begin{aligned}\mathrm{dom}(\kappa) \cap \mathbb{P}^* &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : \pi_-\tau_- + \pi_+\tau_+ \neq 1\right\} \cap \mathbb{P}^* \\ &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)}\right\} \cap \mathbb{P}^* \\ &= \mathrm{dom}(R_I) \cap \mathbb{P}^* \,.\end{aligned}$$

3. When $\pi_+ = 1$, which implies that $TN = 0 \wedge FP = 0$, we have

$$\begin{aligned}\mathrm{dom}(\kappa) \cap \mathbb{P}^* &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : \pi_-\tau_- + \pi_+\tau_+ \neq 1\right\} \cap \mathbb{P}^* \\ &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : \tau_+ \neq 1\right\} \cap \mathbb{P}^* \\ &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : FP + TP \neq 1\right\} \cap \mathbb{P}^* \\ &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : TP \neq 1\right\} \cap \mathbb{P}^* \\ &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : \mathbf{E}_P[I] \neq 0\right\} \cap \mathbb{P}^* \\ &= \mathrm{dom}(R_I) \cap \mathbb{P}^* \,.\end{aligned}$$

The restricted domains are thus equal in all cases:

$$\mathrm{dom}(\kappa) \cap \mathbb{P}^* = \mathrm{dom}(R_I) \cap \mathbb{P}^* \,.$$

We now check the equality of the values taken by both scores. By definition, for all $P \in \mathrm{dom}(\kappa)$,

$$\begin{aligned}\kappa(P) &= \frac{A - EA}{1 - EA} \text{ avec } \begin{cases} A = P(\{tn, tp\}) & \text{(the accuracy)} \\ EA = \pi_-\tau_- + \pi_+\tau_+ & \text{(the expected accuracy)} \end{cases} \\ &= \frac{(TN + TP) - (\pi_-TN + \pi_+FP + \pi_-FN + \pi_+TP)}{(TN + FP + FN + TP) - (\pi_-TN + \pi_+FP + \pi_-FN + \pi_+TP)} \\ &= \frac{(1-\pi_-)TN - \pi_+FP - \pi_-FN + (1-\pi_+)TP}{(1-\pi_-)TN + (1-\pi_+)FP + (1-\pi_-)FN + (1-\pi_+)TP} \\ &= \frac{\pi_+TN - \pi_+FP - \pi_-FN + \pi_-TP}{\pi_+TN + \pi_-FP + \pi_+FN + \pi_-TP}\end{aligned}$$

Thus,

$$(\pi_-^2 + \pi_+^2)\kappa(P) + 2\pi_-\pi_+ = \frac{\lambda_{tn}TN + \lambda_{fp}FP + \lambda_{fn}FN + \lambda_{tp}TP}{\pi_+TN + \pi_-FP + \pi_+FN + \pi_-TP}$$

with

$$\begin{aligned}\lambda_{tn} &= \pi_+(\pi_-^2 + \pi_+^2) + \pi_+(2\pi_-\pi_+) \\ &= \pi_+(\pi_-^2 + 2\pi_-\pi_+ + \pi_+^2) \\ &= \pi_+(\pi_- + \pi_+)^2 \\ &= \pi_+\end{aligned}$$

$$\begin{aligned}\lambda_{fp} &= -\pi_+(\pi_-^2 + \pi_+^2) + \pi_-(2\pi_-\pi_+) \\ &= \pi_+(\pi_-^2 - \pi_+^2) \\ &= \pi_+(\pi_- - \pi_+)(\pi_- + \pi_+) \\ &= \pi_+(\pi_- - \pi_+)\end{aligned}$$

$$\begin{aligned}\lambda_{fn} &= -\pi_-(\pi_-^2 + \pi_+^2) + \pi_+(2\pi_-\pi_+) \\ &= \pi_-(\pi_+^2 - \pi_-^2) \\ &= \pi_-(\pi_+ - \pi_-)(\pi_+ + \pi_-) \\ &= \pi_-(\pi_+ - \pi_-)\end{aligned}$$

$$\begin{aligned}\lambda_{tp} &= \pi_-(\pi_-^2 + \pi_+^2) + \pi_-(2\pi_-\pi_+) \\ &= \pi_-(\pi_-^2 + 2\pi_-\pi_+ + \pi_+^2) \\ &= \pi_-(\pi_- + \pi_+)^2 \\ &= \pi_-\end{aligned}$$

We continue by eliminating FP and FN from the equation. For all $P \in \mathrm{dom}(\kappa) \cap \mathbb{P}^*$,

$$\begin{aligned}(\pi_-^2 + \pi_+^2)\kappa(P) + 2\pi_-\pi_+ &= \frac{\lambda_{tn}TN + \lambda_{fp}(\pi_- - TN) + \lambda_{fn}(\pi_+ - TP) + \lambda_{tp}TP}{\pi_+TN + \pi_-(\pi_- - TN) + \pi_+(\pi_+ - TP) + \pi_-TP} \\ &= \frac{(\lambda_{tn} - \lambda_{fp})TN + (\lambda_{fp}\pi_- + \lambda_{fn}\pi_+) + (\lambda_{tp} - \lambda_{fn})TP}{(\pi_+ - \pi_-)TN + (\pi_-^2 + \pi_+^2) + (\pi_- - \pi_+)TP}\end{aligned}$$

We have

$$\begin{aligned}\lambda_{tn} - \lambda_{fp} &= \pi_+ - \pi_+(\pi_- - \pi_+) \\ &= \pi_+(\pi_- + \pi_+) - \pi_+(\pi_- - \pi_+) \\ &= \pi_+\pi_- + \pi_+^2 - \pi_+\pi_- + \pi_+^2 \\ &= 2\pi_+^2\end{aligned}$$

$$\begin{aligned}\lambda_{fp}\pi_- + \lambda_{fn}\pi_+ &= \pi_+(\pi_- - \pi_+)\pi_- + \pi_-(\pi_+ - \pi_-)\pi_+ \\ &= \pi_-\pi_+(\pi_- - \pi_+ + \pi_+ - \pi_-) \\ &= 0\end{aligned}$$

$$\begin{aligned}\lambda_{tp} - \lambda_{fn} &= \pi_- - \pi_-(\pi_+ - \pi_-) \\ &= \pi_-(\pi_+ + \pi_-) - \pi_-(\pi_+ - \pi_-) \\ &= \pi_-\pi_+ + \pi_-^2 - \pi_-\pi_+ + \pi_-^2 \\ &= 2\pi_-^2\end{aligned}$$

So, for all $P \in \mathrm{dom}(\kappa) \cap \mathbb{P}^*$,

$$(\pi_-^2 + \pi_+^2)\kappa(P) + 2\pi_-\pi_+ = \frac{2\pi_+^2TN + 2\pi_-^2TP}{(\pi_+ - \pi_-)TN + (\pi_-^2 + \pi_+^2) + (\pi_- - \pi_+)TP}$$

Let us now rework the denominator.

$$\begin{aligned}&(\pi_+ - \pi_-)TN + (\pi_-^2 + \pi_+^2) + (\pi_- - \pi_+)TP \\ =&(\pi_+ - \pi_-)(\pi_+ + \pi_-)TN + (\pi_-^2 + \pi_+^2) + (\pi_- - \pi_+)(\pi_- + \pi_+)TP \\ =&(\pi_+^2 - \pi_-^2)TN + (\pi_-^2 + \pi_+^2) + (\pi_-^2 - \pi_+^2)TP \\ =&(\pi_+^2 - \pi_-^2)TN + (\pi_-^2 + \pi_+^2)(\pi_- + \pi_+) + (\pi_-^2 - \pi_+^2)TP \\ =&(\pi_+^2 - \pi_-^2)TN + (\pi_-^2 + \pi_+^2)\pi_- + (\pi_-^2 + \pi_+^2)\pi_+ + (\pi_-^2 - \pi_+^2)TP \\ =&2\pi_+^2TN + (\pi_-^2 + \pi_+^2)(\pi_- - TN) + (\pi_-^2 + \pi_+^2)(\pi_+ - TP) + 2\pi_-^2TP \\ =&2\pi_+^2TN + (\pi_-^2 + \pi_+^2)FP + (\pi_-^2 + \pi_+^2)FN + 2\pi_-^2TP\end{aligned}$$

And thus, for all $P \in \mathrm{dom}(\kappa) \cap \mathbb{P}^*$,

$$\begin{aligned}(\pi_-^2 + \pi_+^2)\kappa(P) + 2\pi_-\pi_+ =& \frac{2\pi_+^2 TN + 2\pi_-^2 TP}{2\pi_+^2 TN + (\pi_-^2 + \pi_+^2)FP + (\pi_-^2 + \pi_+^2)FN + 2\pi_-^2 TP} \\ =& \frac{\frac{\pi_+^2}{\pi_-^2+\pi_+^2}TN + \frac{\pi_-^2}{\pi_-^2+\pi_+^2}TP}{\frac{\pi_+^2}{\pi_-^2+\pi_+^2}TN + \frac{1}{2}FP + \frac{1}{2}FN + \frac{\pi_-^2}{\pi_-^2+\pi_+^2}TP} \\ =& \frac{I(tn)TN + I(tp)TP}{I(tn)TN + I(fp)FP + I(fn)FN + I(tp)TP} \\ =& R_I(P)\end{aligned}$$

In conclusion, $(\pi_-^2+\pi_+^2)\kappa+2\pi_-\pi_+ = R_I$ on $\mathrm{dom}(\kappa)\cap\mathbb{P}^* = \mathrm{dom}(R_I)\cap\mathbb{P}^*$. Moreover, as $I(tn)+I(tp) = I(fp)+I(fn)$, R_I is canonical. □

A.3.4 Placing performance orderings induced by the score WA on the Tile

Lemma 9 (The weighted accuracies)**.** *In two-class classification, let the priors of the negative and positive classes be fixed, strictly positive, and denoted, respectively, by π_- and π_+. In this case, all weighted accuracies $WA = (1-\alpha)TNR+\alpha TPR$, $\alpha \in [0,1]$, are canonical ranking scores: $WA = R_I$ with*

$$\begin{cases} I(tn) = I(fp) = \frac{\frac{1-\alpha}{\pi_-}}{\frac{1-\alpha}{\pi_-}+\frac{\alpha}{\pi_+}} \\ I(fn) = I(tp) = \frac{\frac{\alpha}{\pi_+}}{\frac{1-\alpha}{\pi_-}+\frac{\alpha}{\pi_+}} \end{cases} .$$

Proof. Let us begin by introducing the set of performances with fixed and strictly positive priors:

$$\mathbb{P}^* = \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : P(\{fn, tp\}) = \pi_+\right\} \qquad \text{with } \pi_+ \in (0,1) .$$

For the sake of concision, we pose $TN = P(\{tn\})$, $FP = P(\{fp\})$, $FN = P(\{fn\})$, and $TP = P(\{tp\})$.

We first check the equality of the restricted domains. The domain of WA is $\mathrm{dom}(TNR) \cap \mathrm{dom}(TPR)$, that is

$$\mathrm{dom}(WA) = \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : P(\{fn, tp\}) \notin \{0,1\}\right\} ,$$

and thus the restricted domain of the weighted accuracy WA is

$$\mathrm{dom}(WA) \cap \mathbb{P}^* = \mathbb{P}^* .$$

For the domain of R_I, we have to take into account the fact that I should be well defined ($\pi_- \neq 0$ and $\pi_+ \neq 0$) and that the mathematical expectation of the importance should be non-zero, which is always the case:

$$\begin{aligned}\mathbf{E}_P[I] \neq 0 &\Leftrightarrow \frac{\frac{1-\alpha}{\pi_-}}{\frac{1-\alpha}{\pi_-}+\frac{\alpha}{\pi_+}}(TN+FP) + \frac{\frac{\alpha}{\pi_+}}{\frac{1-\alpha}{\pi_-}+\frac{\alpha}{\pi_+}}(FN+TP) \neq 0 \\ &\Leftrightarrow \frac{1-\alpha}{\pi_-}(TN+FP) + \frac{\alpha}{\pi_+}(FN+TP) \neq 0 \\ &\Leftrightarrow \frac{1-\alpha}{\pi_-}\pi_- + \frac{\alpha}{\pi_+}\pi_+ \neq 0 \\ &\Leftrightarrow 1 \neq 0\end{aligned}$$

Thus, the restricted domain of the ranking score R_I is

$$\mathrm{dom}(R_I) \cap \mathbb{P}^* = \mathbb{P}^* .$$

$$\mathrm{dom}(R_I) = \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : \mathbf{E}_P[I] \neq 0\right\}$$

The restricted domains are thus equal:

$$\mathrm{dom}(WA) \cap \mathbb{P}^* = \mathrm{dom}(R_I) \cap \mathbb{P}^* \,.$$

We now check the equality of the values taken by both scores. For all $P \in \mathrm{dom}(WA) \cap \mathbb{P}^*$, we have:

$$\begin{aligned}
WA(P) &= (1-\alpha)TNR(P) + \alpha TPR(P) \\
&= (1-\alpha)\frac{TN}{\pi_-} + \alpha\frac{TP}{\pi_+} \\
&= \frac{1-\alpha}{\pi_-}TN + \frac{\alpha}{\pi_+}TP \\
&= \frac{\frac{1-\alpha}{\pi_-}TN + \frac{\alpha}{\pi_+}TP}{\frac{1-\alpha}{\pi_-}\pi_- + \frac{\alpha}{\pi_+}\pi_+} \\
&= \frac{\frac{1-\alpha}{\pi_-}TN + \frac{\alpha}{\pi_+}TP}{\frac{1-\alpha}{\pi_-}(TN+FP) + \frac{\alpha}{\pi_+}(FN+TP)} \\
&= \frac{\frac{1-\alpha}{\pi_-}TN + \frac{\alpha}{\pi_+}TP}{\frac{1-\alpha}{\pi_-}TN + \frac{1-\alpha}{\pi_-}FP + \frac{\alpha}{\pi_+}FN + \frac{\alpha}{\pi_+}TP} \\
&= \frac{I(tn)TN + I(tp)TP}{I(tn)TN + I(fp)FP + I(fn)FN + I(tp)TP} \\
&= R_I(P)
\end{aligned}$$

In conclusion, $WA = R_I$ on $\mathrm{dom}(WA) \cap \mathbb{P}^* = \mathrm{dom}(R_I) \cap \mathbb{P}^*$. Moreover, as $I(tn) + I(tp) = I(fp) + I(fn)$, R_I is canonical. □

The previous lemma can be particularized for the balanced accuracy and for the accuracy. Taking $\alpha = \frac{1}{2}$, we obtain $BA = WA = R_I$ with $I(tn) = I(fp) = \pi_+$ and $I(fn) = I(tp) = \pi_-$. Taking $\alpha = \pi_+$, we obtain $A = WA = R_I$ with $I(tn) = I(fp) = 1/2$ and $I(fn) = I(tp) = 1/2$.

A.3.5 Placing performance orderings induced by some other non-probabilistic scores on the Tile

Lemma 10 (Other particular scores)**.** *Let us consider performance orderings induced by scores by the mechanism described in the 1st theorem of paper A [28]. In two-class classification, Youden's index J_Y leads to the same performance ordering as the balanced accuracy BA, and Jaccard's coefficient for the positive class J_+ leads to the same performance ordering as F_1. Moreover, when the class priors are fixed and non-zero, the standardized negative predictive value $SNPV$ leads to the same performance ordering as the negative predictive value NPV when $P(\{tn, fn\}) \neq 0$, the negative likelihood ratio leads to the dual performance ordering of NPV, the standardized positive predictive value $SPPV$ leads to the same performance ordering as the positive predictive value PPV when $P(\{tp, fp\}) \neq 0$, and the positive likelihood ratio PLR leads also to the same performance ordering as PPV.*

Proof. For the sake of concision, let us pose $TN = P(\{tn\})$, $FP = P(\{fp\})$, $FN = P(\{fn\})$, and $TP = P(\{tp\})$.

- Youden's index is defined as $J_Y = TNR + TPR - 1$, and the balanced accuracy as $BA = 1/2 TNR + 1/2 TPR$. They have the same domain: $\mathrm{dom}(J_Y) = \mathrm{dom}(BA) = \mathrm{dom}(TNR) \cap \mathrm{dom}(TPR)$. Trivially,

$$J_Y = 2BA - 1 \,.$$

Thus,

$$\frac{\partial J_Y}{\partial BA} = 2 > 0 \,.$$

As there is a strictly increasing relationship between J_Y and BA, these two scores lead to same performance ordering.

- Jaccard's coefficient for the positive class is defined as $J_+ = X_{\{tp\}|\{fp,fn,tp\}}$ and the F-one score as $F_1 : P \mapsto \frac{2TP}{FP+FN+2TP}$. Their respective domains

$$\begin{aligned}\mathrm{dom}(J_+) &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : P(\{fp, fn, tp\}) \neq 0\right\} \\ &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : FP \neq 0 \vee FN \neq 0 \vee TP \neq 0\right\}\end{aligned}$$

and

$$\begin{aligned}\mathrm{dom}(F_1) &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : FP + FN + 2TP \neq 0\right\} \\ &= \left\{P \in \mathbb{P}_{(\Omega,\Sigma)} : FP \neq 0 \vee FN \neq 0 \vee TP \neq 0\right\}\end{aligned}$$

are equal. Trivially,

$$F_1 = \frac{2J_+}{1+J_+}\,.$$

Thus,

$$\frac{\partial F_1}{\partial J_+} = \frac{2}{(1+J_+)^2} > 0\,.$$

As there is a strictly increasing relationship between F_1 and J_+, these two scores lead to same performance ordering.

- The standardized negative predictive value is defined as $SNPV = \frac{TNR}{TNR+(1-TPR)}$ and the negative predictive value as $NPV = X_{\{tn\}|\{tn,fn\}}$. Let us consider the performances such that $TN + FN \neq 0$, $TN + FP = \pi_- \neq 0$, and $FN + TP = \pi_+ \neq 0$. We have:

$$\begin{aligned}SNPV = \frac{TNR}{TNR+(1-TPR)} &= \frac{\frac{TN}{\pi_-}}{\frac{TN}{\pi_-} + \frac{FN}{\pi_+}} \\ &= \frac{\frac{TN}{TN+FN}\frac{1}{\pi_-}}{\frac{TN}{TN+FN}\frac{1}{\pi_-} + \frac{FN}{TN+FN}\frac{1}{\pi_+}} = \frac{NPV\frac{1}{\pi_-}}{NPV\frac{1}{\pi_-} + (1-NPV)\frac{1}{\pi_+}}\,.\end{aligned}$$

Thus,

$$\frac{\partial SNPV}{\partial NPV} = \frac{\frac{1}{\pi_-}\frac{1}{\pi_+}}{\left(NPV\frac{1}{\pi_-} + (1-NPV)\frac{1}{\pi_+}\right)^2} > 0\,.$$

As there is a strictly increasing relationship between $SNPV$ and NPV, these two scores lead to same performance ordering.

- The negative likelihood ratio is defined as $NLR = \frac{1-TPR}{TNR}$ and the negative predictive value as $NPV = X_{\{tn\}|\{tn,fn\}}$. Let us consider the performances such that $TN \neq 0$, $TN + FP = \pi_-$, and $FN + TP = \pi_+ \neq 0$. We have:

$$NLR = \frac{1-TPR}{TNR} = \frac{\frac{FN}{\pi_+}}{\frac{TN}{\pi_-}} = \frac{FN}{TN}\frac{\pi_-}{\pi_+} = \frac{1-NPV}{NPV}\frac{\pi_-}{\pi_+}\,.$$

Thus,

$$\frac{\partial NLR}{\partial NPV} = -\frac{1}{NPV^2}\frac{\pi_-}{\pi_+} < 0\,.$$

As there is a strictly decreasing relationship between NLR and NPV, these two scores lead to dual performance orderings.

- The standardized positive predictive value is defined as $SPPV = \frac{TPR}{(1-TNR)+TPR}$ and the positive predictive value as $PPV = X_{\{tp\}|\{tp,fp\}}$. Let us consider the performances such that $TP + FP \neq 0$, $TN + FP = \pi_- \neq 0$, and $FN + TP = \pi_+ \neq 0$. We have:

$$\begin{aligned}SPPV = \frac{TPR}{(1-TNR)+TPR} &= \frac{\frac{TP}{\pi_+}}{\frac{FP}{\pi_-} + \frac{TP}{\pi_+}} \\ &= \frac{\frac{TP}{TP+FP}\frac{1}{\pi_+}}{\frac{FP}{TP+FP}\frac{1}{\pi_-} + \frac{TP}{TP+FP}\frac{1}{\pi_+}} = \frac{PPV\frac{1}{\pi_+}}{(1-PPV)\frac{1}{\pi_-} + PPV\frac{1}{\pi_+}}\end{aligned}$$

Thus,

$$\frac{\partial SPPV}{\partial PPV} = \frac{\frac{1}{\pi_-}\frac{1}{\pi_+}}{\left((1-PPV)\frac{1}{\pi_-} + PPV\frac{1}{\pi_+}\right)^2} > 0\,.$$

As there is a strictly increasing relationship between $SPPV$ and PPV, these two scores lead to same performance ordering.

- The positive likelihood ratio is defined as $PLR = \frac{TPR}{1-TNR}$ and the positive predictive value as $PPV = X_{\{tp\}|\{tp,fp\}}$. Let us consider the performances such that $FP \neq 0$, $TN + FP = \pi_-$, and $FN + TP = \pi_+ \neq 0$. We have:

$$PLR = \frac{TPR}{1-TNR} = \frac{\frac{TP}{\pi_+}}{\frac{FP}{\pi_-}} = \frac{TP}{FP}\frac{\pi_-}{\pi_+} = \frac{PPV}{1-PPV}\frac{\pi_-}{\pi_+}\,.$$

Thus,

$$\frac{\partial PLR}{\partial PPV} = \frac{1}{(1-PPV)^2}\frac{\pi_-}{\pi_+} > 0\,.$$

As there is a strictly increasing relationship between PLR and PPV, these two scores lead to same performance ordering. □

A.4. Supplementary material about Sec. 4.3

A.4.1 Algorithmic contributions

It is easy to calculate, for a given importance I, which entity ϵ is the best among a set $\mathbb{E}$, based on their performances: it is the entity that has the highest score value. By performing this calculation at numerous points on the Tile, it is possible to get a good idea of the set of importances for which a given entity is the best, occupies the r-th place, or is ranked last. However, it is very interesting to be able to give an explicit representation of this set, in an exact way, by computing it efficiently. We now present some algorithmic contributions that make this possible. We present them for the particular case in which all compared performances have the same class priors.

When $\pi_- = \pi_+$. Let us first consider the case in which two performances with balanced priors, P_1, P_2, are compared. In our toy example (Fig. 5), we made the observation that the regions where the different classifiers are ranked first are convex polygons when $\pi_- = \pi_+$. In fact, one can show that the canonical importances $I_{a,b}$ for which P_1 is better or equivalent than P_2 are given by a linear inequality in the variables a and b:

$$P_1 \gtrsim P_2 \Leftrightarrow \lambda_a(P_1, P_2)a + \lambda_b(P_1, P_2)b + \lambda_0(P_1, P_2) \geq 0 \tag{6}$$

with

$$\begin{cases} \lambda_a(P_1, P_2) = FPR(P_1)FNR(P_2) - FNR(P_1)FPR(P_2) \\ \lambda_b(P_1, P_2) = TPR(P_1)TNR(P_2) - TNR(P_1)TPR(P_2) \\ \lambda_0(P_1, P_2) = TNR(P_1) - TNR(P_2) \end{cases} . \tag{7}$$

Given that the Tile is bounded by four other linear inequalities in the variables a and b,

$$\begin{cases} (1)a + (0)b + (0) \geq 0\,. \\ (-1)a + (0)b + (1) \geq 0\,. \\ (0)a + (1)b + (0) \geq 0\,. \\ (0)a + (-1)b + (1) \geq 0\,. \end{cases} \tag{8}$$

the zone in the Tile where P_1 is better or equivalent than P_2 is the intersection between 5 half-planes, thus either an empty set or a convex polygon. This can be trivially generalized to the computation of the zone in which all performances in a set Π_1 are better or equivalent to all performances in a set Π_2. In this case, we obtain either an empty set or a convex polygon that is the intersection of $|\Pi_1||\Pi_2| + 4$ half planes. If needed, the representation of this polygon as a set of linear inequalities can be converted in an equivalent representation based on its vertices and edges.

When $\pi_- \neq \pi_+$. In this case, we proceed in three steps: (1) we apply a tartget/prior shift to the performances to fall back in the case of uniform priors, then (2) we calculate the polygon as described above, and finally (3) we cancel the effect of a tartget/prior shift by applying a deformation to this polygon according to what was described in Appendix A.2.1. This transformation is continuous and invertible, so that the border of the computed zone corresponds to the contour of the polygon. For that reason, it is convenient to represent the polygons based on their vertices and edges. Applying the transformation to them can be done by simply discretizing the edges and applying the transformation to the points. In this way, the resulting zone is approximated as a polygon.

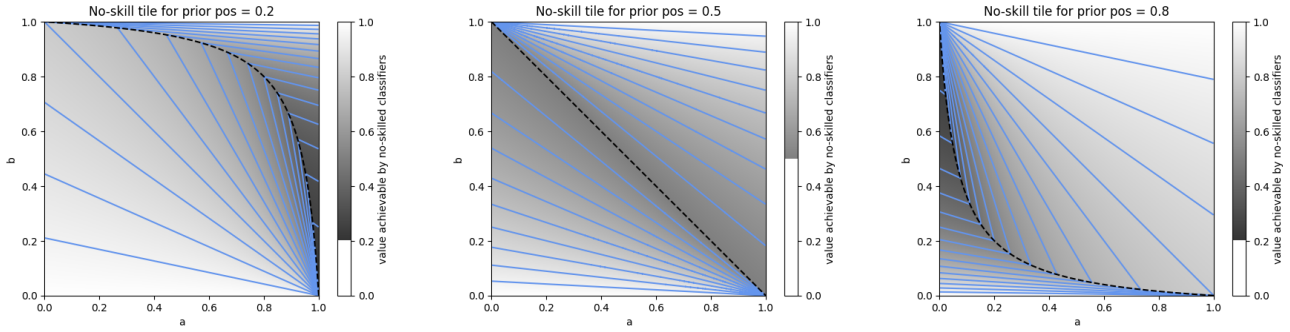


Figure 9. Tiles showing the values of canonical ranking scores achievable by no-skill classifiers when the prior of the positive class is 0.2 (left), 0.5 (center), and 0.8 (right). In all cases, the Tile is divided into two parts. On the bottom left part, the best no-skill classifier is the one predicting always the negative class. In the upper-right part, the best no-skill classifier is the one predicting always the positive class. Note that the limit between the two parts is the curve γ_π given in Fig. 6.

A.5. Supplementary material about Sec. 4.4

A.5.1 What is achievable by no-skilled classifiers?

Figure 9 uses the Tile to depict, for all canonical ranking scores, the maximum value achievable by the no-skill performances.

A.6. Averaging all canonical ranking scores: the *Volume Under Tile*

A.6.1 Definition

Let us study the score VUT (*Volume Under Tile*):

$$VUT : \mathbb{P}_{(\Omega,\Sigma)} \to [0,1] : P \mapsto VUT(P) = \int_{a=0}^{1} \int_{b=0}^{1} R_{I_{a,b}}(P)\,\mathrm{d}b\,\mathrm{d}a\,. \tag{9}$$

A.6.2 Closed-form expression

It can be showed that:
1. When $P(\{tp\}) = P(\{tn\})$ and $P(\{fn\}) = P(\{fp\})$:

$$\begin{aligned} VUT(P) &= R_{I_{a,b}}(P)\,\forall(a,b) \in [0,1] && (10)\\ &= A(P) = TNR(P) = TPR(P) && (11)\\ &= NPV(P) = PPV(P) = F_\beta(P) = \dots && (12)\end{aligned}$$

2. When $P(\{tp\}) = P(\{tn\})$ and $P(\{fn\}) \neq P(\{fp\})$:

$$VUT(P) = \frac{P(\{tn\})}{P(\{fn\}) - P(\{fp\})}\left(\ln\left(P(\{tn, fn\})\right) - \ln\left(P(\{tn, fp\})\right)\right) \tag{13}$$

3. When $P(\{tp\}) \neq P(\{tn\})$ and $P(\{fn\}) = P(\{fp\})$:

$$VUT(P) = 1 - \frac{P(\{fn\})}{P(\{tp\}) - P(\{tn\})}\left(\ln\left(P(\{tp, fn\})\right) - \ln\left(P(\{tn, fn\})\right)\right) \tag{14}$$

4. When $P(\{tp\}) \neq P(\{tn\})$ and $P(\{fn\}) \neq P(\{fp\})$:

$$\begin{aligned} VUT(P) = \frac{1}{2} - \frac{1}{2}\frac{\begin{array}{rl} & \left(P(\{tn\})^2 - P(\{fn\})^2\right)\ln\left(P(\{tn, fn\})\right)\\ + & \left(P(\{tp\})^2 - P(\{fp\})^2\right)\ln\left(P(\{tp, fp\})\right)\\ + & \left(P(\{fp\})^2 - P(\{tn\})^2\right)\ln\left(P(\{fp, tn\})\right)\\ + & \left(P(\{fn\})^2 - P(\{tp\})^2\right)\ln\left(P(\{fn, tp\})\right)\end{array}}{\left(P(\{tp\}) - P(\{tn\})\right)\left(P(\{fn\}) - P(\{fp\})\right)} \end{aligned} \tag{15}$$

Proof of Eq. (15). For the sake of concision, let us rewrite under volume under the Tile as the following double integral, where a, b, c, d are positive numbers:
$\int_{x=0}^{1}\int_{y=0}^{1} \frac{(1-x)a+xd}{(1-x)a+(1-y)b+yc+xd}dxdy$
$= \int_{x=0}^{1}\int_{y=0}^{1} \frac{(d-a)x+a}{(d-a)x+a+b+(c-b)y}dxdy$
Let's substitute: $p = (d-a), q = a, r = (d-a) = p, s = a+b+(c-b)y$
Knowing that $\int \frac{px+q}{rx+s}dx = \int f(x)dx = \frac{px}{r} + \frac{1}{r}\left(q - \frac{ps}{r}\right)\ln|rx+s| + C$, we have
$\int_0^1 f(x)dx = \frac{p}{r} + \frac{1}{r}\left(q - \frac{ps}{r}\right)\ln|r+s| - \frac{1}{r}\left(q - \frac{ps}{r}\right)\ln|s|$
As $p = r$, we can simplify the equation as:
$\int_0^1 f(x)dx = 1 + \left(\frac{q}{r} - \frac{s}{r}\right)\ln|r+s| - \left(\frac{q}{r} - \frac{s}{r}\right)\ln|s|$
$= 1 + \frac{a-a-b+(b-c)y}{d-a}\ln|d-a+a+b+(c-b)y| - \left(\frac{-b+(b-c)y}{d-a}\right)\ln|a+b+(c-b)y|$
$= 1 + \left[\frac{(b-c)}{d-a}y - \frac{b}{d-a}\right]\ln|b+d+(c-b)y| - \left[\frac{(b-c)}{d-a}y - \frac{b}{d-a}\right]\ln|a+b+(c-b)y|$
Let's now substitute: $\alpha = \frac{b-c}{d-a}, \beta = -\frac{b}{d-a}, \gamma = b+d, \delta = (c-b), \varepsilon = a+b$
$\to \int_0^1 1dy + \int_0^1(\alpha y+\beta)\ln\underbrace{|\delta y+\gamma|}_{\geqslant 0}dy - \int_0^1(\alpha y+\beta)\ln\underbrace{|\delta y+\varepsilon|}_{\geqslant 0}dy$
Knowing that $\int(ax+b)\ln(cx+d)dx = \int g(x)dx$
$= \frac{1}{4c^2}[2(cx+d)\ln(cx+d)(acx-ad+2bc) - cx(acx-2ad+4bc)] + C$,
we have

$$\int_0^1 g(x)dx = \frac{1}{4c^2}[2(c+d)\ln(c+d)(ac-ad+2bc) - c(ac-2ad+4bc)]$$
$$- \frac{1}{4c^2}[2d\ln(d)(2bc-ac)]$$

$$= 1 + \frac{1}{4\delta^2}\left[2(\delta+\gamma)\ln(\delta+\gamma)(\alpha\delta-\alpha\gamma+2\beta\delta)\right.$$
$$\left. - \delta\left(\alpha\delta-2\alpha\gamma+4\beta\delta\right) - 2\gamma\ln(2\beta\delta-\alpha\delta)\right]$$
$$- \frac{1}{4\delta^2}\left[2(\delta+\varepsilon)\ln(\delta+\varepsilon)(\alpha\delta-\alpha\varepsilon+2\beta\delta)\right.$$
$$\left. - \delta\left(\alpha\delta-2\alpha\varepsilon+4\beta\delta\right) - 2\varepsilon\ln(2\beta\delta-\alpha\delta)\right]$$
$$= 1 + \frac{1}{4(c-b)^2(d-a)}\left[2(c+d)\ln(c+d)\left[-(b-c)^2-(b-c)(b+d)+2b(b-c)\right]\right.$$
$$- 2(c+a)\ln(c+a)\left[-(b-c)^2-(b-c)(b+a)+2b(b-c)\right]$$
$$- 2(b+d)\ln(b+d)[2(c-b)(-b)+(c-b)(b+d)]$$
$$+ 2(a+b)\ln(a+b)[2(c-b)(-b)+(c-b)(b+a)]$$
$$- (c-b)\left[-(c-b)^2-2(b-c)(b+d)-4b(c-b)\right]$$
$$\left. + (c-b)\left[-(c-b)^2-2(b-c)(b+a)-4b(c-b)\right]\right]$$
$$= 1 + \frac{1}{4(c-b)(d-a)}[2(c+d)\ln(c+d)[(b-c)+(b+d)-2b]$$
$$- 2(c+a)\ln(c+a)[(b-c)+(b+a)-2b]$$
$$- 2(b+d)\ln(b+d)[d-b]$$
$$+ 2(a+b)\ln(a+b)[a-b]$$
$$+ \left[(c-b)^2+2(b-c)(b+d)+4b(c-b)\right]$$
$$- \left[(c-b)^2+2(b-c)(b+a)+4b(c-b)\right]$$

The volume under the Tile is thus analytically expressed as:

$$VUT = \frac{1}{2} - \frac{1}{2(c-b)(d-a)}\left[\left(c^2-d^2\right)\ln(c+d)\right.$$
$$+ \left(a^2-c^2\right)\ln(a+c)$$
$$+ \left(d^2-b^2\right)\ln(b+d)$$
$$\left. + \left(b^2-a^2\right)\ln(a+b)\right]$$

□ □

A.6.3 Discussion: can we use it to rank?

The ordering induced by VUT does not satisfy Axiom 3. Nevertheless, it is interesting to note that it has a very high rank correlation with the accuracy A (Spearman's ρ is about 0.996 for a uniform performance distribution, *i.e.* a Dirichlet distribution with all concentration parameters set to $\alpha = 1$).