Machine Learning (2025) 114:41
https://doi.org/10.1007/s10994-024-06654-4

Adapting performance metrics for ordinal classification to interval scale: length matters

Giulia Binotto[1] · Rosario Delgado[1]

Received: 16 May 2024 / Revised: 18 November 2024 / Accepted: 19 December 2024 /
Published online: 27 January 2025

Abstract

In the field of supervised machine learning, accurate evaluation of classification models is a critical factor for assessing their performance and guiding model selection. This paper delves into the domain of ordinal classification and raises the question of adapting ordinal metrics to the interval scale. In scenarios where measurements are recorded at intervals, not only the order but also their length assume significance, and this promotes the adoption of novel performance metrics. Initially, we revisit two existing confusion matrix-based ordinal metrics and introduce a normalization technique to render them comparable and enhance their practical utility. We extend our focus to classification by intervals, proposing a robust framework for adapting ordinal metrics to the interval scale, and applying it to the aforementioned ordinal metrics. We address the challenge of unbounded rightmost intervals, a common issue in practical applications, from both theoretical and simulation perspectives, by providing a solution that enhances the applicability of the proposed metrics. To further explore practical implications, we conducted experiments on real-world datasets. The results reveal a promising trend in the use of interval-scale metrics to guide hyper-parameter tuning for improving model performance.

Keywords Ordinal classification · Interval-scale classification · Performance metrics · Cost-sensitive metrics · Hyper-parameter tuning

Editor: Willem Waegeman.

✉ Giulia Binotto
Giulia.Binotto@uab.cat

Rosario Delgado
Rosario.Delgado@uab.cat

[1] Department of Mathematics, Autonomous University of Barcelona UAB, Av. de l’Eix Central s/n, 08193 Cerdanyola del Vallès, Spain

1 Introduction

Classification is one of the tasks in supervised machine learning. Its primary objective is to predict the suitable class or label for given input data, accomplished through the utilization of a predictive model, commonly referred to as a classifier. Model validation is performed using a previously unseen test data set: by comparing the predicted labels to the actual observations within the test set, a relevant metric can be formulated and employed to assess the predictive efficacy of the model.

In *multi-class* classification, where labels lack a natural ordering, a *nominal scale* is employed. Examples include car brand, type of job, political party, pets, sport, and more. However, in cases where classes possess an inherent ordering, and only their rank matters, we use an *ordinal scale*. For instance, when gauging customers' opinions in online shopping, we might consider categories like `would not recommend`, `would recommend`, or `would highly recommend`. This exemplifies a Likert-type scale, typically comprising 3, 5, or 7 points that respondents use to express their level of agreement or disagreement with a statement. Other examples include `completely disagree`, `disagree`, `neutral`, `agree` and `completely agree`, or `always`, `often` and `sometimes`. This scale was introduced by Likert (1932) and is widely employed in sentiment analysis, satisfaction surveys, opinion mining, and, more recently, information retrieval processes within the context of recommender systems (Pang and Lee, 2008). Responses on an ordinal scale can be scored or ranked, but the gaps between categories are not quantifiable. In other words, using an example, it is not meaningful to compare the "distance" between `often` and `always` with the "distance" between `often` and `sometimes`.

After the nominal and ordinal scales, the *interval scale* represents the third level of measurement. The interval scale is an ordered quantitative measurement scale that records measurements at intervals, not necessarily of equal length, along a common scale with endpoints that can be any real numbers. The differences between these intervals hold meaning.

To illustrate, consider a straightforward example. Let's say we are classifying individuals' heights (in cm) into three categories: `short`, `average` and `tall`. We have a data set comprising 100 height measurements, which have been categorised according to two different scales:

Scale 1: Equal Interval Lengths: 161–170, 171–180, and 181–190
Scale 2: Unequal Interval Lengths: 161–165, 166–175, and 176–190

Assume that the number of individuals assigned to any of the categories is the same on both scales: 40 are `short`, 30 are `average` and the remaining 30 are `tall`. Our goal is to compare classifiers A and B using their respective confusion matrices. In these matrices, rows represent the predicted category, while columns represent the observed category, for each of the 100 measurements:

$$C_A = \left(\begin{array}{ccc} & \text{observed} & \\ \text{short} & \text{average} & \text{tall} \\ 35 & \mathbf{5} & 0 \\ 5 & 25 & 0 \\ 0 & 0 & 30 \end{array}\right) \quad C_B = \left(\begin{array}{ccc} & \text{observed} & \\ \text{short} & \text{average} & \text{tall} \\ 35 & 0 & 0 \\ 5 & 25 & 0 \\ 0 & \mathbf{5} & 30 \end{array}\right)$$

With both scales, the two classifiers achieve the same accuracy, correctly predicting 90% of the heights. However, the confusion matrices yield different comparisons depending on the scale. Using scale 1, which is equivalent to the ordinal scale, both classifiers perform similarly. With scale 2, instead, classifier B exhibits poorer predictive performance. Specifically, each misclassified `average` height for C_A results in an error of at most 15 cm. In contrast, for C_B, the error can be as high as 25 cm. This disparity arises from the differing lengths of the `short` and `tall` intervals, which are 5 cm. and 15 cm., respectively, while the `average` interval is 10 cm. in both cases. This underscores the critical importance of considering interval length when evaluating classifiers on an interval scale.

Interval data, characterised by its simplicity and quantifiability, is useful across various domains, including business, social and physical sciences, and healthcare. In statistics, the interval scale is preferable to the qualitative nominal and ordinal scales due to its ability to assign numerical values to intervals, typically represented by their midpoints. Consequently, it facilitates the calculation of measures related to central tendency and dispersion, along with the creation of graphical representations of data. This quantitative approach extends to the assessment of subjective perceptions, such as emotions, sentiments, and other arbitrary evaluations.

1.1 More examples and related works

It is not uncommon to encounter classification problems where the target variable is derived by binning a count variable. For instance, the length of a patient's stay in the Intensive Care Unit of a hospital could be categorised as: `short` (1-2 days), `moderate` (3-7 days), `long` (8-14 days), or `extremely long` (more than 14 days). Likewise, the number of times a criminal offender re-enters the penal system can be grouped into categories such as `none`, `one`, `two`, or `more than two`. Similarly, it's quite common to predict a continuous variable within grouped value intervals. For instance, consider predicting a person's income categorised into intervals like `less than 20,000$`, `between 20,000$ and 40,000$`, and `more than 40,000$`. Grouped predictions in these cases offer greater practicality and interpretability. Grouping income into intervals facilitates a more meaningful analysis than using precise individual income values. It aids in identifying income trends, disparities, and specific population segments that may necessitate targeted interventions or policies. Additionally, grouping mitigates the impact of outliers and data variability, resulting in more robust and practical predictions for decision-making.

Discretization and binning entail the grouping of data, inevitably leading to information loss. Nevertheless, when we employ a classifier as a predictive model, this drawback is offset by the advantage of avoiding assumptions required by other predictive models, such as regression. Unlike regression models, which require the diagnosis of model assumptions, classifiers offer a more straightforward approach.

Among the multitude of works dedicated to addressing such scenarios, we would like to highlight a selection as illustrative examples, encompassing two distinct areas.

- *Customer Experience* (CX): The customer experience has evolved into a cornerstone of marketing strategy, with its primary focus on fostering customer loyalty and value. The best practices in the field of CX centre around the adoption of a CX interval-scale metric known as the *Net Promoter Score* (NPS) (Baehre et al., 2022). This metric is employed to gauge customers' sentiment by asking them how likely they are to recommend a product or company. Responses are recorded on an integer scale ranging from 0 to 10. Based on their responses, customers are categorised into one of three groups:

`Promoters`: These are customers who express high levels of satisfaction and rate the product or service with a score of 9 or 10.
`Passives`: Customers providing ratings of 7 or 8 are categorised as passives. They have a neutral experience, neither spreading negative word-of-mouth nor actively promoting the brand.
`Detractors`: Detractors are customers who rate the product or service below or equal to 6 and are dissatisfied with their experience.

To calculate NPS, the difference between the number of `promoters` and `detractors` is computed and divided by the total number of customers. This metric provides valuable insights into customer satisfaction and loyalty, aiding businesses in improving their operations and customer relationships. Apart from the NPS metric, business surveys also assess customer satisfaction regarding various CX attributes, such as product experience, value perception, touch-point experiences (e.g., call centre, website, mobile app, physical stores, etc.), and critical customer journeys (e.g., billing, service purchases). Analysing these surveys yields useful insights into a company's market position and areas ripe for improvement or differentiation. Linking the NPS metric with CX attribute performance, referred to as *key drivers analysis*, presents a multi-class classification challenge with the goal of predicting the interval-scale metric NPS.

Recognizing the inherent limitations of performance metrics within the multi-class setting, especially when compared to binary contexts, underscores the need for a specialized metric tailored to interval-scale classification. This is precisely the issue we address in our work. Other approaches have been considered previously. For example, the authors in (Jeske et al., 2011) employ multinomial Logistic Regression as a classification methodology to discern the key drivers of NPS in the context of a case study involving a healthcare company in Canada. They use *accuracy* as a metric to measure model performance on the multi-class confusion matrix, although this approach oversimplifies the problem. In contrast, in Markoulidakis et al. (2021), the authors take an indirect approach, avoiding a direct confrontation. Instead of proposing an appropriate metric, as we do, they introduce an innovative method designed to meet this need. Their approach involves oversimplifying the complexity of a multi-class confusion matrix through class collapsing. Specifically, they condense the 3 × 3 confusion matrices from the NPS classification problem into a series of binary confusion matrices.

- *Age Prediction:* Estimating a person's age is difficult due to the substantial variations in ageing patterns among individuals. These variations stem from factors like lifestyle, environmental conditions, and health. Numerous researchers have implemented age prediction models based on various individual characteristics. Here are some notable examples:

  - *Biometric traits.* Estimation of the age of a person based on facial features like wrinkles, freckles, and skin spots is a challenging endeavor. Age prediction using biometric traits has seen several noteworthy developments. Indeed, in 2013, Erbilek et al. (2013) introduced age prediction based on pupils (iris biometrics), categorizing individuals into three age groups: <25, 25–60, and >60. More recently, Sharma et al. (2021) explored age prediction from facial wrinkles, defining age groups as 20–40, 41–60, and >60. In another approach, the authors of Gowroju et al. (2022) devised an image recognition model based on deep neural networks to predict age groups from facial images, using pupil information. They grouped ages into six categories of unequal lengths: 10–20, 21–27, 28–45, 46–65, and 66–100. It is worth noting that the selection of an upper limit of 100 years appears arbitrary. While this choice doesn't impact their study when employing the ordinal scale metric MAE for comparisons, it would have held significant consequences had they used a metric designed for the interval scale, one that considers interval lengths.

- *Oral speech.* In response to the substantial demand from various applications, such as personal voice assistants, criminal tracking, and user profiling, for models capable of swiftly and accurately estimating a speaker's age, other researchers have delved into predicting age groups based on speech features. For instance, in Ravishankar et al. (2020), a multi-layer perceptron is employed as a predictive model, categorizing individuals into age groups as follows: <19, 19–29, 30–39, 40–49, 50–59, 60–69, 70–79 and 80–89.
- *Writing.* Another domain where automatic age prediction finds applications is social media. In Peersman et al. (2011), a text categorization approach is employed to predict both age and gender. This study utilizes a corpus of chat texts sourced from the Belgian social networking site Netlog, with age groups classified as `10 s`, `20 s`, `30 s`, and `Plus40s`. Meanwhile, in Morgan-López et al. (2017), authors develop a model leveraging various linguistic and metadata features to predict the age of Twitter users. The age categories used in this model are `youth` (13–17), `young adults` (18–24), and `adults` (≥ 25).

1.2 Problem statement and objectives

After establishing the rationale for the necessity of performance metrics tailored to interval-scale classification, a natural question arises: How can we construct such metrics using pre-existing ones designed for the ordinal scale?

On one hand, including information about the interval lengths in the definition of the metric contradicts the fundamental concept of ordinal classification, according to which the intervals that define the classes are assumed to lack intrinsic meaning (Stevens, 1946). Conversely, this approach aligns with what seems intuitively logical. This paradox has captivated our attention, representing the novelty and intrigue of the problem at hand. Here, the lengths of the intervals defining categories not only can but must play a pivotal role in metric construction. To the best of our knowledge, such an endeavor remains unexplored in the existing literature. To bridge this gap, our objectives in this work are as follows:

1. Introducing a comprehensive and easy to apply methodology for adapting any ordinal metric defined on the confusion matrix to the interval scale.

   We elaborate on this general methodology in Sect. 3, illustrating this application with two metrics: the Mean Absolute Error (MAE), a frequently used metric in ordinal classification, and the cost-sensitive metric referred to as Total misclassification Cost (TC) (George et al., 2016). To provide a thorough understanding, both of these metrics are defined in advance in Sect. 2.
2. Introducing a wide-ranging normalization method to constrain any metric defined on the confusion matrix within the interval $[0, 1]$.

   In Sect. 3, we apply this method to both the original ordinal-scale metrics and their interval-scale counterparts, enabling meaningful comparisons.

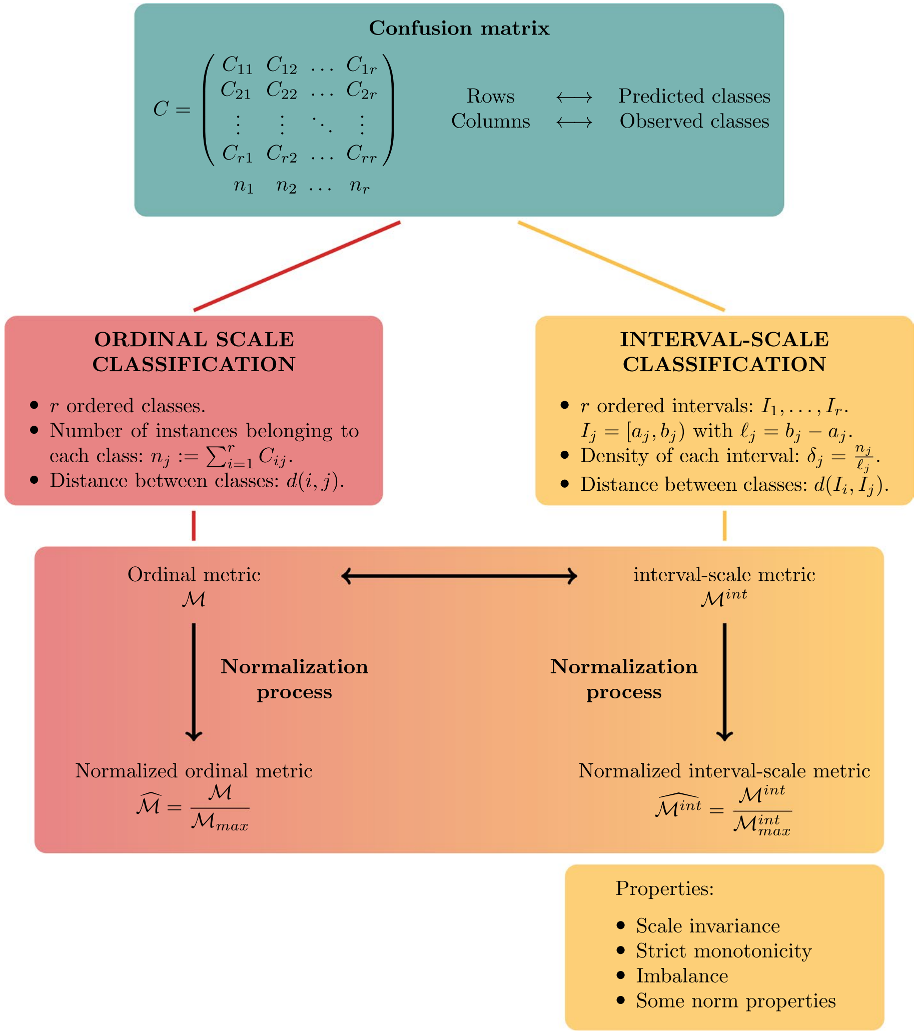

Confusion matrix
$C = \begin{pmatrix} C_{11} & C_{12} & \dots & C_{1r} \\ C_{21} & C_{22} & \dots & C_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ C_{r1} & C_{r2} & \dots & C_{rr} \end{pmatrix}$
$n_1 \quad n_2 \quad \dots \quad n_r$
Rows ⟷ Predicted classes
Columns ⟷ Observed classes
ORDINAL SCALE CLASSIFICATION
• r ordered classes.
• Number of instances belonging to each class: $n_j := \sum_{i=1}^{r} C_{ij}$.
• Distance between classes: $d(i, j)$.
INTERVAL-SCALE CLASSIFICATION
• r ordered intervals: $I_1, \dots, I_r$. $I_j = [a_j, b_j)$ with $\ell_j = b_j - a_j$.
• Density of each interval: $\delta_j = \frac{n_j}{\ell_j}$.
• Distance between classes: $d(I_i, I_j)$.
Ordinal metric
$\mathcal{M}$
interval-scale metric
$\mathcal{M}^{int}$
Normalization process
Normalization process
Normalized ordinal metric
$\widehat{\mathcal{M}} = \frac{\mathcal{M}}{\mathcal{M}_{max}}$
Normalized interval-scale metric
$\widehat{\mathcal{M}^{int}} = \frac{\mathcal{M}^{int}}{\mathcal{M}^{int}_{max}}$
Properties:
• Scale invariance
• Strict monotonicity
• Imbalance
• Some norm properties

Fig. 1 Schematic representation of the methodologies presented in this study

Diagram in Fig. 1 provides a visual representation of this two procedures. To strengthen them from a theoretical perspective, we demonstrate some desirable properties (Amigó et al., 2020; Sebastiani, 2015) fulfilled by the metrics constructed in this way.

Another significant challenge we faced is as follows: the methodology introduced to adapt metrics from ordinal to interval scale inherently relies on the length of intervals. However, what should be done when the rightmost interval lacks an assigned length, rendering it unconstrained? This scenario commonly arises when discretizing a continuous variable or binning a count variable, particularly in cases where there is no readily defined upper bound, as exemplified in previous examples. Addressing this challenge has been the primary focus of Sect. 4. After presenting a toy example to elucidate the problem, we approach it from an analytical perspective in Sect. 4.1. Specifically, we delve into the search for the optimal length to be assigned to the unbounded rightmost interval for the purpose of computing the performance metric. Here, we establish an upper bound for this optimal interval length in the general case (Theorems 1, 2). Furthermore, we provide a comprehensive solution by determining its value in the balanced case involving three classes (Proposition 6). In Sect. 4.2, we demonstrate the practical application of the algorithm we developed to provide an approximate numerical solution for the optimal length. Additionally, to explore the practical implications of the proposed interval-scale metrics, we conducted experiments detailed in Sect. 5. These experiments compared interval-scale metrics with traditional metrics, such as Accuracy and MAE, to evaluate their effectiveness in guiding the grid search process for hyper-parameter tuning to improve classifier performance. The study involved three real-world datasets and evaluated two classifiers: random forest and k-nearest neighbors ($k$nn).

The structure of the remainder of this manuscript unfolds as follows: we start by introducing mainstream metrics MAE and TC for ordinal classification based on the confusion matrix in Sect. 2. In Sect. 3, we present the methods discussed earlier and apply them, step-by-step, to metrics MAE and TC, ultimately obtaining their normalized interval-scale versions. Noteworthy properties of these newly derived metrics are also presented. Their proofs are given in Appendix A. Sect. 6 serves as the concluding segment of this manuscript, providing some final remarks. In Appendix B, we furnish a proof for Proposition 6 from Sect. 4.

2 Metrics for ordinal classification

Despite its practical applications, as explained in the introduction, ordinal classification has been less developed to date than nominal scale classification. Even so, there are different performance metrics known in the literature that allow classifiers to be validated and compared when the class variable is ordinal, and we will recall some of them below. Evaluation metrics differ in how they handle classification errors. *Association metrics* rely on the agreement between two raters when classifying instances into ordered categories, with *Kendall's correlation coefficient* being among the most frequently employed. However, in this study, our primary emphasis lies on the commonly utilized metrics rooted in the confusion matrix. This section recalls the core set, encompassing a cost-sensitive metric introduced as TC in (George et al., 2016).

2.1 Notations

For ordinal scale classification, without loss of generality, we can assume that the classes $1, \dots, r$, $r \geq 3$, are in this order, and this implies that we assume as more serious the error of classifying an instance of class i as belonging to class $i+2$ or $i-2$ than as belonging to class $i+1$ or $i-1$, respectively.

We denote by $C = (C_{ij})_{i,j=1,\dots,r}$ a general confusion matrix obtained from any *validation* procedure, where C_{ij} is the number of instances in the test data set that belong to class j and have been assigned to class i by the classifier. Also, denote by $N = \sum_{i=1}^{r} \sum_{j=1}^{r} C_{ij}$ the total number of instances in the validation set, and by $n_j = \sum_{i=1}^{r} C_{ij}$ the number of instances belonging to class j, for $j = 1, \dots, r$, with $N = \sum_{j=1}^{r} n_j$. We assume that $n_j > 0$ for any $j = 1, \dots, r$ (otherwise, the class would be removed and we would be left with the remaining $r-1$ classes).

2.2 Mainstream metrics based on the confusion matrix

- Error Rate: It is simply defined as the fraction of incorrect predictions (that is, 1−*Accuracy*). It has the disadvantage that all errors are treated equally and therefore does not penalize classifiers who make flagrant errors. However, we will also consider this metric as a reference. The formula is: Error Rate $(C) = 1 - \frac{\sum_{i=1}^{r} C_{ii}}{N}$ and ranges between 0 and 1, the first corresponding to perfect classification.
- *Mean Absolute Error* (MAE): This very popular metric for ordinal classification is defined as the average of the absolute errors between the ground truth and the estimated class. The literature in several studies concludes that MAE is one of the best performance metrics in ordered classification. For example, in Gaudette and Japkowicz (2009) the authors experimentally show that for the unbalanced data set studied, MSE (Mean Squared Error) and MAE perform the best, but while MSE is better in situations where the severity of the errors is more important, MAE shows to be better in situations where the tolerance for small errors is lower. This is despite the fact that neither of these measures is truly ordinal by design. In Cano and García (2017) the authors use MAE as a performance metric for monotonic ordinal classification, to show the usefulness of selecting the training set to obtain more accurate and efficient models. More recently, MAE is used in Liu et al. (2020) to evaluate the medical diagnose task on two medical ordinal benchmarks, as well as the face age prediction, and in Vargas et al. (2020) it is employed to estimate the performance of a deep convolutional neural network model for ordinal classification.

  This metric penalizes classification errors proportionally to the distance between the categories, so the lower the metric value, the better the performance of the classifier. Its definition is the following:

$$\text{MAE}\ (C) = \frac{1}{N} \sum_{i,j=1}^{r} C_{ij}\ |i-j|$$

  where |*x*| denotes the absolute value of $x \in \mathbb{R}$. Note that we can write

$$\text{MAE}\ (C) = \sum_{j=1}^{r} \text{MAE}_j(C), \quad \text{with} \quad \text{MAE}_j(C) = \frac{1}{N} \sum_{i=1}^{r} C_{ij}\ |i-j| \tag{1}$$

  There are other measures for ordinal classification that are variations of MAE, such as the *Weighted Average of Mean Absolute Error* (AMAE) or the *Maximum of Mean Absolute Error* (MMAE) (Cruz-Ramírez et al., 2011), which are useful if class sizes are unbalanced.

The problem with MAE and its variants is that they assume that all classes are equidistant, which does not have to be true when performing an ordinal classification task. For example, classification on a scale

`very bad`, `bad`, `acceptable`, `good` or `very good`

is consequence of a subjective appreciation that will hardly correspond to equidistant numerical values. Works using these metrics, explicitly or implicitly assume that “misclassification costs are always proportional to the absolute difference between the actual and the predicted label” (Waegeman et al., 2006, *expressis verbis*). But this assumption goes against the basic principle of meaninglessness of the numerical values in ordinal classification, beyond their ordering (Amigó et al., 2020).

2.3 A cost-sensitive metric based on the confusion matrix

In George et al. (2016) a new cost-sensitive metric is introduced, named **Total misclassification Cost (TC)**. The proposed measure accounts for inherent ordinal data structure, the total misclassification cost of a classifier, and the unbalanced class distribution, and shows good performance in identifying the best ordinal classifier with some real data sets and simulation studies. Its definition is:

$$\mathrm{TC}\,(C) = \sum_{j=1}^{r} \mathrm{TC}_j(C), \quad \text{with} \quad \mathrm{TC}_j(C) = \sum_{i=1}^{r} C_{ij}\, \gamma_{ij}\, |i-j|, \tag{2}$$

where

$$\gamma_{ij} = \frac{\sum_{k\neq j}^{r} n_k}{n_i} = \frac{N-n_j}{n_i} \quad \text{(which is } \geq 1 \text{ if } i \neq j).$$

(Compare with expression (1) for MAE, in which γ_{ij} is just $1/N$.)

The rationale behind this definition is as follows: this measure uses information from the class distribution and domain knowledge about the ordinal class structure, and is the sum of the misclassification costs for any class $j = 1, \dots, r$. Indeed, for instances that are of class j, the misclassification cost takes into account not only the distance between predicted class i and the true class labels, $|i-j|$ (the higher, the higher cost), but the size of the classes, in the sense that the smaller the size of the class i, or of the class j, the higher the cost. That is, this measure penalizes more cases of misclassification when assigning a small class than of a large one, and penalizes more misclassification when assigning from a small class than from a large one. This is captured in (2) by γ_{ij}, which is defined as the inverse of the probability of misclassifying an object in class i given that the object is of class j and has been misclassified, if the classification is done randomly (i.e: a label chosen randomly from the available labels has been assigned).

Remark 1 We can introduce the **cost matrix** $W = (\omega_{ij})_{i,j=1,\dots,r}$ by

$$\omega_{ij} = \gamma_{ij}\, |i-j|$$

that is, ω_{ij} is the cost associated to misclassify an instance belonging to class j in class i. When is the cost matrix W symmetric? W is symmetric if for all $i \neq j$, $\gamma_{ij} = \gamma_{ji}$, which is equivalent to say that

$$\sum_{k\neq i,j} \frac{n_k}{n_i} = \sum_{k\neq i,j} \frac{n_k}{n_j} \Leftrightarrow n_i = n_j.$$

Then, matrix W is symmetric if and only if there is a perfect balance between classes in the sense that $n_i = n_j$ for all $i,j = 1, \dots, r$ (that is, $n_i = N/r$ for all $i = 1, \dots, r$). With this notation, by (2),

$$\mathrm{TC}\,(C) = \sum_{i,j=1}^{r} C_{ij}\,\omega_{ij} = sum(C \odot W) \tag{3}$$

where $\odot$ denotes the element-wise (Hadamard) or Schur matrix product.

2.4 Other ordinal scale metrics: state-of-the-art

The authors of Baccianella et al. (2009) address the problem of imbalance, when certain classes are considerably more common than others, in the context of performance measures for ordinal classification. In this case, using a metric designed for balanced data sets may result in a situation in which a classifier that assigns always the majority class outperforms highly sophisticated classification systems. To overcome this problem, they introduce macro-averaged versions of the most common ordinal classification measures, which are more resilient to imbalance and equivalent to the standard versions when the data sets are balanced. The same problem has been considered by Yilmaz and Demirhan (2023), proposing a solution based on weighted agreement measures, such as Cohen's κ, Scott's π, Gwet and Brennan-Prediger, where the weighting schemes considered are linear, quadratic, ordinal, radical and bipolar weights, concluding from the experimental phase with real data sets that Cohen's κ and Scott's π with quadratic weights perform better than the other considered metrics. The *quadratic weighted kappa* QWK metric was previously introduced in Ben-David (2008) as a method to addres cost-sensitive classification, especially concerning ordinal classes, and has recently been employed by Liu et al. (2020) and Vargas et al. (2020), among others.

The lack of adaptability to imbalance is not the only issue that the standard metrics for ordinal classification exhibit. In fact, as evidenced by Cardoso and Sousa (2011), where an alternative measure that prevents this flaws is proposed, some of them, such as MAE and its derivatives, have the disadvantage of being dependent on the numbers chosen to represent the classes, whereas Kendall's τ_b outperforms this problem at the price of losing information about absolute predictions. As a result, this metric is better suited for assessing preference learning than ordinal classification.

Finally, in Cruz-Ramírez et al. (2011, 2014) two new measures for ordinal classification are introduced: the maximum and the minimum of mean absolute error of all the classes, which take into account the per-class distribution of patterns as well as the magnitude of the error, and they propose using the first of them, jointly with the mean absolute error, as a pair of metrics to drive a multi-objective evolutionary algorithm, since they are competitive objectives.

3 Adapting an ordinal metric based on the confusion matrix to the interval scale

The aim here is to adapt the performance metrics employed for ordinal classification to the interval scale. After recalling the essential features of ordinal metrics, we give a definition for the case in which the classes are represented by intervals, highlighting the symmetry between the two scales. Then, we define a normalization process for both ordinal and interval-scale classification to force the metric to reside in the interval [0, 1]. This is a crucial step to enable performance metrics comparison. This whole procedure is schematized in Fig. 1. In order to illustrate the procedure in more detail, we apply it step-by-step to metrics MAE and TC. At each stage, we investigate the properties of the new metrics to gain insight of their performance.

In what comes next, unless otherwise indicated, we will assume that r, the number of classes, and $n_1, \dots, n_r$, the number of cases in the test set that belong to any of the classes, are positive and fixed, with $N = \sum_{j=1}^{r} n_j$. When we refer to ordinal classification, the classes are $1 < 2 < \cdots < r$, while when referring to the interval scale, $1, \dots, r$ are assumed to be the sub-indexes of the classes-intervals $I_1, \dots, I_r$, such that $I_j = [a_j, b_j)$ with $a_j < b_j$ for $j = 1, \dots, r$ and $a_j = b_{j-1}$ for $j = 2, \dots, r$. Their corresponding lengths are $\ell_j = b_j - a_j > 0$, $j = 1, \dots, r$, and we are assuming that the intervals are ordered, in the sense that

$$\forall x \in I_i,\ y \in I_j, \qquad \text{if } i < j \text{ then } x < y.$$

3.1 Transition to the interval scale

Let $\mathcal{M}$ be an arbitrary ordinal metric on the classes $1, \dots, r$ based on the confusion matrix, which measures in some sense the error in classification. Then, $\mathcal{M}$ operates on the (finite) set of matrices:

$$\mathcal{S} = \left\{ r \times r \text{ matrices } A = (a_{ij})_{i,j=1,\dots,r} \text{ with } a_{ij} \in \mathbb{N} \text{ and } \sum_{i=1}^{r} a_{ij} = n_j,\ j = 1, \dots, r \right\}$$

When needed, the dependence of $\mathcal{S}$ on the values $n_1, \dots, n_r$ shall be specified in the notation as $\mathcal{S}^{n_1,\dots,n_r}$. Following (3), we also assume that for any $A \in \mathcal{S}$, $\mathcal{M}(A)$ can be expressed as $sum(A \odot \mathcal{W})$ where $\mathcal{W} = (w_{ij})_{i,j=1\dots,r}$ is a generic cost matrix that depends on the classes only through their distances, say $d(i, j)$ for $i,j = 1, \dots, r$ (for example, $d(i,j) = |i - j|$ or $(i - j)^2$), and also depends on the quantities $n_1, \dots, n_r$ associated to $\mathcal{S}$, but of no other quantity related to matrix A.

Definition 1 Given such an ordinal metric $\mathcal{M}$ on $\mathcal{S}$, we introduce its counterpart metric for the interval scale on $\mathcal{S}$, and denote it by $\mathcal{M}^{int}$, by

$$\mathcal{M}^{int}(A) = sum(A \odot \mathcal{W}^{int}) \tag{4}$$

for any $A \in \mathcal{S}$, where $\mathcal{W}^{int}$ is a cost matrix obtained by modifying the cost matrix $\mathcal{W}$ associated to $\mathcal{M}$ in this way:

- The distance between real numbers $d(i, j)$ is substituted by a distance between intervals, $d(I_i, I_j)$, such as the Hausdorff distance (although other distances could also be considered), which is defined by

$$d(I_i, I_j) = \max\{\,|a_j - a_i|,\ |b_j - b_i|\,\}\ ; \tag{5}$$

- n_j, which is the number of instances in the j-class, is substituted by δ_j, which represents the idea of "density" of interval I_j, defined as the number of instances belonging to that interval divided by its length, that is,

$$\delta_j = \frac{n_j}{\ell_j}. \tag{6}$$

3.1.1 The interval-scale metrics MAE^{int} and TC^{int}

We proceed by adapting the ordinal metrics MAE and TC to the interval scale following the ideas of Definition 1. In what follows, we assume that $A \in \mathcal{S}$.

Definition 2 The interval-scale version of metric MAE is

$$\mathrm{MAE}^{int}(A) := \frac{1}{N} \sum_{i,j=1}^{r} a_{ij}\, d(I_i, I_j),$$

where the distances $d(I_i, I_j)$ between the intervals are given by (5).

Observe that, for this new metric, the cost matrix $\mathcal{W}^{int}$ introduced in (4) has components

$$w_{ij} = \frac{1}{N} d(I_i, I_j), \qquad i,j = 1, \dots, r.$$

Definition 3 The interval-scale version of metric TC is

$$\mathrm{TC}^{int}(A) := \sum_{i,j=1}^{r} a_{ij}\, \gamma_{ij}^{int}\, d(I_i, I_j)$$

where

$$\gamma_{ij}^{int} = \frac{\sum_{k \neq j}^{r} \delta_k}{\delta_i} \quad \text{(which is } \geq 1 \text{ when } i \neq j), \tag{7}$$

the distances $d(I_i, I_j)$ between the intervals are given by (5), and the "densities" δ_i are defined by (6).

In this case, the cost matrix $\mathcal{W}^{int}$ introduced in (4) has components

$$w_{ij} = \gamma_{ij}^{int}\, d(I_i, I_j), \qquad i,j = 1, \dots, r.$$

The following proposition show some properties of the newly introduced metrics. Let $\mathcal{A}$ denote $\mathcal{A} = \{r \times r \text{ matrices } A = (a_{ij})_{i,j=1,\dots,r} \text{ with } a_{ij} \in \mathbb{N}\}$.

Proposition 1 MAE^{int} *verifies the following properties of a* **norm** *on* $\mathcal{A}$:

1. *Non-negativity/Positiveness*:

$$\mathrm{MAE}^{int}(A) \geq 0, \textit{ for all } A \in \mathcal{A}.$$

2. *Positive definiteness/Maximal agreement*:

$$\text{For } A \in \mathcal{A},\ \mathrm{TC}^{int}(A) = 0 \Longleftrightarrow A \text{ is diagonal.}$$

Note that if $A \in \mathcal{S}^{n_1,\dots,n_r}$, A is diagonal means that diag$(A) = (n_1, \dots, n_r)$.

3. *Homogeneity* (*of degree* 0):

$$\text{For } A \in \mathcal{A} \text{ and } k > 0,\ \mathrm{MAE}^{int}(kA) = \mathrm{MAE}^{int}(A).$$

4. *Subadditivity/Triangle inequality*:

$$\text{For } A, B \in \mathcal{A},\ \mathrm{MAE}^{int}(A + B) \leq \mathrm{MAE}^{int}(A) + \mathrm{MAE}^{int}(B).$$

TC^{int} *verifies properties 1. and 2., but instead of property 3. it is homogeneous of degree 1, and the triangle inequality is satisfied with some restrictions*:

3'. *Homogeneity* (*of degree 1*):

$$\text{For } A \in \mathcal{A} \text{ and } k > 0,\ \mathrm{TC}^{int}(kA) = k\,\mathrm{TC}^{int}(A).$$

4'. (*Restricted*) *Subadditivity/Triangle inequality (in fact, the equality holds)*:

$$\text{For } A, B \in \mathcal{A},\ \text{if for any } j = 1, \dots, r,\ \sum_{i=1}^{r} a_{ij} = \sum_{i=1}^{r} b_{ij},\ \text{then}$$

$$\mathrm{TC}^{int}(A + B) \leq \mathrm{TC}^{int}(A) + \mathrm{TC}^{int}(B).$$

The proofs of these properties can be found in Appendix A.1.

3.2 Normalization process on $\mathcal{S}$

The following procedure is used to normalize any metric defined on $\mathcal{S}$, forcing it to live in the interval [0, 1]. Such a metric could be of ordinal ($\mathcal{M}$) or interval ($\mathcal{M}^{int}$) scale. With an abuse of notation, in this subsection we will use $\mathcal{M}$ to refer to this metric, although all definitions are also valid for $\mathcal{M}^{int}$. We first define an equivalence relation in $\mathcal{S}$ by

$$A, B \in \mathcal{S} \text{ belong to the same class } (A \sim B) \text{ if } \mathcal{M}(A) = \mathcal{M}(B).$$

$[A]$ denotes the equivalence class of a matrix $A \in \mathcal{S}$, that is, $[A] = \{C \in \mathcal{S} \ :\ C \sim A\}$. The set of the equivalence classes of $\mathcal{S}$ with this relation is the quotient set denoted by $\mathcal{S}/\sim$, where we can define a total ordering $\preceq$ in this way: for $[A], [B] \in \mathcal{S}/\sim$,

$$[A] \preceq [B] \Longleftrightarrow \mathcal{M}(A) \leq \mathcal{M}(B)$$

This set is finite because $\mathcal{S}$ is. Indeed, for any column of matrix $A \in \mathcal{S}$, say column j, there is a finite number of ways to accommodate n_j among the r positions in the column. Since $\mathcal{S}/\sim$ is finite and it is a totally ordered set with $\preceq$, then there exists a (unique) maximum of $\left(\mathcal{S}/\sim, \preceq\right)$, say class $\mathcal{S}_{max}$. Let us define

$$\mathcal{M}_{max} = \mathcal{M}(A_{max}) \tag{8}$$

being A_{max} any matrix representing the class $\mathcal{S}_{max}$, that is, such that $[A_{max}] = \mathcal{S}_{max}$. Note that A_{max} depends on the metric $\mathcal{M}$ and on $n_1, \dots, n_r$. When necessary, we will make it explicit in the notation as $A_{max}^{n_1,\dots,n_r}$ (if the metric $\mathcal{M}$ is understood).

By definition, the value $\mathcal{M}(A_{max})$ is unique and therefore, $\mathcal{M}_{max}$ is well defined and only depends on quantities $n_1, \dots, n_r$, and not on the confusion matrix itself. Then, for any confusion matrix $C \in \mathcal{S}$, $\mathcal{M}(C) \leq \mathcal{M}_{max}$. This allows us to introduce the following definition:

Definition 4 Given a metric $\mathcal{M}$ on $\mathcal{S}$, we define its corresponding **normalized version**, and denote it by $\widehat{\mathcal{M}}$, as the metric in $[0,\ 1]$ obtained by dividing the metric by its maximum value, that is,

$$\widehat{\mathcal{M}} = \frac{\mathcal{M}}{\mathcal{M}_{max}} \in [0,\ 1],$$

where $\mathcal{M}_{max}$ is defined by (8). Said another way, for any $A \in \mathcal{S}$, $\widehat{\mathcal{M}}(A)$ is the proportion of $\mathcal{M}_{max}$ that represents $\mathcal{M}(A)$.

We have just described the procedure to normalize a metric by constraining it to the interval $[0,\ 1]$. This approach can be applied to both ordinal and interval-scale metrics. In what follows, we apply it to metrics MAE and TC and their interval-scale versions.

3.2.1 The normalized (ordinal) metrics $\widehat{\mathrm{MAE}}$ and $\widehat{\mathrm{TC}}$

We apply the normalization of a general metric on $\mathcal{S}$ introduced in Sect. 3.2 to the ordinal metrics MAE and TC. This yields normalized versions of these metrics constrained to the interval $[0,\ 1]$. As outlined in Definition 4, we achieve normalization by dividing MAE and TC by their respective maximum values. The normalized versions of the metrics MAE and TC, constrained within $[0,\ 1]$, are then defined as follows.

Definition 5 For any matrix $A \in \mathcal{S}$,

$$\widehat{\mathrm{MAE}}(A) = \frac{\mathrm{MAE}(A)}{\mathrm{MAE}_{max}} \quad \text{and} \quad \widehat{\mathrm{TC}}(A) = \frac{\mathrm{TC}(A)}{\mathrm{TC}_{max}}.$$

The following result provides practical expressions for calculating their maximum values.

Proposition 2 MAE_{max} *and* TC_{max} *can be expressed as the sum of one term for any of the classes as follows*:

(a) $\mathrm{MAE}_{max} = \sum_{j=1}^{r} G_j$, *with* $G_j = \frac{n_j}{N}\,|g_j - j|$ *and* $g_j = \arg\max_{\ell=1,\dots,r} |\ell - j|$.

(b) $\mathrm{TC}_{max} = \sum_{j=1}^{r} K_j$, *with* $K_j = \frac{n_j\,(N-n_j)}{n_{k_j}}\,|k_j - j|$ *and* $k_j = \arg\max_{\ell=1,\dots,r} \frac{|\ell-j|}{n_\ell}$.

Remark 2 Note that although k_j or g_j could not be unique, K_j and G_j are well defined.

Proof We only do the proof for b), because case a) is analogous.

By definition, $\mathrm{TC}_{max} = \mathrm{TC}(A_{max})$ where $A_{max} = (a_{ij})_{i,j=1,\dots,r}$ is given by:

$$\text{for any } j = 1, \dots, r, \quad a_{ij} = \begin{cases} n_j & \text{if } i = k_j \\ 0 & \text{otherwise} \end{cases}$$

and

$$k_j = \arg \max_{\ell=1,\dots,r} w_{\ell j} = \arg \max_{\ell=1,\dots,r} \gamma_{\ell j} \, |\ell - j| = \arg \max_{\ell=1,\dots,r} \frac{|\ell - j|}{n_\ell}.$$

Therefore,

$$\mathrm{TC}_{max} = \mathrm{TC}(A_{max}) = \sum_{i,j=1}^{r} a_{ij} \, \gamma_{ij} \, |i - j| = \sum_{j=1}^{r} n_j \, \gamma_{k_j j} \, |k_j - j| = \sum_{j=1}^{r} K_j.$$

□

We employ the expressions derived in the previous proposition to compute MAE_{max} and TC_{max} in two specific scenarios: the binary case and the balanced case. The proofs are provided in Appendix A.2.

Corollary 1 *In the binary case,*

$$\mathrm{MAE}_{max} = 1 \quad \text{and} \quad \mathrm{TC}_{max} = N.$$

Corollary 2 *In the balanced case $n_j = N/r$ for $j = 1, \dots, r$,*

$$\mathrm{MAE}_{max} = \frac{r-1}{2} + \frac{(r-h)\,h}{r}, \quad \mathrm{TC}_{max} = N \left(\frac{(r-1)^2}{2} + \frac{(r-1)\,(r-h)\,h}{r} \right)$$

where

$$h = \begin{cases} r/2 & \text{if } r \text{ is even,} \\ \lfloor r/2 \rfloor + 1 & \text{if } r \text{ is odd.} \end{cases}$$

Note that this implies that, in this scenario, $\mathrm{TC}_{max} = N\,(r-1)\,\mathrm{MAE}_{max}$.

As a result of these corollaries, we arrive at the following outcome.

Corollary 3 *In the balanced case,*

$$\mathrm{MAE}_{max} \begin{cases} = 1 & \text{if } r = 2 \\ > 1 & \text{if } r > 2 \end{cases} \quad \text{and} \quad \mathrm{TC}_{max} \begin{cases} = N & \text{if } r = 2 \\ > N & \text{if } r > 2, \end{cases}$$

with both MAE_{max} *and* TC_{max} *being strictly increasing functions of r.*

3.2.2 The normalized interval-scale metrics $\widehat{\mathrm{MAE}^{int}}$ and $\widehat{\mathrm{TC}^{int}}$

This step completes the adaptation process of ordinal metrics to the interval scale. By merging Definitions 1 and 4, the subsequent definition provide the normalized versions to the interval scale of the ordinal metrics MAE and TC.

Definition 6 For any matrix $A \in \mathcal{S}$,

$$\widehat{\mathrm{MAE}^{int}}(A) = \frac{\mathrm{MAE}^{int}(A)}{\mathrm{MAE}^{int}_{max}} \quad \text{and} \quad \widehat{\mathrm{TC}^{int}}(A) = \frac{\mathrm{TC}^{int}(A)}{\mathrm{TC}^{int}_{max}},$$

where MAE^{int} and TC^{int} have been introduced in Definitions 2 and 3, respectively.

Remark 3 Note that if all the intervals have the same length, say λ, therefore $\mathcal{M}^{int} = \lambda\,\mathcal{M}$ with $\mathcal{M}$ =MAE or TC, since in this case, for any $i,j = 1,\dots,r$, $d(I_i, I_j) = \lambda\,|i-j|$ and $\gamma^{int}_{ij} = \gamma_{ij}$. As a consequence, $\widehat{\mathcal{M}^{int}}$ coincides with the corresponding normalized ordinal metric $\widehat{\mathcal{M}}$.

Remark 4 Also note that with $\mathcal{M}$ =MAE or TC, $\widehat{\mathcal{M}^{int}}(A) = 0 \Longleftrightarrow A$ is diagonal (*Positive definiteness/Maximal agreement*), as will be explained in Proposition 5. On the other hand, it also holds that $\widehat{\mathcal{M}^{int}}(A) = 1 \Longleftrightarrow A \in \mathcal{S}_{max}$ (*Minimal agreement*). Recall that $\mathcal{S}_{max}$ depends on the metric, although we did not specify this dependency in the notation to lighten it.

Matrix $\mathcal{W}^{int} = (w^{int}_{ij})_{i,j=1,\dots,r}$ in Definition 1 is defined by

$$w^{int}_{ij} = \begin{cases} d(I_i, I_j)\,/\,N & \text{for } \mathcal{M} = \mathrm{MAE} \\ \gamma^{int}_{ij}\, d(I_i, I_j) & \text{for } \mathcal{M} = \mathrm{TC} \end{cases}$$

where γ^{int}_{ij} is introduced in (7). We wonder when is this cost matrix symmetric. The answer is: always for $\mathcal{M}$ =MAE, since the distance between intervals is. In the case of $\mathcal{M} = \mathrm{TC}$, $\mathcal{W}^{int}$ is symmetric if for all $i \neq j$, $\gamma^{int}_{ij} = \gamma^{int}_{ji}$, which is equivalent to say that

$$\sum_{k\neq i,j} \frac{\delta_k}{\delta_i} = \sum_{k\neq i,j} \frac{\delta_k}{\delta_j} \Leftrightarrow \delta_i = \delta_j,$$

that is, $\mathcal{W}^{int}$ is symmetric in case of homogeneous density of the intervals. In this case, if we denote by δ the common value of δ_i, $i = 1,\dots,r$, then, $\delta = n_i/\ell_i$ for all i or, equivalently, $n_i = \delta\,\ell_i$. That is, matrix $\mathcal{W}^{int}$ is symmetric if and only if the number of instances belonging to any interval is proportional to the length of the interval. We name this case *the proportional case* (not to confuse with the *balanced case*, in which the number of cases in all intervals coincide, that is, $n_i = n_j$ for all $i,j = 1,\dots,r$). Below (Corollary 5), we will establish that in the *proportional case*, metrics $\widehat{\mathrm{TC}^{int}}$ and $\widehat{\mathrm{MAE}^{int}}$ coincide.

Before proving this, and analogously to the ordinal scenario, we first derive expressions for MAE^{int}_{max} and TC^{int}_{max}, which will be useful for further analysis of the metrics $\widehat{\mathrm{MAE}^{int}}$ and $\widehat{\mathrm{TC}^{int}}$.

Proposition 3 MAE^{int}_{max} *and* TC^{int}_{max} *can be expressed as the sum of one term for any of the classes as follows*:

(a) $\mathrm{MAE}^{int}_{max} = \sum_{j=1}^{r} \widehat{G}_j$, *with* $\widehat{G}_j = \frac{n_j}{N} d(I_{\widehat{g}_j}, I_j)$ *and* $\widehat{g}_j = \arg\max_{\ell=1,\dots,r} d(I_\ell, I_j)$.

(b) $\mathrm{TC}^{int}_{max} = \sum_{j=1}^{r} \widehat{K}_j$, *with* $\widehat{K}_j = n_j \frac{\sum_{\ell \neq j} \delta_\ell}{\delta_{\widehat{k}_j}} d(I_{\widehat{k}_j}, I_j)$ *and* $\widehat{k}_j = \arg\max_{\ell=1,\dots,r} \frac{d(I_\ell, I_j)}{\delta_\ell}$.

The proof of this result follows similarly to that of Proposition 2, and is therefore omitted.

The following corollaries study the behavior of these maximum values in both the binary and proportional cases.

Corollary 4 *In the binary case* $r = 2$,

$$\mathrm{MAE}^{int}_{max} = \max(\ell_1, \ell_2) \quad \text{and} \quad \mathrm{TC}^{int}_{max} = N \times \max(\ell_1, \ell_2).$$

Corollary 5 *In the proportional case*, $\widehat{\mathrm{TC}^{int}} = \widehat{\mathrm{MAE}^{int}}$.

Refer to Appendix A.3 for the proofs of these results.

To further explore the normalized interval-scale metrics derived from MAE and TC, we first consider the desirable properties that a metric should satisfy, as outlined in Amigó et al. (2020), and then investigate their fundamental norm properties. The proofs of these results can be found in Appendix A.3.

Proposition 4 *The metric* $\widehat{\mathcal{M}^{int}}$ *verifies the following two properties for both* $\mathcal{M} = \mathrm{MAE}$ *and* TC:

1. **Scale invariance**: $\widehat{\mathcal{M}^{int}}$ *remains unchanged when the scale of the original data is modified.*
2. **Strict Monotonicity**: *If predictions are moved farther away from the true category* (*while maintaining the same density*), $\widehat{\mathcal{M}^{int}}$ *will increase.*

Furthermore, the metric $\widehat{\mathrm{TC}^{int}}$ *also satisfies the following additional property*:

3. **Imbalance**: $\widehat{\mathrm{TC}^{int}}$ *is more sensitive to changes in items that are moved farther from* (*or to*) *a low-density class than a high-density class.*

The fundamental norm properties satisfied by our interval-scale metrics are presented in the following result:

Proposition 5 $\widehat{\mathcal{M}^{int}}$ *verifies the following properties of a* **norm** *on* $\mathcal{A}$, *both for* $\mathcal{M} = \mathrm{MAE}$ *and* TC.

1. *Non-negativity/Positiveness*:

$$\widehat{\mathcal{M}^{int}}(A) \geq 0 \text{ for all } A \in \mathcal{A}.$$

2. *Positive definiteness/Maximal agreement*:

$$\text{For } A \in \mathcal{A},\ \widehat{\mathcal{M}^{int}}(A) = 0 \Longleftrightarrow A \text{ is diagonal.}$$

3. *Homogeneity* (*of degree* 0):

$$\text{For } A \in \mathcal{A} \text{ and } k > 0,\ \widehat{\mathcal{M}^{int}}(kA) = \widehat{\mathcal{M}^{int}}(A).$$

4. (*Restricted*) *Subadditivity/Triangle inequality*:

$$\text{For } A, B \in \mathcal{A}, \text{ if for any } j = 1, \ldots, r,\ \sum_{i=1}^{r} a_{ij} = \sum_{i=1}^{r} b_{ij}, \text{ then}$$

$$\widehat{\mathcal{M}^{int}}(A+B) \leq \widehat{\mathcal{M}^{int}}(A) + \widehat{\mathcal{M}^{int}}(B). \quad \text{In fact,}$$

$$\widehat{\mathcal{M}^{int}}(A+B) = \frac{1}{2}\left(\widehat{\mathcal{M}^{int}}(A) + \widehat{\mathcal{M}^{int}}(B)\right).$$

4 What length to assign to the rightmost interval?

As was said in the introduction, a problem occurs when binning a count variable or discretizing a continuous variable since frequently the rightmost interval is unconstrained, which means it cannot be assigned a preset length. To use the measures described in Sect. 3, however, this interval must have a length defined in advance. How can this issue be resolved?

An initial naïve approximation may be to take the maximum of the observations in the given database as the upper limit of the interval. Nothing, however, prevents this sample limit from being surpassed in the future. Furthermore, this approach has a concerning lack of definition, such that one might take the greatest value plus one unit, or plus two, and so on.

After ruling out the first approximation as unsatisfactory, we investigated an alternative strategy based on two pillars. The first is the simple but powerful notion of decoupling the calculation of the number of observations in the rightmost interval from its length. That is, regardless of the length we assign it to compute the metrics introduced in Sect. 3, the number of observations in the interval will be the same, equal to all those that exceed its set lower limit. The second pillar is the concept of determining the length of the rightmost interval, given the lengths of the other intervals and the number of instances in each interval are known, in such a way that $\mathcal{M}^{int}_{max}$ is minimized, with $\mathcal{M}$ being MAE or TC (or any other ordinal metric). This is backed by the rationale that minimizing $\mathcal{M}^{int}_{max}$ would maximize the impact of a confusion matrix improvement/worsening on $\widehat{\mathcal{M}^{int}}$, enhancing its capacity to discern differences among classifiers.

Additionally, this procedure can be interpreted as a cost-sensitive approximation, where the assigned length serves as a specific cost associated with the misclassification errors involving the rightmost interval. By treating it as such, we separate the decision about the interval's length from the actual distribution of observations, allowing for an optimization criterion that enhances the metric's sensitivity without being tied to whether observations may fall outside the interval.

Following the rightmost unbounded scenario, we can also explore problems involving an unbounded leftmost interval or where both extreme intervals are unbounded. For cases where the unbounded class is the leftmost, the approach mirrors that of the rightmost interval: we simply apply the method to the first interval instead of the last. When both the first and last intervals are unbounded, however, the process becomes slightly more complex. One method is to treat these two classes as a single, merged one, applying the aforementioned procedure to determine a combined length. This total length is then divided between the two intervals. The split can be either equal or weighted proportionally to the number of instances in each class, allowing for flexibility in distribution. However, the choice of how to allocate the length ultimately depends on the specific context and dataset.

Let us provide an example to illustrate this procedure in the unbounded rightmost interval scenario.

4.1 A toy example

Classifiers A and B give the following confusion matrices for the same validation data set with $N = 15$ instances. Data is classified into $r = 3$ intervals, I_1, I_2 and I_3, with corresponding undetermined lengths $\ell_1 = \ell_2 = 1$ and $\ell_3 = x$. The validation set is balanced, with $n_1 = n_2 = n_3 = 5$ instances in each interval.

$$C_A = \begin{pmatrix} & \text{observed} & \\ I_1 & I_2 & I_3 \\ 3 & 2 & \mathbf{1} \\ 2 & 2 & \mathbf{2} \\ 0 & 1 & 2 \end{pmatrix} \quad C_B = \begin{pmatrix} & \text{observed} & \\ I_1 & I_2 & I_3 \\ 3 & 2 & \mathbf{2} \\ 2 & 2 & \mathbf{1} \\ 0 & 1 & 2 \end{pmatrix}$$

Note that since $\delta_1 = \delta_2$ and the discrepancy between C_A and C_B consists of a difference of an element of I_3 misclassified into I_1 (classifier B) or into I_2 (classifier A), intuition tells us that classifier A is better than B and, according to this, we should obtain that $\widehat{\mathrm{TC}^{int}}(C_B) > \widehat{\mathrm{TC}^{int}}(C_A)$. We will see that this is indeed so. But also, we will find out the value for the undetermined x that makes this difference more evident.

For it, first, we calculate TC^{int} by using that

$$\mathrm{TC}^{int}(C) = \sum_{i,j=1}^{3} C_{ij}\, w_{ij}^{int} = sum(C \odot \mathcal{W}^{int})$$

where $w_{ij}^{int} = \gamma_{ij}^{int}\, d(I_i, I_j)$, with $d(I_1, I_2) = 1$, $d(I_1, I_3) = 1 + \max(1, x)$, $d(I_2, I_3) = \max(1, x)$, and $\delta_1 = n_1 = 5$, $\delta_2 = n_2 = 5$, $\delta_3 = n_3/x = 5/x$. Using that $\gamma_{ij}^{int} = \frac{\sum_{\ell \neq j}^{3} \delta_\ell}{\delta_i}$ we have that matrix $\mathcal{W}^{int}$ can be written as

$$\mathcal{W}^{int} = \begin{pmatrix} 0 & 1 + \frac{1}{x} & 2\,(1 + \max(1,x)) \\ 1 + \frac{1}{x} & 0 & 2\,\max(1,x) \\ (1+x)\,(1 + \max(1,x)) & (1+x)\,\max(1,x) & 0 \end{pmatrix},$$

and therefore,

$$\mathrm{TC}^{int}(C_A) = sum(C_A \odot \mathcal{W}^{int}) = 6 + \frac{4}{x} + (7+x)\,\max(1,x),$$
$$\mathrm{TC}^{int}(C_B) = sum(C_B \odot \mathcal{W}^{int}) = \ \mathrm{TC}^{int}(C_A) + 2\,,$$

obtaining that $\mathrm{TC}^{int}(C_B) - \ \mathrm{TC}^{int}(C_A) = 2$, which is independent of x.

Since $\widehat{\mathrm{TC}^{int}}(C) = \frac{\mathrm{TC}^{int}(C)}{\mathrm{TC}^{int}_{max}}$, with TC^{int}_{max} independent of the values of matrix C, given the column sums are fixed, it is obvious that the value of x that minimizes TC^{int}_{max} at the same time maximizes the relative difference between $\mathrm{TC}^{int}(C_A)$ and $\mathrm{TC}^{int}(C_B)$.

By Proposition 6 below, we get in this example that the value of x that minimizes TC^{int}_{max} is $x_{min} = 1/\sqrt{2}$ and that $\mathrm{TC}^{int}_{max}(x_{min}) = 5\,(2\,\sqrt{2}+7)$. Then, we assign a length of $x_{min} = 1/\sqrt{2}$ to interval I_3 for the purposes of computing the metric, and

$$\widehat{\mathrm{TC}^{int}}(C_A) = \frac{\mathrm{TC}^{int}(C_A)}{\mathrm{TC}^{int}_{max}} = \frac{6 + \frac{4}{x_{min}} + (7 + x_{min})\,\max(1, x_{min})}{5\,(2\,\sqrt{2}+7)} = \frac{73 + \frac{11}{\sqrt{2}}}{205}\,,$$

$$\widehat{\mathrm{TC}^{int}}(C_B) = \frac{\mathrm{TC}^{int}(C_B)}{\mathrm{TC}^{int}_{max}} = \frac{\mathrm{TC}^{int}(C_A) + 2}{\mathrm{TC}^{int}_{max}} = \frac{87 + \frac{3}{\sqrt{2}}}{205}\,,$$

from what we finally get that $\widehat{\mathrm{TC}^{int}}(C_B) - \widehat{\mathrm{TC}^{int}}(C_A) = \frac{14 - \frac{8}{\sqrt{2}}}{205} \approx 0.0407$. With any other value of x, the difference between the two would be less.

4.2 Optimal interval length: useful bounds and a particular case

Denote by $\mathcal{M}^{int}_{max}(x)$, where $\mathcal{M}$ could be TC or MAE, the function that returns the value of $\mathcal{M}^{int}_{max}$ for a length x assigned to the rightmost interval: $x = \ell_r$. The following results help to find the value of x that minimizes this function providing an upper bound for it.

First, we consider the case $\mathcal{M} = \mathrm{TC}$.

Theorem 1 *Let* $x_{min} := \arg\min_{x>0} \mathrm{TC}^{int}_{max}(x)$. *Then,* $x_{min} \leq n_r \sum_{j=1}^{r-1} \ell_j$.

In the balanced case ($n_1 = \cdots = n_r$), *the upper bound for* x_{min} *drops to* $\sum_{j=1}^{r-1} \ell_j$.

As a consequence, the search for a minimum of $\mathrm{TC}^{int}_{max}(x)$ can be restricted to the interval $(0, n_r \sum_{j=1}^{r-1} \ell_j]$ (respectively, to $(0, \sum_{j=1}^{r-1} \ell_j]$ in the balanced case). Observe that x_{min} could be not unique. This result is a direct consequence of the following lemma.

Lemma 1 *If* $x > n_r \sum_{j=1}^{r-1} \ell_j$, *then* $\mathrm{TC}^{int}_{max}(x)$ *is a monotonically increasing function.*

In the balanced case, the interval in which $\mathrm{TC}^{int}_{max}(x)$ *is monotonically increasing expands to* $x > \sum_{j=1}^{r-1} \ell_j$.

Proof of Lemma 1 Assume that $x > n_r \sum_{j=1}^{r-1} \ell_j$. From Proposition 3 we have that $\mathrm{TC}_{max}^{int} = \sum_{j=1}^{r} \widehat{K}_j$, with $\widehat{K}_j = n_j \frac{\sum_{\ell \neq j} \delta_\ell}{\delta_{\widehat{k}_j}} d(I_{\widehat{k}_j}, I_j)$ and $\widehat{k}_j = \arg\max_{\ell=1,\dots,r} \frac{d(I_\ell, I_j)}{\delta_\ell}$. We define the following quantities that will be useful in the proof:

$$\alpha_j := \frac{1}{n_r} \sum_{\ell \neq j,r} \delta_\ell \qquad \text{for } j = 1, \dots, r-1$$

and

$$\beta_j := \begin{cases} \sum_{k=j+1}^{r-1} \ell_k & \text{if } j = 1, \dots, r-2 \\ 0 & \text{if } j = r-1. \end{cases}$$

Observe that they are all non-negative and that, for $j = 1, \dots, r-1$, $d(I_j, I_r) = \beta_j + x$.

Step 1: We first prove that for $j = 1, \dots, r-1$, $\widehat{k}_j = r$. Indeed, for all $k = 1, \dots, r-1$,

$$\frac{1}{\delta_r} = \frac{x}{n_r} > \sum_{j=1}^{r-1} \ell_j > \ell_k \geq \frac{\ell_k}{n_k} = \frac{1}{\delta_k}.$$

In addition, we also have the following:

- For $k < j$,

$$d(I_k, I_j) = \sum_{i=k+1}^{j-1} \ell_i + \max(\ell_k, \ell_j) < \sum_{i=k}^{j} \ell_i < x$$

  and

$$d(I_r, I_j) = \sum_{i=j+1}^{r-1} \ell_i + \max(\ell_r, \ell_j) = \sum_{i=j+1}^{r-1} \ell_i + x \geq x,$$

  with the convention that the sums are zero if the lower limit for the index is greater than the upper limit, as usual. Then, $d(I_r, I_j) > d(I_k, I_j)$. As a consequence, if $k < j$,

$$\frac{d(I_r, I_j)}{\delta_r} > \frac{d(I_k, I_j)}{\delta_k}.$$

- Similar arguments lead to the same result for $k > j$.

Therefore, $\widehat{k}_j = r$ for $j = 1, \dots, r-1$.

On the other hand, $\widehat{k}_r$ depends on $\ell_1, \cdots, \ell_{r-1}$ and $n_1, \cdots, n_{r-1}$, and hence its value cannot be determined in general.

Step 2: Let us now determine the values of $\widehat{K}_j$. For $j = 1, \dots, r-1$,

$$\begin{aligned}\widehat{K}_j =& n_j \frac{\sum_{\ell \neq j} \delta_\ell}{\delta_r} d(I_r, I_j) = n_j \frac{x}{n_r}\left(\sum_{\ell \neq j,r} \delta_\ell + \frac{n_r}{x}\right)(\beta_j + x)\\ =& n_j\left[\alpha_j x^2 + (\alpha_j\beta_j + 1)x + \beta_j\right],\\ \widehat{K}_r =& n_r\left(\frac{1}{\delta_{\widehat{k}_r}} \sum_{\ell \neq \widehat{k}_r, r} \delta_\ell + 1\right)\left(\beta_{\widehat{k}_r} + x\right) = n_r\left[\left(\frac{n_r}{\delta_{\widehat{k}_r}}\alpha_{\widehat{k}_r} + 1\right)x + \left(\frac{n_r}{\delta_{\widehat{k}_r}}\alpha_{\widehat{k}_r} + 1\right)\beta_{\widehat{k}_r}\right].\end{aligned}$$

Step 3: Finally, we have the following expression for $\mathrm{TC}^{int}_{max}(x)$:

$$\begin{aligned}\mathrm{TC}^{int}_{max}(x) =& \left(\sum_{j=1}^{r-1} n_j\alpha_j\right)x^2 + \left(\sum_{j=1}^{r-1} n_j(\alpha_j\beta_j + 1) + n_r\left(\frac{n_r}{\delta_{\widehat{k}_r}}\alpha_{\widehat{k}_r} + 1\right)\right)x\\ &+ \left(\sum_{j=1}^{r-1} n_j\beta_j + n_r\left(\frac{n_r}{\delta_{\widehat{k}_r}}\alpha_{\widehat{k}_r} + 1\right)\beta_{\widehat{k}_r}\right).\end{aligned}$$

It is a polynomial of second order with positive coefficients, then it is monotonically increasing for $x > n_r \sum_{j=1}^{r-1} \ell_j$.

The proof in the balanced case is similar and therefore omitted. □

In Sect. 4.2, we present a series of simulations that incorporate the upper bound established in Theorem 1. These simulations illustrate the practical application of the bound in various scenarios. To further highlight its usefulness, we provide some examples in Table 1, which demonstrate how this upper bound can be employed effectively.

Table 1 Value of x_{min} (up to three decimals) and its upper bound provided by Theorem 1 for $r = 3$ and different $\vec{n}$, with $\vec{\ell} = (1, 0.4)$

	n_1	n_2	n_3	x_{min}	Bound for x_{min}
Balanced case	4	4	4	0.535	5.6
n_1 varies	1	7	4	1.000	5.6
	2	7	4	1.000	5.6
	3	7	4	0.713	5.6
	4	7	4	0.535	5.6
	5	7	4	0.428	5.6
	6	7	4	0.356	5.6
n_2 varies	7	1	4	0.770	5.6
	7	2	4	0.571	5.6
	7	3	4	0.381	5.6
	7	4	4	0.305	5.6
	7	5	4	0.305	5.6
	7	6	4	0.305	5.6
n_3 varies	4	7	1	0.134	1.4
	4	7	2	0.267	2.8
	4	7	3	0.401	4.2
	4	7	4	0.535	5.6
	4	7	5	0.668	7.0
	4	7	6	0.802	8.4

It is important to emphasize that the existence of such an upper bound for x_{min} narrows its search field. This simplifies its computation and makes the overall process more efficient.

For the case $\mathcal{M} = \mathrm{MAE}$, we have the following result proportioning an upper bound for the value of x that minimizes function $\mathrm{MAE}^{int}_{max}(x)$.

Theorem 2 *Let* $x_{min} := \arg\min_{x>0} \mathrm{MAE}^{int}_{max}(x)$. *Then,* $x_{min} \leq \sum_{j=1}^{r-1} \ell_j$.

Proof The proof is obtained as for Theorem 1 proving that, if $x > \sum_{j=1}^{r-1} \ell_j$, then $\mathrm{MAE}^{int}_{max}(x)$ is a monotonically increasing function. This is an equivalent version of Lemma 1 adapted for the metric MAE, so the proof follows the same ideas.

Assume then that $x > \sum_{j=1}^{r-1} \ell_j$. Recall that by Proposition 3, $\mathrm{MAE}^{int}_{max} = \sum_{j=1}^{r} \widehat{G}_j$ with $\widehat{G}_j = \frac{n_j}{N} d(I_{\widehat{g}_j}, I_j)$, and $\widehat{g}_j = \arg\max_{\ell=1,\dots,r} d(I_\ell, I_j)$. It is easy to derive that

$$\widehat{g}_j := \begin{cases} r & \text{if } j = 1, \dots, r-1 \\ 1 & \text{if } j = r. \end{cases}$$

Therefore,

$$\widehat{G}_j = \frac{n_j}{N}(\beta_j + x) \quad \text{for } j = 1, \dots, r-1 \quad \text{and} \quad \widehat{G}_r = \frac{n_r}{N}(\beta_1 + x).$$

Then,

$$\mathrm{MAE}^{int}_{max}(x) = x + \frac{1}{N}\left(\sum_{j=1}^{r-1} n_j\beta_j + n_r\beta_1\right),$$

which is a straight line with positive slope, so it is a monotonically increasing function for $x > \sum_{j=1}^{r-1} \ell_j$. □

4.3 The balanced case with $r = 3$

The following result, whose proof is in Appendix B, states the length of the rightmost interval that minimizes TC^{int}_{max} in a particular case that, precisely, we have used in the toy example. This particular case fully covers the situation of having $r = 3$ intervals when the number of cases in each interval is the same. Although it is a particular scenario, the *balanced case* is of great interest in practice since one of the most used methods for the discretization of continuous variables, if not the most, consists of dividing into intervals in such a way the number of cases in each interval be the same, used by default in the functions that implement the discretization algorithms in R programming language.[1] It is the case of `method="frequency"` in function `arules::discretize`, and of `method="quantile"` in function `bnlearn::discretize`.

Proposition 6 *If* $r = 3$, $\ell_1 = 1$ *and* $\ell_2 = L > 0$ *known, and if* $n_1 = n_2 = n_3$, *then the value of* $x = \ell_3$ *that minimizes* TC^{int}_{max} *is given by:*

[1] R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

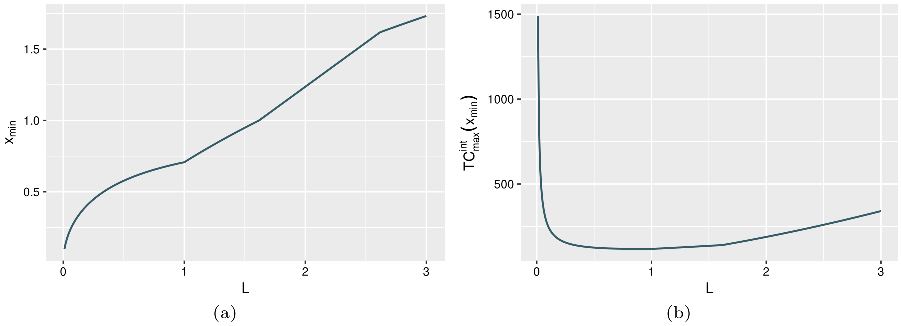


Fig. 2 Plot of x_{min} (a) and $\mathrm{TC}^{int}_{max}(x_{min})$ (b) as function of the parameter L, for $L \in (0,\ 3]$ and $n = 12$

$$x_{min} = \begin{cases} \sqrt{\frac{L}{L+1}} & \text{if } L \le 1,\ \mathrm{TC}^{int}_{max}(x_{min}) = n\left(2\sqrt{\frac{L+1}{L}} + 2L + 4 + \frac{1}{L}\right) \\ \frac{L}{\sqrt{L+1}} & \text{if } 1 < L \le \frac{1+\sqrt{5}}{2},\ \mathrm{TC}^{int}_{max}(x_{min}) = n\left(2\sqrt{L+1} + 3L + 3 + \frac{1}{L}\right) \\ \frac{L(\sqrt{5}-1)}{2} & \text{if } \frac{1+\sqrt{5}}{2} < L \le \frac{3+\sqrt{5}}{2},\ \mathrm{TC}^{int}_{max}(x_{min}) = n\frac{L}{2}\left((\sqrt{5}+1)L + \sqrt{5} + 7\right) \\ \sqrt{L} & \text{if } L > \frac{3+\sqrt{5}}{2},\ \mathrm{TC}^{int}_{max}(x_{min}) = n\,L\left(2\sqrt{L} + L + 3\right). \end{cases}$$

where n is the common number of instances at any interval ($n = n_i,\ i = 1,2,3$).

Note that considering ℓ_1 as a reference, L represents the ratio ℓ_2/ℓ_1. Proposition 6 provides the optimal value for $x = \ell_3/\ell_1$ as a function of L, with Fig. 2 providing a visual depiction of this outcome. Figure 2a displays the progression of x_{min} concerning L, while Fig. 2b presents the plot of function $\mathrm{TC}^{int}_{max}(x_{min})$ with respect to the parameter L.

Remark 5 Another significant consideration concerns the upper bound of x_{min}. Theorem 1 states in the balanced case that the abscissa of the minimum is never greater than the sum of the lengths of all intervals except the last one. In the balanced scenario with $r = 3$ described in Proposition 6, this sum corresponds to the value $L + 1$. However, in this particular case it is easy to check from the expression for x_{min} obtained in Proposition 6 that x_{min} is always smaller than the length of the largest interval, that is, $x_{min} < \max(1, L)$. Compared to the upper bound given in Theorem 1, this provides a tighter constraint.

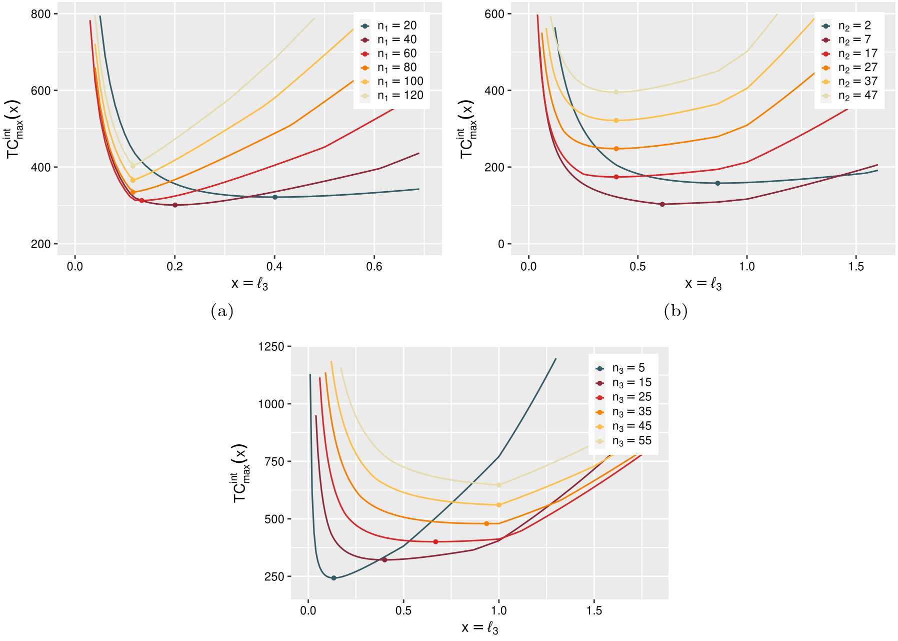
4.4 Simulations

The issue we address in this section is as follows: in order to utilize the performance measures outlined in Sect. 3, the rightmost interval must have a predefined length. When this is not the case, we propose to determine the value for this length that optimally minimizes the maximum value of the metric. Recall that we have denoted this value as x_{min}. The analytic results from Sect. 4.1 provide an upper bound for x_{min} for both metrics TC^{int}_{max} (Theorem 1) and MAE^{int}_{max} (Theorem 2). Furthermore, Proposition 6 gives us, for $r = 3$ intervals in the balanced case, the value of x_{min} as a function of the lengths of the other two intervals for the first of these metrics.

The utility of the upper bounds is that they provide us with a range to focus on for obtaining a numerical approximation when exact computation of x_{min} is not feasible. This is what we will do in this section. Specifically, in various scenarios for the case of $r = 3$ intervals (unbalanced) and for $r = 4$ intervals, we will carry out a procedure to obtain a numerical approximation of the value of x_{min}, based on the lengths of the remaining intervals $\vec{\ell} = (\ell_1, \dots, \ell_{r-1})$, and the number of instances in each of them, $\vec{n} = (n_1, \dots, n_r)$. We will focus on TC^{int}_{max}, as the procedure for MAE^{int}_{max} is analogous.

The objective is not solely to obtain results in specific scenarios, as will demonstrate, but rather to showcase the utility of the algorithm we have implemented. Given an initial interval for the value of x_{min} acquired through the upper bound provided in Theorem 1, we utilize the algorithm to obtain a numerical approximation.

The algorithm is implemented within the `R` programming environment. We employ the definitions of ordinal and interval-scale metrics introduced in Sect. 2 and Sect. 3.2.2, respectively. The approximate minimum is obtained by using the `optimize` function from the `R` Stats Package, which contains a variety of helpful statistical functions. The numerical method implemented into `optimize` is specifically designed for continuous functions and combines golden section search with successive parabolic interpolation. For further details, please refer to the R Documentation.



(a)

(b)

(c)

Fig. 3 Plot of TC^{int}_{max} for $r = 3$ as function of $x = \ell_3$, with $\vec{\ell} = (1, 0.4)$. When not specified, the number of instances of each class is fixed at the following values: $n_1 = 20$, $n_2 = 37$ and $n_3 = 15$. (a) Different values of n_1. (b) Different values of n_2. (c) Different values of n_3

Table 2 Minimum of TC^{int}_{max} for $r = 3$ and different $\vec{n}$, with $\vec{\ell} = (1, 0.4)$

a) $n_2 = 37,\ n_3 = 15$. Bound for x_{min}: 21.				b) $n_1 = 20,\ n_3 = 15$. Bound for x_{min}: 21.			
n_1	x_{min}	$\mathrm{TC}^{int}_{max}(x_{min})$	Increment	n_2	x_{min}	$\mathrm{TC}^{int}_{max}(x_{min})$	Increment
20	0.40089	321.5663		2	0.86603	157.9660	
40	0.20045	301.0038	−20.5625	7	0.61224	102.9500	−55.0160
60	0.13363	312.8163	11.8125	17	0.40089	174.2332	71.2832
80	0.11583	334.1750	21.3587	27	0.40089	247.8997	73.6665
100	0.11583	365.3400	31.1650	37	0.40089	321.5663	73.6666
120	0.11583	402.1167	36.7767	47	0.40089	395.2329	73.6666

c) $n_1 = 20,\ n_2 = 37$.				
n_3	x_{min}	Bound for x_{min}	$\mathrm{TC}^{int}_{max}(x_{min})$	Increment
5	0.13363	7	242.8163	
15	0.40089	21	321.5663	78.7500
25	0.66815	35	400.3163	78.7500
35	0.93541	49	479.0663	78.7500
45	1.00000	63	560.1806	81.1143
55	1.00000	77	646.9659	86.7853

The upper bound for x_{min} provided by Theorem 1 is also recorded

Table 3 Behavior of x_{min}, $\mathrm{TC}_{max}^{int}(x_{min})$, and the slope of function TC_{max}^{int} after its minimum, for the examples in Table 2

Case	x_{min}	$\mathrm{TC}_{max}^{int}(x_{min})$	Slope after x_{min}
a) $n_1 \nearrow$	$\searrow$	$\searrow \nearrow$	$\nearrow$
b) $n_2 \nearrow$	$\searrow$	$\searrow \nearrow$	$\nearrow$
c) $n_3 \nearrow$	$\nearrow$	$\nearrow$	$\searrow$

When not specified, the parameters are fixed at the following values: $n_1 = 20, n_2 = 37, n_3 = 15, \ell_1 = 1$, and $\ell_2 = 0.4$

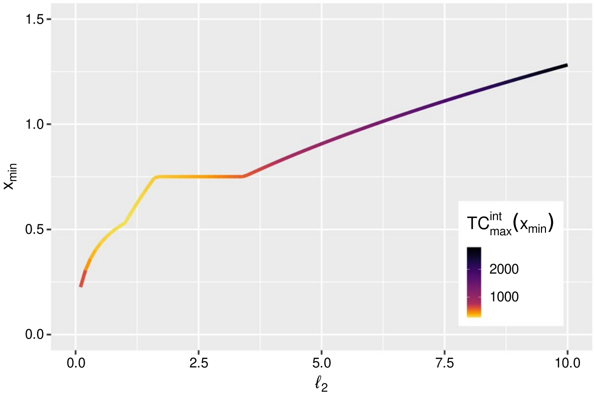


Fig. 4 Plot of x_{min} as function of ℓ_2 with $r = 3, \vec{n} = (20, 37, 15)$ fixed and $\ell_1 = 1$. The color gradient is given by the magnitude of $\mathrm{TC}_{max}^{int}(x_{min})$

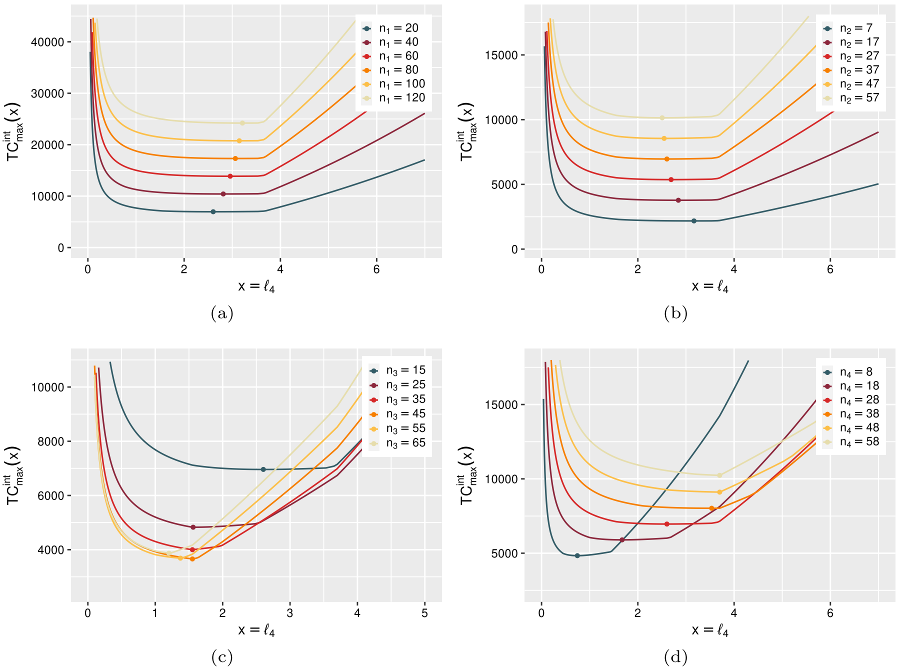
Given the vast range of values that the parameters $\vec{\ell}$ and $\vec{n}$ can encompass, fixed r, and thus the multitude of potential scenarios, we give up being exhaustive and instead provide a limited number of situations as examples in this section.

• **Case $r = 3$.** This scenario is represented in Fig. 3, where we plot TC_{max}^{int} as function of ℓ_3 when the length of the rest of intervals is fixed to $\vec{\ell} = (1, 0.4)$.

The approximated minimum of the function, which is x_{min}, is indicated with a dot. Observe that we set $\ell_1 = 1$, as in Proposition 6. Here, we take $\ell_2 < 1$, specifically, $\ell_2 = 0.4$. Similarly, we could have chosen $\ell_2 > 1$, yielding similar results. Note that in all the plots in Fig. 3, the simulation with $\vec{n} = (20, 37, 15)$ and $\vec{\ell} = (1, 0.4)$ has been considered.

Sub-tables a), b) and c) in Table 2 provide information about the minimum of TC_{max}^{int} and correspond, respectively, to sub-figures (a), (b) and (c) in Fig. 3. The last column indicates the difference between the values of $\mathrm{TC}_{max}^{int}(x_{min})$ of two consecutive cases. The corresponding upper bounds for x_{min} given by Theorem 1 are also indicated. As they depend on n_3, in sub-table c) an extra column has been added.

In general, the first thing we notice is that the function TC_{max}^{int} is convex, and that it reaches its minimum in an abscissa x_{min} much lower than the upper bound of Theorem 1. Furthermore, x_{min} is smaller than the length of the largest interval, as in the balanced case (see Remark 5). While here we can only assert this observation within the context of these examples and cannot extrapolate it as a generalized conclusion, it suggests a tighter upper bound compared to that proposed in Theorem 1. We observe in Table 2 a) and b) that x_{min} decreases until stabilizing at a certain value, while $\mathrm{TC}_{max}^{int}(x_{min})$ increases after a slight decrease. We do not detect any regularity in this growth. After the minimum is reached, the



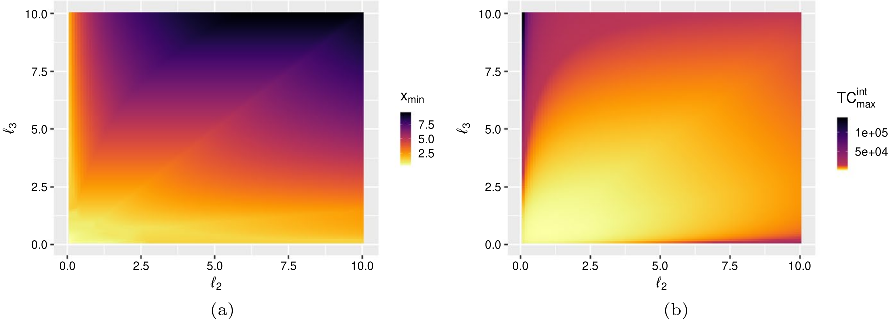
(a)

(b)

(c)

(d)

Fig. 5 Plot of TC^{int}_{max} for $r = 4$ as function of $x = \ell_4$, with $\vec{\ell} = (1, 0.4, 3.7)$. When not specified, the number of instances of each class is fixed at the following values: $n_1 = 20$, $n_2 = 37$, $n_3 = 15$ and $n_4 = 28$. (a) Different values of n_1. (b) Different values of n_2. (c) Different values of n_3. (d) Different values of n_4



(a)

(b)

Fig. 6 Heatmap of (a) x_{min} and (b) $TC^{int}_{max}(x_{min})$ as function of ℓ_2 and ℓ_3 for $r = 4$, $\vec{n} = (20, 37, 15, 28)$, $\ell_1 = 1$

slope of the curve increases as n_1 (respectively, n_2) arises. Although the behavior is similar in both cases, we highlight two facts for b). First, value 0.40089 is reached by x_{min} quite fast. Secondly, after $x_{min} = 0.40089$ is hit, TC^{int}_{max} seems to grow regularly.

As for Table 2 c), the behavior of TC^{int}_{max} as n_3 varies, differs significantly from the previous cases, which is not surprising since n_3 is the number of instances in the unbounded rightmost interval. In this scenario, as n_3 arises, both x_{min} and $TC^{int}_{max}(x_{min})$ increase, while the slope of the curve after the minimum is reached decreases (see Fig. 3 (c)). Upon reaching value 1, x_{min} remains there even for values of n_3 larger than those displayed (which are not shown here for space reasons), while $\mathrm{TC}^{int}_{max}(x_{min})$ first grows regularly, and then with increasing gap.

The behavior of x_{min}, $\mathrm{TC}^{int}_{max}(x_{min})$, and the slope of function TC^{int}_{max} after its minimum in these examples is summarized in Table 3. Here, arrows ↘ and ↗ denote increase and decrease, respectively.

To finish this small study on the sensitivity of x_{min} and function TC^{int}_{max} with respect to the parameters $\vec{\ell}$ and $\vec{n}$, we plot in Fig. 4x_{min} with respect to ℓ_2 in the particular case in which the number of instances of each class is given by $\vec{n} = (20, 37, 15)$ and $\ell_1 = 1$.

The color gradient in Fig. 4 indicates the values of $\mathrm{TC}^{int}_{max}(x_{min})$. The behavior of both x_{min} and $\mathrm{TC}^{int}_{max}(x_{min})$ is similar to that of the balanced case, illustrated in Fig. 2: x_{min} increases while $\mathrm{TC}^{int}_{max}(x_{min})$ first decreases and then increases.

• **Case $\boldsymbol{r} = 4$**. Figs. 5 and 6 illustrate the case with $r = 4$ classes.

Plots in Fig. 5 represent TC^{int}_{max} as function of ℓ_4 when the length of the intervals is fixed and the number of instances in each class varies. We set $\vec{\ell} = (1, 0.4, 3.7)$: ℓ_1 is taken as reference, the other intervals are smaller and larger than 1, respectively. On the other hand, the minimum of TC^{int}_{max} is represented as a function of ℓ_2 and ℓ_3, the number of instances of each class being fixed, in Fig. 6. In all the graphics, the simulation with $\vec{n} = (20, 37, 15, 28)$ and $\vec{\ell} = (1, 0.4, 3.7)$ has been taken into account.

From Fig. 5 we can draw the same general considerations as in case $r = 3$: the function TC^{int}_{max} is convex and its minimum is reached much earlier than the upper bound provided by Theorem 1. Once more, we note that the length of the largest interval serves as an upper limit for the minimum. Specifically, the first two cases are similar. When n_1 grows (Fig. 5a), x_{min} increases, while when n_2 grows (Fig. 5b), it decreases. Apart from that, in both cases, $\mathrm{TC}^{int}_{max}(x_{min})$ increases regularly. The slope of the curves after the minimum also increases. In general, the curves present the same shape. The case when n_3 grows (Fig. 5c) is the least regular. The value of x_{min} drops, while $\mathrm{TC}^{int}_{max}(x_{min})$ first decreases and then increases. No discernible patterns can be found in the behavior of the increments. As in the previous cases, the slope of the curves after the minimum increases. When n_4 grows (Fig. 5d), both x_{min} and $\mathrm{TC}^{int}_{max}(x_{min})$ increase, while the slope of the curves after $x = x_{min}$ is reached decreases.

Heatmaps in Fig. 6 show how the minimum of TC^{int}_{max} varies as a function of ℓ_2 and ℓ_3 in the particular case in which the number of instances of each class is fixed at $\vec{n} = (20, 37, 15, 28)$ and $\ell_1 = 1$. Precisely, Fig. 6a represents the magnitude of x_{min}, while Fig. 6b depicts the values of $\mathrm{TC}^{int}_{max}(x_{min})$. It is evident from both graphics that the minimum increases as the intervals become larger. This same behavior has been also detected in the $r = 3$ scenario.

5 Experimentation with real-world data

This section presents the experimental phase, designed to evaluate the utility of interval-scale metrics, such as $\widehat{\mathrm{MAE}}^{int}$, for improve classifiers through hyper-parameter tuning. Additionally, we investigate the metric's sensitivity to the arbitrary length assigned to the rightmost interval, which is required within an interval-scale framework. We use three real-world datasets for this purpose. For simplicity, we refer to $\widehat{\mathrm{MAE}}$ and $\widehat{\mathrm{MAE}}^{int}$ as SMAE and SMAE.int, respectively, throughout this section.

5.1 The datasets

- **Facial Age dataset** (Kaggle[2]). This dataset contains 9,673 facial images organized into 99 folders labeled by age. Each image, in PNG format, represents a human face, allowing for age-based feature analysis. We preprocess the images by resizing them to 32×32 pixels and converting them to grayscale, resulting in 1,024 features per image. The target variable, `age`, is grouped into six ordinal intervals: < 2, $[2, 10)$, $[10, 15)$, $[15, 35)$, $[35, 60)$, and ≥ 60. For the rightmost unbounded interval, we assign five different endpoint values: 80, 90, 100, 110, and 120, resulting in corresponding lengths of 20, 30, 40, 50, and 60, respectively.
- **Abalone dataset** (UC Irvine Machine Learning Repository[3]). This dataset consists of 4,177 instances and 8 features, used to predict the age of abalones based on physical measurements. The target variable, `Rings`, is converted into five intervals: < 8, [8, 10), [10, 11), [11, 14), and ≥ 14. We assign five different endpoints to the rightmost unbounded interval: 20, 25, 30, 35, and 40, corresponding to lengths of 6, 11, 16, 21, and 26.
- **Parkinson dataset** (UC Irvine Machine Learning Repository[4]). This dataset includes biomedical voice measurements from 42 individuals with early-stage Parkinson's disease, collected over a six-month period for tele-monitoring purposes. The dataset comprises 5,875 voice recordings, each described by 16 biomedical voice features, with the goal of predicting the motor Unified Parkinson's Disease Rating Scale (UPDRS), a widely used score to track disease progression. The target variable, `motor_UPDRS`, has been discretized into intervals: < 13, [13, 18), [18, 24), [24, 29), and ≥ 29. For the rightmost interval, we explore five endpoint values: 35, 40, 45, 50, and 60, resulting in respective lengths of 6, 11, 16, 21, and 31.

5.2 Description of the experiment

To examine the impact of using SMAE.int metric for hyper-parameter tuning, along with its sensitivity to different lengths assigned to the rightmost interval, we applied two tuning procedures, each on a different classifier:

(a) We employed the `train` function from the R package `caret`[5] to train random forest models, tuning the **mtry** hyper-parameter, which represents the number of features randomly sampled at each split. A modest ensemble of three trees was used. Hyper-parameter tuning was conducted using 3-fold cross-validation with a random search of 10 iterations. Although `caret` typically optimizes for Accuracy, we introduced SMAE and SMAE.int as custom metrics via the `summaryFunction` argument in `trainControl`, setting minimization as the objective.

[2] https://www.kaggle.com/datasets/frabbisw/facial-age

[3] https://archive.ics.uci.edu/dataset/1/abalone

[4] https://archive.ics.uci.edu/dataset/189/parkinsons+telemonitoring

[5] Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, Vol. 28(5), pp. 1–26. https://www.jstatsoft.org/index.php/jss/article/view/v028i05

Table 4 Facial age dataset

Perform. metric	Optimization metric Random Forest						
	Accuracy	SMAE	SMAE. int.20	SMAE. int.30	SMAE. int.40	SMAE. int.50	SMAE. int.60
SMAE. int.30				–			0.09162$^{\bullet}$
SMAE. int.40					–		0.08637$^{\bullet}$
SMAE. int.60				**0.08523**$^{\bullet}$	**0.08523**$^{\bullet}$	–	
Perform. metric	Error function *k*nn						
	Error rate	SMAE	SMAE. int.20	SMAE. int.30	SMAE. int.40	SMAE. int.50	SMAE. int.60
SMAE. int.20		**0.01861***	–			**0.05919**$^{\bullet}$	
SMAE. int.30		**0.02161***		–	0.02661*	**0.02425***	
SMAE. int.40		**0.01186***		**0.02919***	–	**0.00421****	**0.08008**$^{\bullet}$
SMAE. int.50	0.05254$^{\bullet}$		0.06126$^{\bullet}$	0.02237*	0.00818**	–	
SMAE. int.60		**0.07751**$^{\bullet}$					–

p values comparing SMAE.int metrics for random forest models optimized with different metrics and *k*nn models tuned with various error functions. *p* values in regular indicate a lower SMAE.int metric (i.e., better performance) for models tuned with the same metric compared to others; *p* values in bold represent the opposite hypothesis. Only significant *p* values are reported

(b) For *k*-nearest neighbors (*k*nn) classifiers, we used the `tune.knn` function from the R package `e1071`,[6] a tuning wrapper function that uses `e1071::tune`. Here, the hyper-parameter *k* determines the number of nearest (in Euclidean distance) vectors used for classification, where ties are broken randomly. If there are ties for the *k*th nearest vector, all candidates are included in the vote. Cross-validation with three folds was used to explore *k* values from 1 to 20. In addition to the default error function (which is the Error rate = 1− Accuracy), we employed SMAE and SMAE.int as custom error functions, specified with `tune.control`.

For model validation, we applied 10-fold cross-validation, dividing the dataset into 10 randomly partitioned folds, using each fold once as a test set, with the others forming the training set. Given the computational demands, we used a random subsample of 2,000 instances for the "Facial age" dataset for each training set. Hyper-parameters were tuned

[6] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F. (2024). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-16, https://CRAN.R-project.org/package=e1071

Table 5 Abalone dataset

Perform. metric	Optimization metric Random Forest						
	Accuracy	SMAE	SMAE. int.6	SMAE. int.11	SMAE. int.16	SMAE. int.21	SMAE. int.26
SMAE. int.26		0.09223$^{\bullet}$					–
Perform. metric	Error function *k*nn						
	Error rate	SMAE	SMAE. int.6	SMAE. int.11	SMAE. int.16	SMAE. int.21	SMAE. int.26
SMAE. int.6	0.04460*	0.05361$^{\bullet}$	–			0.00588**	
SMAE. int.11	0.00819*			–		0.02681*	0.05830$^{\bullet}$
SMAE. int.21			**0.05349**$^{\bullet}$	**0.04602***		–	
SMAE. int.26		**0.09295**$^{\bullet}$	**0.03494***	**0.01664***			–

p values comparing SMAE.int metrics for random forest models optimized with different metrics and *k*nn models tuned with various error functions. *p* values in regular indicate a lower SMAE.int metric (i.e., better performance) for models tuned with the same metric compared to others; *p* values in bold represent the opposite hypothesis. Only significant *p* values are reported

Table 6 Parkinson dataset

Perform. metric	Optimization metric Random Forest						
	Accuracy	SMAE	SMAE. int.6	SMAE. int.11	SMAE. int.16	SMAE. int.21	SMAE. int.31
SMAE. int.6	0.08308$^{\bullet}$		–	**0.09235**$^{\bullet}$			
SMAE. int.11	0.04060*		0.07945$^{\bullet}$	–			
Perform. metric	Error function *k*nn						
	Error rate	SMAE	SMAE. int.6	SMAE. int.11	SMAE. int.16	SMAE. int.21	SMAE. int.31
SMAE. int.6			–		**0.05972**$^{\bullet}$		
SMAE. int.16			0.03957*		–		

p values comparing SMAE.int metrics for random forest models optimized with different metrics and *k*nn models tuned with various error functions. *p* values in regular indicate a lower SMAE.int metric (i.e., better performance) for models tuned with the same metric compared to others; *p* values in bold represent the opposite hypothesis. Only significant *p* values are reported

with `caret::train` and `e1071::tune.knn` functions across five interval lengths for the SMAE.int metric, as outlined previously.

Predictions were generated from the best-performing models identified in each tuning procedure. For random forest, we used the generic `predict` function in R, while for *k*nn,

Table 7 Facial age dataset

Tuned random forest	20	30	40	50
30	$9.4 \times 10^{-8***}$	–	–	–
40	$1.5 \times 10^{-6***}$	0.00057^{***}	–	–
50	$3.7 \times 10^{-5***}$	0.00633^{**}	$0.05192^{\bullet}$	–
60	$3.2 \times 10^{-5***}$	0.00148^{**}	0.00276^{**}	$3.7 \times 10^{-5***}$
Tuned *k*nn	20	30	40	50
30	$4.2 \times 10^{-6***}$	–	–	–
40	$1.0 \times 10^{-6***}$	$1.2 \times 10^{-5***}$	–	–
50	$4.2 \times 10^{-7***}$	$3.4 \times 10^{-5***}$	0.00016^{***}	–
60	$1.4 \times 10^{-7***}$	$1.4 \times 10^{-5***}$	$3.4 \times 10^{-5***}$	0.00865^{**}

Adjusted *p* values comparing SMAE.int metric across tuned random forest and *k*nn models, for each of the five lengths assigned to the rightmost interval. The *p* values indicate that the mean SMAE.int value for the model in the row is significantly lower (and thus better) than that of the model in the column

Table 8 Abalone dataset

Tuned random forest	6	11	16	21
11	$8.0 \times 10^{-7***}$	–	–	–
16	$1.2 \times 10^{-9***}$	$1.8 \times 10^{-11***}$	–	–
21	$4.2 \times 10^{-10***}$	$1.8 \times 10^{-11***}$	$1.8 \times 10^{-11***}$	–
26	$4.1 \times 10^{-10***}$	$2.5 \times 10^{-11***}$	$5.2 \times 10^{-11***}$	$4.2 \times 10^{-10***}$
Tuned *k*nn	6	11	16	21
11	$2.6 \times 10^{-5***}$	–	–	–
16	$1.8 \times 10^{-7***}$	$3.1 \times 10^{-7***}$	–	–
21	$9.0 \times 10^{-9***}$	$2.8 \times 10^{-9***}$	$1.8 \times 10^{-7***}$	–
26	$8.4 \times 10^{-9***}$	$2.2 \times 10^{-8***}$	$8.4 \times 10^{-9***}$	$3.8 \times 10^{-5***}$

Adjusted *p* values comparing SMAE.int metric across tuned random forest and *k*nn models, for each of the five lengths assigned to the rightmost interval. The *p* values indicate that the mean SMAE.int value for the model in the row is significantly lower (and thus better) than that of the model in the column

we applied the `knn` function from the R package `class`.[7] The performance metric SMAE.int was calculated for each of the 10 confusion matrices obtained from cross-validation.

Finally, we performed statistical analyses to assess model performance. The Shapiro-Wilk test evaluated normality, while paired Student's t-tests and Wilcoxon signed-rank tests compared models. When applicable, *p* values were adjusted using the Holm-Bonferroni method to account for multiple comparisons. As usual, significance levels are denoted as follows: $^{\bullet}$ for 10%, for 5%, for 1% and for 1‰.

[7] Venables, W.N., Ripley, B.D. (2002). Modern Applied Statistics with S, Fourth edition. Springer, New York. ISBN 0-387-95457-0, https://www.stats.ox.ac.uk/pub/MASS4/

5.3 Results

To evaluate the effectiveness of the SMAE.int metric in the interval-scale framework hyper-parameter tuning, we compare the performance of models obtained by optimizing different metrics in random forest, and using the metrics as error functions in *k*nn, with SMAE.int serving as the benchmark for performance measurement.

For the “Facial age” dataset, Table 4 shows *p* values (in black) testing the hypothesis that the average SMAE.int metric for models optimized with the same metric is lower (thus better) than for models tuned with other metrics. Conversely, *p* values in red test the opposite hypothesis. Only significant *p* values are presented. Notably, models optimized with Accuracy/Error rate never outperform those optimized with SMAE.int. On the other hand, the SMAE.int metric using a rightmost interval length of 50 performs comparably well. For other lengths, SMAE-based tuning yields better outcomes. Tables 5 and 6 provide similar comparisons for the “Abalone” and “Parkinson” datasets, where SMAE.int yields better results with short-length intervals: 6 and 11 for “Abalone”, and 11 and 16 for “Parkinson”.

Overall, the results show variations depending on the dataset, classifier, and the assigned length to the rightmost interval. Although further experimentation would be beneficial, these preliminary results are promising and highlight the potential of SMAE.int as a practical metric for hyper-parameter tuning in the interval scale.

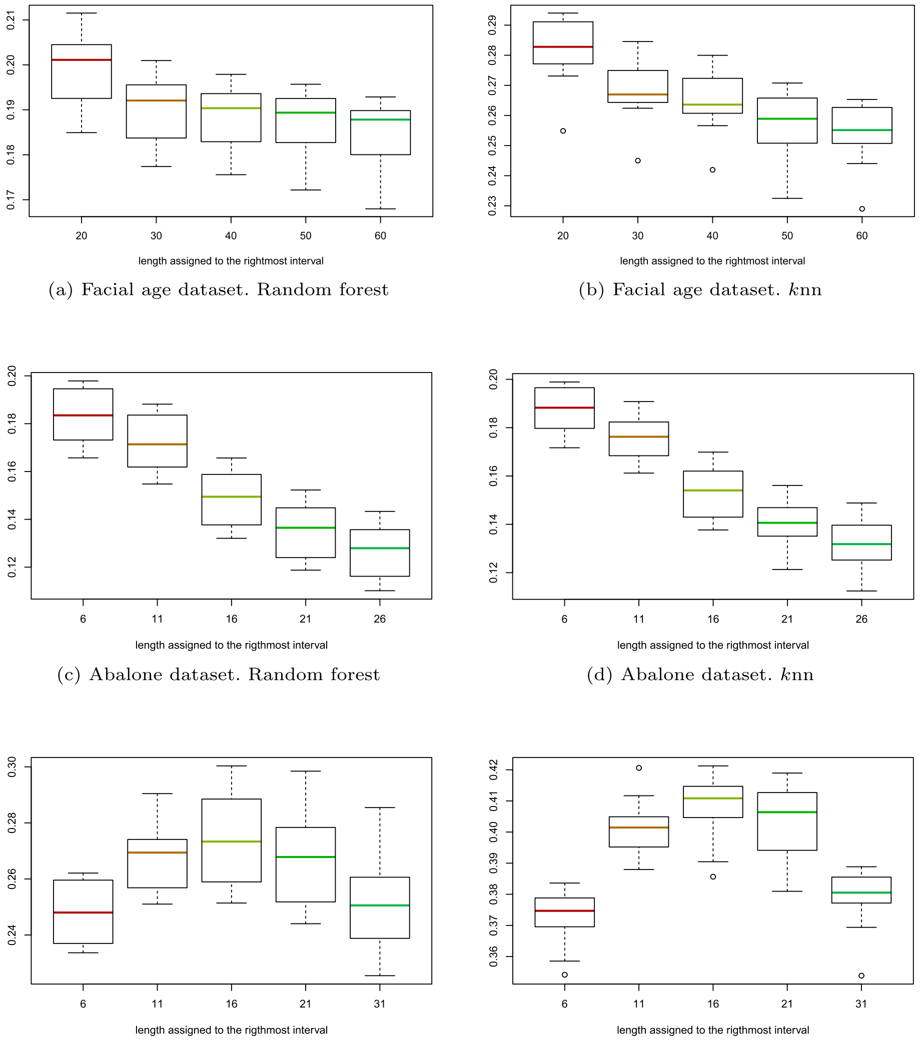
To further investigate the impact of the length assigned to the rightmost interval, we conducted multiple comparisons of the SMAE.int metric across models tuned with SMAE.int as the error function (for *k*nn) or as the optimization metric (for random forest). Tables 7, 8 and 9 report the adjusted *p* values from one-sided Student’s t-test or Wilcoxon signed-rank test, as appropriate, for comparing the mean SMAE.int across different assigned interval lengths for tuned random forest and *k*nn models. Figure 7 complements this analysis by illustrating boxplots of SMAE.int values across the interval lengths. The *p* values indicate that the mean SMAE.int value for the model in the row is significantly lower (and thus performs better) than that of the model in the column.

The boxplots in Fig. 7 reveal a trend where the SMAE.int metric tends to decrease (indicating improved performance) as the assigned length of the rightmost interval increases for the “Facial age” and “Abalone” datasets. In contrast, for the “Parkinson” dataset, an initial

Table 9 Parkinson dataset

Tuned random forest	11	16	21
6	$2.0 \times 10^{-6***}$	$6.2 \times 10^{-5***}$	0.0013^{**}
11	–	$0.0785^{\bullet}$	–
21	–	$8.5 \times 10^{-5***}$	–
31	0.00016^{***}	$7.6 \times 10^{-6***}$	$2.4 \times 10^{-5***}$
Tuned *k*nn	11	16	21
6	$1.7 \times 10^{-6***}$	$1.7 \times 10^{-6***}$	$4.9 \times 10^{-5***}$
31	0.00162^{**}	$2.4 \times 10^{-6***}$	0.00026^{***}

Adjusted *p* values comparing SMAE.int metric across tuned random forest and *k*nn models, for each of the five lengths assigned to the rightmost interval. The *p* values indicate that the mean SMAE.int value for the model in the row is significantly lower (and thus better) than that of the model in the column



(a) Facial age dataset. Random forest

(b) Facial age dataset. $k$nn

(c) Abalone dataset. Random forest

(d) Abalone dataset. $k$nn

(e) Parkinson dataset. Random forest

(f) Parkinson dataset. $k$nn

Fig. 7 Boxplots of the SMAE.int metric as an indicator of prediction performance for models tuned using this metric, with lower values indicating better performance. The figure compares different lengths assigned to the rightmost interval across three datasets and two classifiers

increase is observed before the metric decreases with longer intervals. These patterns align with the findings in Tables 7-9 suggesting that optimal interval length may vary depending on the dataset's characteristics.

6 Conclusions

In this research, our primary aim has been to address the lack of specific performance metrics for classification in interval scale. To tackle this issue, we have devised a methodology aimed at extending any ordinal classification performance metric based on a confusion matrix to this scenario. This methodology involves replacing the class distances in the ordinal metric with interval distances (such as Hausdorff distance) and substituting the number of instances in each class or interval with its "density", obtained by dividing by the length of the interval.

Furthermore, we have also introduced a procedure to normalize the metrics so that they fall within the interval [0, 1]. By applying both the methodology and the normalization procedure to two well-known ordinal classification metrics, MAE and TC, we obtain their respective normalized metrics for interval scale. For these metrics, we demonstrate that they satisfy certain norm properties (positiveness, maximal agreement, homogeneity, and triangle inequality), as well as other desirable properties such as scale invariance, strict monotonicity, and imbalance. The fact that the introduced metrics exhibit such favorable properties reinforces the validity of the methodology used for their construction.

However, in cases where the data on the interval scale originates from a discretized continuous variable or from binning a count variable, it is common for the rightmost interval to be unbounded. A problem arises thereafter, since the metrics introduced earlier require the length of this interval to be defined. We propose the following as a solution for this issue: Firstly, we dissociate the number of instances in the rightmost interval, which will be determined by the available data, from the length that can be arbitrarily assigned to this interval for the purpose of calculating the performance metric. Secondly, we posit that the optimal length to assign to this interval is the one that minimizes the maximum achievable value of the metric (previous to normalization). It is important to note that the value of the metric depends on the interval's length, and its maximum value serves as the divisor for normalizing the metric to the interval [0, 1]. Consequently, we maximize the sensitivity of the metric to changes in the confusion matrix, thereby enhancing its ability to distinguish differences among classifiers, which aligns with the overarching objective of the metric itself.

In this regard, we have indeed identified the optimal value for the length of the unbounded rightmost interval in a particular case (balanced with 3 classes), and we have derived an upper bound in the general case. This upper bound has enabled us to implement a numerical procedure that yields an approximation of the sought-after value, which we have tested across various scenarios. While this bound may not seem tight in the examples utilized, it is still valuable as it provides a range within which the solution to the numerical problem must necessarily lie. This allows us to implement an iterative procedure capable of converging towards the solution, and this would not be possible without having such an upper bound available. Ultimately, applying our numerical approximation procedure to various scenarios involving 3 and 4 classes highlights the practical utility of our algorithm. This approach not only yields specific results but also underscores the effectiveness of our implemented methodology in achieving accurate solutions.

In addition to the theoretical contributions, we investigated the practical impact of the proposed interval-scale metrics through experiments on three real-world datasets and two classifiers. These experiments compared the interval-scale metrics with traditional metrics like Accuracy and MAE to assess their utility in guiding hyper-parameter tuning. While the results are not definitively conclusive, they reveal a promising trend toward enhanced model performance when using interval-scale metrics, suggesting their potential value for practical applications in classification tasks. This evidence points to a direction that merits further exploration.

Appendix A: New metrics: proofs of their properties

In this appendix, we provide the proofs of the properties satisfied by the metrics obtained from the adaptations of MAE and TC, which have been introduced in Sect. 3.

Proofs for the interval-scale metrics MAE^{int} and TC^{int}

Recall that MAE^{int} and TC^{int}, as formulated in Definitions 2 and 3, represent the interval-scale, non-normalized versions of the metrics MAE and TC, respectively. Several of their properties are stated in Proposition 1 and proved below.

Proof of Proposition 1 We will prove all the different properties of TC^{int}, and also property *4.* of MAE^{int}. Similar reasoning leads to the proof of properties *1.*, *2.* and *3.* of MAE^{int}.

1. *Non-negativity/Positiveness.* This property follows directly from Definition 3.

2. *Positive definiteness/Maximal agreement.* The proof of this property is trivial since TC$^{int}(A) = 0$ if and only if A is diagonal, because $\gamma_{ij}^{int}\, d(I_i, I_j) > 0$.

3'. *Homogeneity (of degree 1).* This property is consequence of the fact that if $A = (a_{ij})_{i,j=1,\dots,r} \in \mathcal{A}$, then $A \in \mathcal{S}^{n_1,\dots,n_r}$ where $n_j = \sum_{i=1}^r a_{ij}$ for $j = 1,\dots,r$, and then $k\,A \in \mathcal{S}^{k\,n_1,\dots,k\,n_r}$. Since for any i, j, γ_{ij}^{int} matches for the two spaces $\mathcal{S}^{n_1,\dots,n_r}$ and $\mathcal{S}^{k\,n_1,\dots,k\,n_r}$, we then obtain that

$$\mathrm{TC}^{int}(k\,A) = \sum_{i,j} k\,a_{ij}\,\gamma_{ij}^{int}\,d(I_i, I_j) = k\,\sum_{i,j} a_{ij}\,\gamma_{ij}^{int}\,d(I_i, I_j) = k\,\mathrm{TC}^{int}(A).$$

4'. *(Restricted) Subadditivity/Triangle inequality.* We assume that A and B belong to the same space $S^{n_1,\dots,n_r}$, where $n_j = \sum_{i=1}^r a_{ij} = \sum_{i=1}^r b_{ij}$. Then, $A + B \in \mathcal{S}^{2\,n_1,\dots,2\,n_r}$, and since γ_{ij}^{int} matches for the two spaces $\mathcal{S}^{n_1,\dots,n_r}$ and $\mathcal{S}^{2\,n_1,\dots,2\,n_r}$, consequently we obtain that

$$\begin{aligned}\mathrm{TC}^{int}(A+B) &= \sum_{i,j}(a_{ij} + b_{ij})\,\gamma_{ij}^{int}\,d(I_i, I_j)\\ &= \sum_{i,j} a_{ij}\,\gamma_{ij}^{int}\,d(I_i, I_j) + \sum_{i,j} b_{ij}\,\gamma_{ij}^{int}\,d(I_i, I_j)\\ &= \mathrm{TC}^{int}(A) + \mathrm{TC}^{int}(B).\end{aligned}$$

4. *Subadditivity/Triangle inequality.* Let denote by N_A and N_B, respectively, the sum of the elements of matrices A and B. Then, the sum of elements of matrix $A + B$ is $N_A + N_B$, and in order to obtain the triangle inequality we observe that

$$\mathrm{MAE}^{int}(A+B) \leq \mathrm{MAE}^{int}(A) + \mathrm{MAE}^{int}(B)$$

is equivalent to

$$\begin{aligned}
&\sum_{i,j}(a_{ij}+b_{ij})d(I_i, I_j) \\
&\quad \leq \left(1+\frac{N_B}{N_A}\right)\sum_{i,j} a_{ij}d(I_i, I_j) + \left(1+\frac{N_A}{N_B}\right)\sum_{i,j} b_{ij}d(I_i, I_j) \\
&\quad = \sum_{i,j}(a_{ij}+b_{ij})\, d(I_i, I_j) + \sum_{i,j}\left(\frac{N_B}{N_A}\, a_{ij} + \frac{N_A}{N_B}\, b_{ij}\right) d(I_i, I_j)
\end{aligned} \tag{A1}$$

but (A1) holds trivially since the term $\sum_{i,j}(\frac{N_B}{N_A}\, a_{ij} + \frac{N_A}{N_B}\, b_{ij})\, d(I_i, I_j)$ is positive.

Proofs for the normalized (ordinal) metrics $\widehat{\mathrm{MAE}}$ and $\widehat{\mathrm{TC}}$

Proposition 2 provides expressions for MAE_{max} and TC_{max}. These expressions will be used to examine the behavior of these quantities in both the binary and balanced cases.

Proof of Corollary 1 Applying Proposition 2 to $r=2$, with $N = n_1 + n_2$, we obtain that $k_1 = g_1 = 2$, $k_2 = g_2 = 1$. Then,

$$G_1 = \frac{n_1}{N},\ G_2 = \frac{n_2}{N} \quad \text{and} \quad \mathrm{MAE}_{max} = G_1 + G_2 = 1,$$

and, analogously,

$$K_1 = n_1,\ K_2 = n_2 \quad \text{and} \quad \mathrm{TC}_{max} = K_1 + K_2 = N.$$

Proof of Corollary 2 Observe that, in the *balanced case*,

$$A_{max} = \begin{pmatrix} \underbrace{\begin{matrix} 0 & \dots & 0 \\ 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \\ N/r & \dots & N/r \end{matrix}}_{h \text{ columns}} & \underbrace{\begin{matrix} N/r & \dots & N/r \\ 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \\ 0 & 0 & 0 \end{matrix}}_{r-h \text{ columns}} \end{pmatrix}$$

As above, if r is odd, the first block could be composed by $h-1$ columns (and the second by $r-h+1$ columns, respectively) without causing any change in the value of TC_{max} or MAE_{max}. Since $n_j = N/r$ for $j = 1, \dots, r$, we have that

$$k_j = g_j = \arg\max_{\ell=1,\dots,r} |\ell - j| = \begin{cases} r & \text{if } j = 1, \dots, h, \\ 1 & \text{if } j = h+1, \dots, r, \end{cases}$$

Note that, if r is odd, we can define $k_h = 1$ instead of $k_h = r$ and this does not change the following calculations. Then, we can write

$$
\begin{aligned}
\mathrm{TC}_{max} &= \sum_{j=1}^{r} \frac{N}{r}\, \frac{N-N/r}{N/r}\, |k_j - j| = N\, \frac{r-1}{r} \sum_{j=1}^{r} |k_j - j| \\
&= N\, \frac{r-1}{r} \Big(\sum_{j=1}^{h} (r-j) + \sum_{j=h+1}^{r} (j-1) \Big) = N\, \frac{r-1}{r} \Big(\sum_{i=r-h}^{r-1} i + \sum_{i=h}^{r-1} i \Big) \\
&= N\, \frac{r-1}{r} \Big((r-h)\, h + \frac{r\,(r-1)}{2} \Big) = N \Big(\frac{(r-1)^2}{2} + \frac{(r-1)\,(r-h)\,h}{r} \Big),
\end{aligned}
$$

where we have used that the sum of the n consecutive positive integers between a and b, both included, is $\frac{a+b}{2}\, n$. Analogously, using the same type of argument we can write

$$
\mathrm{MAE}_{max} = \sum_{j=1}^{r} \frac{N}{r}\, \frac{1}{N}\, |k_j - j| = \frac{1}{r} \Big((r-h)\, h + \frac{r\,(r-1)}{2} \Big) = \frac{r-1}{2} + \frac{(r-h)\, h}{r}.
$$

Proof of Corollary 3 By Corollary 1, we know the values of MAE_{max} and TC_{max} if $r = 2$.

On the other hand, we denote by $\mathrm{TC}_{max}(r)$ the value of TC_{max} when the number of classes is r. For any $k \geq 1$, it is sufficient to prove that

$$
\mathrm{TC}_{max}(2k) < \mathrm{TC}_{max}(2k+1) < \mathrm{TC}_{max}(2k+2).
$$

This is a mere verification since by Corollary 2 these inequalities are equivalent to

$$
(3\,k-1)\,(2\,k-1)\,k < (3\,k+2)\,(2\,k)\,k < (3\,k+2)\,(2\,k+1)\,(k+1),
$$

which hold trivially for all $k \geq 1$.

Analogously,

$$
\mathrm{MAE}_{max}(2k) < \mathrm{MAE}_{max}(2k+1) < \mathrm{MAE}_{max}(2k+2)
$$

is equivalent to

$$
(3\,k-1)\,k < (3\,k+2)\,k < (3\,k+2)\,(k+1),
$$

which is obviously fulfilled for all $k \geq 1$, finishing the proof. □

Proofs for the normalized interval-scale metrics $\widehat{\mathrm{MAE}^{int}}$ and $\widehat{\mathrm{TC}^{int}}$

Metrics $\widehat{\mathrm{MAE}^{int}}$ and $\widehat{\mathrm{TC}^{int}}$, which are introduced in Definition 6, have maximum values MAE^{int}_{max} and TC^{int}_{max}, provided in Proposition 3. In this section, we examine how these maximum values behave in the binary and proportional cases (see Corollaries 4 and 5, respectively) and establish several properties of these metrics (Propositions 4 and 5).

Proof of Corollary 4 Applying Proposition 3 to $r = 2$, with $N = n_1 + n_2$, we obtain that $\widehat{k_1} = \widehat{g_1} = 2$, $\widehat{k_2} = \widehat{g_2} = 1$ and $d(I_1, I_2) = d(I_2, I_1) = \max(a_2 - a_1, b_2 - b_1) = \max(\ell_1, \ell_2)$ (since $a_2 = b_1$). Then,

$$
\widehat{G}_1 = \frac{n_1}{N}\, \max(\ell_1, \ell_2), \quad \widehat{G}_2 = \frac{n_2}{N}\, \max(\ell_1, \ell_2)
$$

and

$$\mathrm{MAE}^{int}_{max} = \widehat{G}_1 + \widehat{G}_2 = \max(\ell_1, \ell_2).$$

Analogously,

$$\widehat{K}_1 = n_1 \, \frac{\delta_2}{\delta_2} \, d(I_2, I_1) = n_1 \; \max(\ell_1, \ell_2), \quad \widehat{K}_2 = n_2 \, \frac{\delta_1}{\delta_1} \, d(I_1, I_2) = n_2 \; \max(\ell_1, \ell_2)$$

and finally

$$\mathrm{TC}^{int}_{max} = \widehat{K}_1 + \widehat{K}_2 = N \times \max(\ell_1, \ell_2).$$

□

Proof of Corollary 5 By (7), $\gamma^{int}_{ij} = r - 1$ in the proportional case scenario, and then

$$\mathrm{TC}^{int} = \frac{r-1}{N} \, \mathrm{MAE}^{int}.$$

Moreover, $\widehat{g}_j = \widehat{k}_j$ for $j = 1, \ldots, r$ and then, by Proposition 3 we get that in the *proportional case*,

$$\mathrm{TC}^{int}_{max} = \frac{r-1}{N} \, \mathrm{MAE}^{int}_{max}.$$

As a trivial consequence, then, $\widehat{\mathrm{TC}^{int}} = \widehat{\mathrm{MAE}^{int}}$. □

Proof of Proposition 4

1. **Scale invariance**. Let f denote a change of scale function, that is, a linear function of the form $f(x) = c\,x + d$ with $c > 0$. We use an asterisk to denote the quantities after the scale change on the data. Therefore, we will prove that

$$\widehat{\mathrm{TC}^{int}}^{*} = \widehat{\mathrm{TC}^{int}}.$$

Indeed, the change of scale is monotonically increasing since $c > 0$, and the intervals after the change of scale are $I^*_1, \ldots, I^*_r$ with $I^*_i = [f(a_i), f(b_i)) = [c\,a_i + d,\, c\,b_i + d)$, with length $\ell^*_i = f(b_i) - f(a_i) = c\,(b_i - a_i) = c\,\ell_i$. The Hausdorff distance between intervals I^*_i and I^*_j is:

$$\begin{aligned} d(I^*_i, I^*_j) &= \max\{\,|f(a_j) - f(a_i)|,\; |f(b_j) - f(b_i)|\,\} \\ &= |c| \;\max\{\,|a_j - a_i|,\, |b_j - b_i|\,\} = c\,d(I_i, I_j) \end{aligned}$$

and the densities after the change of scale are $\delta^*_i = \frac{n_i}{\ell^*_i} = \frac{\delta_i}{c}$, and then, $\gamma^{int^*}_{ij} = \gamma^{int}_{ij}$. With this, for any confusion matrix $A = (a_{ij})_{i,j=1,\ldots,r} \in \mathcal{S}$,

$$\mathrm{TC}^{\;int\,*}(A) = \sum_{i,j=1}^{r} a_{ij}\,\gamma^{int^*}_{ij}\; d(I^*_i, I^*_j) = \sum_{i,j=1}^{r} a_{ij}\,\gamma^{int}_{ij}\; c\,d(I_i, I_j) = c\;\; \mathrm{TC}^{\;int}(A).$$

On the other hand, for any $j = 1, \ldots, r$,

$$\widehat{k}_j^* = \arg\max_{\ell=1,\ldots,r} \frac{d(I_\ell^*, I_j^*)}{\delta_\ell^*} = \arg\max_{\ell=1,\ldots,r} c^2\, \frac{d(I_\ell, I_j)}{\delta_\ell} = \widehat{k}_j$$

and

$$\widehat{K}_j^* = n_j\, \frac{\sum_{\ell\neq j} \delta_\ell^*}{\delta_{\widehat{k}_j}^*}\, d(I_{\widehat{k}_j}^*, I_j^*) = n_j\, \frac{\sum_{\ell\neq j} \delta_\ell}{\delta_{\widehat{k}_j}}\, c\, d(I_{\widehat{k}_j}, I_j) = c\, \widehat{K}_j$$

giving that $\mathrm{TC}_{max}^{int\,*} = \sum_{j=1}^r \widehat{K}_j^* = c\ \mathrm{TC}_{max}^{int}$. Finally, then,

$$\widehat{\mathrm{TC}^{int}}^{\,*}(A) = \frac{\mathrm{TC}^{int\,*}(A)}{\mathrm{TC}_{max}^{int\,*}} = \frac{c\ \mathrm{TC}^{int}(A)}{c\ \mathrm{TC}_{max}^{int}} = \widehat{\mathrm{TC}^{int}}(A).$$

2. **Strict Monotonicity**. If we change the prediction of an item of a fixed class j_0 from $i_0 \neq j_0$ to $k_0 \neq j_0$ such that

$$d(I_{k_0}, I_{j_0}) > d(I_{i_0}, I_{j_0}) \quad \text{while} \quad \delta_{k_0} = \delta_{i_0},$$

then $w_{k_0 j_0}^{int} > w_{i_0 j_0}^{int}$ and TC^{int} increases by

$$w_{k_0 j_0}^{int} - w_{i_0 j_0}^{int} = \gamma_{i_0 j_0}^{int} \left(d(I_{k_0}, I_{j_0}) - d(I_{i_0}, I_{j_0})\right) > 0$$

(note that $\gamma_{i_0 j_0}^{int} = \gamma_{k_0 j_0}^{int}$). Since TC_{max}^{int} does not depend on the predictions, that is, of the specific form of the confusion matrix, being that it belongs to $\mathcal{S}$, it remains unchanged, and therefore the increase in TC^{int} translates into an increase in $\widehat{\mathrm{TC}^{int}}$.

Similarly, if we misclassify an item belonging to class j_0 as k_0, instead of correctly categorizing it (case $i_0 = j_0$), the total cost TC^{int} increases by $w_{k_0 j_0}^{int} = \gamma_{j_0 j_0}^{int}\, d(I_{k_0}, I_{j_0}) > 0$. Compare this result with Axiom 1 (MON) (Sebastiani, 2015).

3. **Imbalance**. If we consider classes j_0 and j_1, with $\delta_{j_0} < \delta_{j_1}$, we wonder what it is the effect to misclassify an item of any of these classes to class i_0, assuming that $d(I_{i_0}, I_{j_0}) = d(I_{i_0}, I_{j_1})$ (denote these distances by Δ).

The effect on TC^{int} of misclassify an item of class j_0 (respect., of class j_1) into class i_0 is an increase by quantity:

$$w_{i_0 j_0}^{int} = \gamma_{i_0 j_0}^{int}\, d(I_{i_0}, I_{j_0}) \quad (\text{respect.,} \quad w_{i_0 j_1}^{int} = \gamma_{i_0 j_1}^{int}\, d(I_{i_0}, I_{j_1}))$$

and the difference between them is:

$$w_{i_0 j_0}^{int} - w_{i_0 j_1}^{int} = \Delta \left(\gamma_{i_0 j_0}^{int} - \gamma_{i_0 j_1}^{int}\right) = \Delta\, \frac{\delta_{j_1} - \delta_{j_0}}{\delta_{i_0}} > 0.$$

As a consequence, TC^{int} increases more if the misclassified item belongs to class j_0 than if it belongs to class j_1. Since TC_{max}^{int} does not depend on the specific form of the confusion matrix in $\mathcal{S}$, the same applies to $\widehat{\mathrm{TC}^{int}}$.

Besides, if instead we consider classes i_0 and i_1, with $\delta_{i_0} < \delta_{i_1}$, what is the effect of misclassifying an item of a class j_0 to any of these classes, assuming that $d(I_{i_0}, I_{j_0}) = d(I_{i_1}, I_{j_0})$? Denote now with Δ these distances between intervals. The effect on TC ^{int} of misclassifying an item of class j_0 into class i_0 (respectively, into class i_1) is an increase of quantity:

$$w^{int}_{i_0 j_0} = \gamma^{int}_{i_0 j_0}\, d(I_{i_0}, I_{j_0}) \quad (\text{respect.,} \quad w^{int}_{i_1 j_0} = \gamma^{int}_{i_1 j_0}\, d(I_{i_1}, I_{j_0}))$$

and the difference is:

$$w^{int}_{i_0 j_0} - w^{int}_{i_1 j_0} = \Delta\left(\gamma^{int}_{i_0 j_0} - \gamma^{int}_{i_1 j_0}\right) = \Delta\left(\sum_{k \neq j_0}^{r} \delta_k\right)\left(\frac{1}{\delta_{i_0}} - \frac{1}{\delta_{i_1}}\right) > 0.$$

This is, in this case TC ^{int} increases more if the item is misclassified in class i_0 than in class i_1, which also holds for $\widehat{\text{TC}}^{int}$.

□

Proof of Proposition 5 The proof of each property follows directly from those established in Proposition 1. Specifically, Properties *1 (Non-negativity/Positiveness)* and *2 (Positive definiteness/Maximal agreement)* are straightforwardly satisfied. The remaining properties are justified as follows.

3. *Homogeneity (of degree 0)*. By the proof of property *3'.* in Proposition 1, we know that if $A \in \mathcal{S}^{n_1,\ldots,n_r}$ then $kA \in \mathcal{S}^{k\,n_1,\ldots,k\,n_r}$ and

$$\text{TC}^{int}(kA) = k\,\text{TC}^{int}(A),\ \ \text{MAE}^{int}(kA) = \text{MAE}^{int}(A).$$

Analogously, since $A^{k\,n_1,\ldots,k\,n_r}_{max} = k\,A^{n_1,\ldots,n_r}_{max}$, we have that

$$\text{TC}^{int}(A^{k\,n_1,\ldots,k\,n_r}_{max}) = k\,\text{TC}^{int}(A^{n_1,\ldots,n_r}_{max}),$$
$$\text{MAE}^{int}(A^{k\,n_1,\ldots,k\,n_r}_{max}) = \text{MAE}^{int}(A^{n_1,\ldots,n_r}_{max})$$

and then,

$$\begin{aligned}\widehat{\text{TC}}^{int}(kA) &= \frac{\text{TC}^{int}(kA)}{\text{TC}^{int}(A^{k\,n_1,\ldots,k\,n_r}_{max})}\\ &= \frac{k\,\text{TC}^{int}(A)}{k\,\text{TC}^{int}(A^{n_1,\ldots,n_r}_{max})} = \widehat{\text{TC}}^{int}(A),\\ \widehat{\text{MAE}}^{int}(kA) &= \frac{\text{MAE}^{int}(kA)}{\text{MAE}^{int}(A^{k\,n_1,\ldots,k\,n_r}_{max})}\\ &= \frac{\text{MAE}^{int}(A)}{\text{MAE}^{int}(A^{n_1,\ldots,n_r}_{max})} = \widehat{\text{MAE}}^{int}(A).\end{aligned}$$

4. *(Restricted) Subadditivity/Triangle inequality*. By the proof of property *4'.* in Proposition 1, we know that if A and B are in $S^{n_1,\ldots,n_r}$, then $A + B \in \mathcal{S}^{2\,n_1,\ldots,2\,n_r}$ and $\text{TC}^{int}(A+B) = \text{TC}^{int}(A) + \text{TC}^{int}(B)$. Moreover, $A^{2\,n_1,\ldots,2\,n_r}_{max} = 2\,A^{n_1,\ldots,n_r}_{max}$, and we have that

$$\text{TC}^{int}(A^{2\,n_1,\ldots,2\,n_r}_{max}) = 2\,\text{TC}^{int}(A^{n_1,\ldots,n_r}_{max}),$$

and then,

$$\begin{aligned}\widehat{\mathrm{TC}^{int}}(A+B) &= \frac{\mathrm{TC}^{int}(A+B)}{\mathrm{TC}^{int}(A_{max}^{2\,n_1,\dots,2\,n_r})} = \frac{\mathrm{TC}^{int}(A)+\mathrm{TC}^{int}(B)}{2\,\mathrm{TC}^{int}(A_{max}^{n_1,\dots,n_r})} \\ &= \frac{1}{2}\left(\widehat{\mathrm{TC}^{int}}(A)+\widehat{\mathrm{TC}^{int}}(B)\right).\end{aligned}$$

By the proof of property *4.* in Proposition 1, we have that

$$\mathrm{MAE}^{int}(A+B) \le \mathrm{MAE}^{int}(A)+\mathrm{MAE}^{int}(B),$$

but in this scenario we can say more:

$$\begin{aligned}\mathrm{MAE}^{int}(A+B) &= \frac{1}{2N}\sum_{i,j}(a_{ij}+b_{ij})\,d(I_i, I_j) \\ &= \frac{1}{2}\left(\frac{1}{N}\sum_{i,j}a_{ij}\,d(I_i, I_j)+\frac{1}{N}\sum_{i,j}b_{ij}\,d(I_i, I_j)\right) \\ &= \frac{1}{2}\left(\mathrm{MAE}^{int}(A)+\mathrm{MAE}^{int}(B)\right),\end{aligned}$$

and $\mathrm{MAE}^{int}(A_{max}^{2\,n_1,\dots,2\,n_r}) = \frac{1}{2N}\sum_{j=1}^{r} 2\,n_j\,d(I_{\widehat{g}_j}, I_j) = \mathrm{MAE}^{int}(A_{max}^{n_1,\dots,n_r})$. Then,

$$\begin{aligned}\widehat{\mathrm{MAE}^{int}}(A+B) &= \frac{\mathrm{MAE}^{int}(A+B)}{\mathrm{MAE}^{int}(A_{max}^{2\,n_1,\dots,2\,n_r})} = \frac{\frac{1}{2}\left(\mathrm{MAE}^{int}(A)+\mathrm{MAE}^{int}(B)\right)}{\mathrm{MAE}^{int}(A_{max}^{n_1,\dots,n_r})} \\ &= \frac{1}{2}\left(\widehat{\mathrm{MAE}^{int}}(A)+\widehat{\mathrm{MAE}^{int}}(B)\right).\end{aligned}$$

Appendix B Proof of Proposition 6

To find the value of x that minimizes TC_{max}^{int} we use its expression, as given in Proposition 3, that is,

$$\mathrm{TC}_{max}^{int} = \sum_{j=1}^{3}\widehat{K}_j, \quad \text{with } \widehat{K}_j = n\,\frac{\sum_{\ell\neq j}\delta_\ell}{\delta_{\widehat{k}_j}}\,d(I_{\widehat{k}_j}, I_j)$$

where $\widehat{k}_j = \arg\max_{\ell=1,2,3}\frac{d(I_\ell, I_j)}{\delta_\ell}$. Note that

$$d(I_1, I_2) = \max(1, L), \quad d(I_2, I_3) = \max(L, x), \quad d(I_1, I_3) = L + \max(1, x)$$

so a first division in cases will be according to the value of $\max(1, L)$.

Case a) $L \le 1$ (then, $d(I_1, I_2) = \max(1, L) = 1$).

If $j = 1$, $\widehat{k}_1 = \arg\max_{\ell=1,2,3}\frac{d(I_\ell, I_1)}{\delta_\ell}$. Taking into account that

$$\frac{d(I_\ell, I_1)}{\delta_\ell} = \begin{cases} \frac{L}{n} & \text{if } \ell = 2 \\ \frac{x}{n}\big(L+\max(1, x)\big) = \begin{cases} \frac{x}{n}(L+1) & \text{if } 0 < x \le 1 \\ \frac{x}{n}(x+L) & \text{if } x > 1 \end{cases} & \text{if } \ell = 3 \end{cases}$$

we have that

$$\widehat{k}_1 = \begin{cases} 2 & \text{if } 0 < x \leq \frac{L}{L+1} \\ 3 & \text{if } x > \frac{L}{L+1}. \end{cases}$$

Then,

$$\widehat{K}_1 = n \, \frac{\sum_{\ell=2,3} \delta_\ell}{\delta_{\widehat{k}_1}} \, d(I_{\widehat{k}_1}, I_1) = \begin{cases} n \, \frac{(x+L)}{x} & \text{if } 0 < x \leq \frac{L}{L+1} \\ n \, \frac{(L+1)\,(x+L)}{L} & \text{if } \frac{L}{L+1} < x \leq 1 \\ n \, \frac{(x+L)^2}{L} & \text{if } x > 1. \end{cases}$$

Analogously,

$$\widehat{k}_2 = \begin{cases} 1 & \text{if } 0 < x \leq 1 \\ 3 & \text{if } x > 1 \end{cases}, \qquad \widehat{k}_3 = 1 \ \forall x > 0,$$

and

$$\widehat{K}_2 = n \, \frac{\sum_{\ell=1,3} \delta_\ell}{\delta_{\widehat{k}_2}} \, d(I_{\widehat{k}_2}, I_2) = \begin{cases} n \, (1 + \frac{1}{x}) & \text{if } 0 < x \leq 1 \\ n \, x \, (x + 1) & \text{if } x > 1\,, \end{cases}$$

$$\widehat{K}_3 = n \, \frac{\sum_{\ell=1,2} \delta_\ell}{\delta_{\widehat{k}_3}} \, d(I_{\widehat{k}_3}, I_3) = \begin{cases} n \, \frac{(L+1)^2}{L} & \text{if } 0 < x \leq 1 \\ n \, \frac{(L+1)\,(x+L)}{L} & \text{if } x > 1. \end{cases}$$

Finally,

$$\widehat{\mathrm{TC}}_{max} = \sum_{j=1}^{3} \widehat{K}_j = \begin{cases} n \left(\frac{(L+1)}{x} + L + 4 + \frac{1}{L} \right) & \text{if } 0 < x \leq \frac{L}{L+1} \\ n \left((1 + \frac{1}{L}) x + \frac{1}{x} + 2L + 4 + \frac{1}{L} \right) & \text{if } \frac{L}{L+1} < x \leq 1 \\ n \left((1 + \frac{1}{L}) x^2 + (4 + \frac{1}{L}) x + 2L + 1 \right) & \text{if } x > 1. \end{cases}$$

Note that $\widehat{\mathrm{TC}}_{max}$ is a continuous function of $x > 0$, and that it is piece-wise differentiable. Its minimum at each interval of definition is given by:

$$\min \widehat{\mathrm{TC}}_{max}(x) = \begin{cases} 2n \left(L + 3 + \frac{1}{L} \right) & \text{at } x_{min} = \frac{L}{L+1}, \text{ if } 0 < x \leq \frac{L}{L+1} \\ n \left(2\sqrt{\frac{L+1}{L}} + 2L + 4 + \frac{1}{L} \right) & \text{at } x_{min} = \sqrt{\frac{L}{L+1}}, \text{ if } \frac{L}{L+1} < x \leq 1 \\ 2n \left(L + 3 + \frac{1}{L} \right) & \text{at } x_{min} = 1, \text{ if } x > 1. \end{cases}$$

Since $2\sqrt{\frac{L+1}{L}} + 2L + 4 + \frac{1}{L} < 2\left(L + 3 + \frac{1}{L} \right)$ holds always, we have that $\min \widehat{\mathrm{TC}}_{max}(x) = n \left(2\sqrt{\frac{L+1}{L}} + 2L + 4 + \frac{1}{L} \right)$ and is reached at $x_{min} = \sqrt{\frac{L}{L+1}}$.

Case b) $L > 1$ (then, $d(I_1, I_2) = \max(1, L) = L$).

If $j = 1$, $\widehat{k}_1 = \arg\max_{\ell=1,2,3} \frac{d(I_\ell, I_1)}{\delta_\ell}$ with

$$\frac{d(I_\ell, I_1)}{\delta_\ell} = \begin{cases} \frac{L^2}{n} & \text{if } \ell = 2 \\ \frac{x}{n} \left(L + \max(1, x) \right) = \begin{cases} \frac{x}{n} \, (L + 1) & \text{if } 0 < x \leq 1 \\ \frac{x}{n} \, (x + L) & \text{if } x > 1 \end{cases} & \text{if } \ell = 3 \end{cases}$$

and therefore, distinguishing when $L^2 < x(L + 1)$ if $x \leq 1$, and when $L^2 < x(x + L)$ if $x > 1$, we have that

$$\begin{cases} \text{If } 1 < L \leq \frac{1+\sqrt{5}}{2}, \ \widehat{k}_1 = \begin{cases} 2 & \text{if } 0 < x \leq \frac{L^2}{L+1} \\ 3 & \text{if } x > \frac{L^2}{L+1} \end{cases} \\ \text{If } L > \frac{1+\sqrt{5}}{2}, \qquad \widehat{k}_1 = \begin{cases} 2 & \text{if } 0 < x \leq \frac{L}{2}(\sqrt{5}-1) \\ 3 & \text{if } x > \frac{L}{2}(\sqrt{5}-1). \end{cases} \end{cases}$$

Then, using that $\widehat{K}_1 = n \frac{\sum_{\ell=2,3} \delta_\ell}{\delta_{\widehat{k}_1}} d(I_{\widehat{k}_1}, I_1)$ we have that

$$\begin{cases} \text{If } 1 < L \leq \frac{1+\sqrt{5}}{2}, \ \widehat{K}_1 = \begin{cases} n \frac{L(x+L)}{x} & \text{if } 0 < x \leq \frac{L^2}{L+1} \\ n \frac{(L+1)(x+L)}{L} & \text{if } \frac{L^2}{L+1} < x \leq 1 \\ n \frac{(x+L)^2}{L} & \text{if } x > 1 \end{cases} \\ \text{If } L > \frac{1+\sqrt{5}}{2}, \qquad \widehat{K}_1 = \begin{cases} n \frac{L(x+L)}{x} & \text{if } 0 < x \leq \frac{L}{2}(\sqrt{5}-1) \\ n \frac{(x+L)^2}{L} & \text{if } x > \frac{L}{2}(\sqrt{5}-1). \end{cases} \end{cases}$$

With the same approach, we can get that for $j = 2$,

$$\widehat{k}_2 = \begin{cases} 1 & \text{if } 0 < x \leq 1 \\ 3 & \text{if } x > 1 \end{cases}, \qquad \widehat{K}_2 = \begin{cases} nL(1 + \frac{1}{x}) & \text{if } 0 < x \leq 1 \\ nL(x+1) & \text{if } 1 < x \leq L \\ nx(x+1) & \text{if } x > L, \end{cases}$$

and for $j = 3$,

$$\begin{cases} \text{If } 1 < L \leq \frac{1+\sqrt{5}}{2}, \ \widehat{k}_3 = \begin{cases} 1 & \text{if } 0 < x \leq \frac{L}{L-1} \\ 2 & \text{if } x > \frac{L}{L-1} \end{cases} \\ \text{If } \frac{1+\sqrt{5}}{2} < L \leq 2, \ \widehat{k}_3 = \begin{cases} 2 & \text{if } 0 < x \leq L(L-1) \\ 1 & \text{if } L(L-1) < x \leq \frac{L}{L-1} \\ 2 & \text{if } x > \frac{L}{L-1} \end{cases} \\ \text{If } L > 2, \qquad \widehat{k}_3 = 2 \quad \forall x > 0, \end{cases}$$

$$\begin{cases} \text{If } 1 < L \leq \frac{1+\sqrt{5}}{2}, \ \widehat{K}_3 = \begin{cases} n \frac{(L+1)^2}{L} & \text{if } 0 < x \leq 1 \\ n \frac{(L+1)(x+L)}{L} & \text{if } 1 < x \leq \frac{L}{L-1} \\ n(L+1)x & \text{if } x > \frac{L}{L-1} \end{cases} \\ \text{If } \frac{1+\sqrt{5}}{2} < L \leq 2, \ \widehat{K}_3 = \begin{cases} nL(L+1) & \text{if } 0 < x \leq L(L-1) \\ n \frac{(L+1)(x+L)}{L} & \text{if } L(L-1) < x \leq \frac{L}{L-1} \\ n(L+1)x & \text{if } x > \frac{L}{L-1} \end{cases} \\ \text{If } L > 2, \qquad \widehat{K}_3 = \begin{cases} nL(L+1) & \text{if } 0 < x \leq L \\ n(L+1)x & \text{if } x > L. \end{cases} \end{cases}$$

Then, using that $\widehat{\text{TC}}_{max} = \sum_{j=1}^{3} \widehat{K}_j$, we obtain the following:

Case b.1) $1 < L \leq \frac{1+\sqrt{5}}{2}$

$$\widehat{\mathrm{TC}}_{max} = \begin{cases} n\left((L^2+L)\,\frac{1}{x} + (3\,L+2+\frac{1}{L})\right) & \text{if } 0 < x \le \frac{L^2}{L+1} \\ n\left((1+\frac{1}{L})\,x + \frac{L}{x} + (3\,L+3+\frac{1}{L})\right) & \text{if } \frac{L^2}{L+1} < x \le 1 \\ n\left(\frac{x^2}{L} + (3+L+\frac{1}{L})\,x + (3\,L+1)\right) & \text{if } 1 < x \le L \\ n\left((1+\frac{1}{L})\,x^2 + (4+\frac{1}{L})\,x + (2\,L+1)\right) & \text{if } L < x \le \frac{L}{L-1} \\ n\left((1+\frac{1}{L})\,x^2 + (4+L)\,x + L\right) & \text{if } x > \frac{L}{L-1}, \end{cases}$$

which is a continuous function of $x > 0$ and piece-wise differentiable. Its minimum at each interval of definition is given by:

$$\min \widehat{\mathrm{TC}}_{max}(x) = \begin{cases} 2\,n\left(2\,L+2+\frac{1}{L}\right) & \text{at } x_{min} = \frac{L^2}{L+1},\ \text{if } 0 < x \le \frac{L^2}{L+1} \\ n\left(2\sqrt{L+1} + 3L + 3 + \frac{1}{L}\right) & \text{at } x_{min} = \frac{L}{\sqrt{L+1}},\ \text{if } \frac{L^2}{L+1} \le x \le 1 \\ 2\,n\left(2\,L+2+\frac{1}{L}\right) & \text{at } x_{min} = 1,\ \text{if } 1 \le x \le L \\ n\left(L^2+7\,L+2\right) & \text{at } x_{min} = L,\ \text{if } L \le x \le \frac{L}{L-1} \\ 2\,n\,\frac{L}{(L-1)^2}\left(L^2+L-1\right) & \text{at } x_{min} = \frac{L}{L-1},\ \text{if } x > \frac{L}{L-1}, \end{cases}$$

and comparing the above values with each other, we obtain that in case **b.1)**, $\min \widehat{\mathrm{TC}}_{max}(x) = n\left(2\sqrt{L+1} + 3\,L + 3 + \frac{1}{L}\right)$, which is reached at $x_{min} = \frac{L}{\sqrt{L+1}}$.

Case b.2) $\frac{1+\sqrt{5}}{2} < L \le 2$

$$\widehat{\mathrm{TC}}_{max} = \begin{cases} n\left(L\,(L+1)\,\frac{1}{x} + L\,(L+3)\right) & \text{if } 0 < x \le 1 \\ n\left(L\,x + \frac{L^2}{x} + L\,(L+3)\right) & \text{if } 1 < x \le \frac{L}{2}\,(\sqrt{5}-1) \\ n\left(\frac{x^2}{L} + (L+2)x + L(L+3)\right) & \text{if } \frac{L}{2}(\sqrt{5}-1) < x \le L(L-1) \\ n\left(\frac{x^2}{L} + (L+3+\frac{1}{L})\,x + (3\,L+1)\right) & \text{if } L\,(L-1) < x \le L \\ n\left((1+\frac{1}{L})\,x^2 + (4+\frac{1}{L})\,x + (2\,L+1)\right) & \text{if } L < x \le \frac{L}{L-1} \\ n\left((1+\frac{1}{L})\,x^2 + (4+L)\,x + L\right) & \text{if } x > \frac{L}{L-1}, \end{cases}$$

which again is a continuous function of $x > 0$ and piece-wise differentiable, whose minimum at each interval of definition is given by:

$$\min \widehat{\mathrm{TC}}_{max}(x) = \begin{cases} 2\,n\,L\,(L+2) & \text{at } x_{min} = 1,\ \text{if } 0 < x \le 1 \\ n\,\frac{L}{2}\left((\sqrt{5}+1)\,L + \sqrt{5} + 7\right) & \text{at } x_{min} = \frac{L}{2}\,(\sqrt{5}-1),\ \text{if } 1 \le x \le L\,(L-1) \\ 2\,n\,L\,(L^2+1) & \text{at } x_{min} = L\,(L-1),\ \text{if } L\,(L-1) \le x \le L \\ n\left(L^2+7\,L+2\right) & \text{at } x_{min} = L,\ \text{if } L \le x \le \frac{L}{L-1} \\ 2\,n\,\frac{L}{(L-1)^2}\left(L^2+L-1\right) & \text{at } x_{min} = \frac{L}{L-1},\ \text{if } x > \frac{L}{L-1}. \end{cases}$$

Comparing the above values with each other, we obtain that, in case **b.2)**, $\min \widehat{\mathrm{TC}}_{max}(x) = n\,\frac{L}{2}\left((\sqrt{5}+1)\,L + \sqrt{5} + 7\right)$ and it is reached at $x_{min} = \frac{L}{2}\,(\sqrt{5}-1)$.

Case b.3) $L > 2$

$$\widehat{\mathrm{TC}}_{max} = \begin{cases} n\left(L\,(L+1)\,\frac{1}{x} + L\,(L+3)\right) & \text{if } 0 < x \le 1 \\ n\left(L\,x + \frac{L^2}{x} + L\,(L+3)\right) & \text{if } 1 < x \le \frac{L}{2}\,(\sqrt{5}-1) \\ n\left(\frac{x^2}{L} + (L+2)\,x + L\,(L+3)\right) & \text{if } \frac{L}{2}\,(\sqrt{5}-1) < x \le L \\ n\left((1+\frac{1}{L})\,x^2 + (4+L)\,x + L\right) & \text{if } x > L, \end{cases}$$

which is continuous as function of $x > 0$, independent of n, and piece-wise differentiable, whose minimum at each interval of definition is given by:

$$\left\{ \begin{array}{l} \text{if } 0 < x \le 1,\ \min \widehat{\mathrm{TC}}_{max}(x) = 2\,n\,L\,(L+2) \quad \text{reached at } x_{min} = 1 \\ \text{if } 1 < x \le \frac{L}{2}\,(\sqrt{5}-1),\ \min \widehat{\mathrm{TC}}_{max}(x) = \\ \quad = \begin{cases} n\,\frac{L}{2}\big((\sqrt{5}+1)\,L + \sqrt{5}+7\big) & \text{at } x_{min} = \frac{L}{2}\,(\sqrt{5}-1),\ \text{if } 2 < L \le \frac{3+\sqrt{5}}{2} \\ n\,L\,\big(2\,\sqrt{L} + L + 3\big) & \text{at } x_{min} = \sqrt{L},\ \text{if } L > \frac{3+\sqrt{5}}{2} \end{cases} \\ \text{if } \frac{L}{2}\,(\sqrt{5}-1) < x \le L, \\ \quad \min \widehat{\mathrm{TC}}_{max}(x) = n\,\frac{L}{2}\big((\sqrt{5}+1)\,L + \sqrt{5}+7\big) \quad \text{reached at } x_{min} = \frac{L}{2}\,(\sqrt{5}-1) \\ \text{if } x > L,\ \min \widehat{\mathrm{TC}}_{max}(x) = 2\,n\,L\,(L+3) \quad \text{reached at } x_{min} = L. \end{array} \right.$$

Acknowledgements The authors would like to thank the editorial team and the anonymous reviewers for their valuable feedback on the first version of this paper. Their comments and suggestions enabled us to make significant improvements in the final version. In particular, Sect. 5 was added following their recommendations, and the rearrangement of the results in Sect. 3, along with adjustments to the appendices, have enhanced the readability of the final manuscript.

Author Contributions The two authors have contributed equally.

Funding Open Access Funding provided by Universitat Autonoma de Barcelona. Both authors are supported by Ministerio de Ciencia e Innovación, Gobierno de España, project ref. PID2021-123733NB-I00.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare that they have no conflict of interest nor conflict of interest in relation to the research reported in this paper.

Financial and non-financial interests. None.

Code availability. The R code for calculating the metrics as well as for replicating all examples in Sect. 4.2 is available at https://github.com/giuliabinotto/IntervalScaleClassification. Additionally, the R scripts used to implement the experimental phase detailed in Sect. 5 can be found at https://github.com/RosDelgado/IntervalScaleClassification_Experimentation.

References

Amigó, E., Gonzalo, J., Mizzaro, S., & Carrillo-de-Albornoz, J. (2020). An effectiveness metric for ordinal classification: Formal properties and experimental results. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 3938–3949). https://doi.org/10.18653/v1/2020.acl-main.363

Ben-David, A. (2008). Comparison of classification accuracy using Cohen's Weighted Kappa. *Expert Systems with Applications, 34*(2), 825–832. https://doi.org/10.1016/j.eswa.2006.10.022

Baccianella, S., Esuli, A., & Sebastiani, F. (2009). Evaluation measures for ordinal regression. In *2009 ninth international conference on intelligent systems design and applications* (pp. 283–287). https://doi.org/10.1109/ISDA.2009.230

Baehre, S., O'Dwyer, M., O'Malley, L., & Lee, N. (2022). The use of Net Promoter Score (NPS) to predict sales growth: insights from an empirical investigation. *Journal of the Academy of Marketing Science, 50*(1), 67–84. https://doi.org/10.1007/s11747-021-00790-2

Cano, J. R., & García, S. (2017). Training set selection for monotonic ordinal classification. *Data & Knowledge Engineering, 112*, 94–105. https://doi.org/10.1016/j.datak.2017.10.003

Cruz-Ramírez, M., Hervás-Martínez, C., Sánchez-Monedero, J., & Gutiérrez, P. A. (2011). A preliminary study of ordinal metrics to guide a multi-objective evolutionary algorithm. 2011 11th International Conference on Intelligent Systems Design and Applications, 1176–1181 https://doi.org/10.1109/ISDA.2011.6121818

Cruz-Ramírez, M., Hervás-Martínez, C., Sánchez-Monedero, J., & Gutiérrez, P. A. (2014). Metrics to guide a multi-objective evolutionary algorithm for ordinal classification. *Neurocomputing, 135*, 21–31. https://doi.org/10.1016/j.neucom.2013.05.058

Cardoso, J. S., & Sousa, R. (2011). Measuring the performance of ordinal classification. *International Journal of Pattern Recognition and Artificial Intelligence, 25*(8), 1173–1195. https://doi.org/10.1142/S0218001411009093

Erbilek, M., Fairhurst, M., & Costa-Abreu, M.C.D. (2013). Age Prediction from Iris Biometrics. 5th International Conference on Imaging for Crime Detection and Prevention, ICDP 2013 https://doi.org/10.1049/ic.2013.0258

Gaudette, L., & Japkowicz, N. (2009). Evaluation methods for ordinal classification. *Lecture Notes in Computer Science, 5549*, 207–210. https://doi.org/10.1007/978-3-642-01818-3_25

Gowroju, S., Kumar, S., Aarti, & Ghimire, A. (2022). Deep Neural Network for accurate age group prediction through Pupils using the Optimised UNet model. *Mathematical Problems in Engineering* Article ID 7813701 24 pages https://doi.org/10.1155/2022/7813701

George, N. I., Lu, T.-P., & Chang, C.-W. (2016). Cost-sensitive performance metric for comparing multiple ordinal classifiers. *Artificial Intelligence Research, 5*(1), 135–143. https://doi.org/10.5430/air.v5n1p135

Jeske, D. R., Callanan, T. P., & Guo, L. (2011). Identification of key drivers of net promoter score using a statistical classification model. *Efficient Decision Support Systems*, IntechOpen, Chapter 8 https://doi.org/10.5772/16954

Liu, X., Fan, F., Kong, L., Diao, Z., Xie, W., Lu, J., & You, J. (2020). Unimodal regularized neuron stick-breaking for ordinal classification. *Neurocomputing, 388*, 34–44. https://doi.org/10.1016/j.neucom.2020.01.025

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*(140), 5–55.

Morgan-López, A. A., Kim, A. E., Chew, R. F., & Ruddle, P. (2017). Predicting age groups of Twitter users based on language and metadata features. *PLoS ONE, 12*(8), 1–12. https://doi.org/10.1371/journal.pone.0183537

Markoulidakis, J., Rallis, I., Georgoulas, I., Kopsiaftis, G., Doulamis, A., & Doulamis, N. (2021). Multi-class confusion matrix reduction method and its application on net promoter score classification problem. In *Proceedings of the 14th peripheral technologies related to assistive environments conference* (*PETRA '21*)(Vol. 9(4), 81) https://doi.org/10.3390/technologies9040081

Peersman, C., Daelemans, W., & Vaerenbergh, L. V. (2011). Predicting age and gender in online social networks. In *Proceedings of the 3rd international CIKM workshop on search and mining user-generated contents* (pp. 37–44). https://doi.org/10.1145/2065023.2065035

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Ravishankar, S., Kumar, P., Patage, V. V., Tiwari, S., & Goyal, S. (2020). Prediction of age from speech features using a multi-layer perceptron model. In *11th international conference on computing, communication and networking technologies* (*ICCCNT*). https://doi.org/10.1109/ICCCNT49239.2020.9225390

Sebastiani, F. (2015). An axiomatically derived measure for the evaluation of classification algorithms. In *Proceedings of the 2015 international conference on the theory of information retrieval* (pp. 11–20). https://doi.org/10.1145/2808194.2809449

Sharma, N., Sharma, R., & Jindal, N. (2021). Prediction of face age progression with generative adversarial networks. *Multimedia Tools and Applications, 80*(25), 33911–33935. https://doi.org/10.1007/s11042-021-11252-w

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science, New Series, 103*(2684), 677–680.

Vargas, V. M., Gutiérrez, P. A., & Hervás-Martínez, C. (2020). Cumulative link models for deep ordinal classification. *Neurocomputing, 401*, 48–58. https://doi.org/10.1016/j.neucom.2020.03.034

Waegeman, W., Baets, B. D., & Boullart, L. (2006). A comparison of different ROC measures for ordinal regression. In *Proceedings of the CML 2006 workshop on ROC analysis in machine learning*

Yilmaz, A. E., & Demirhan, H. (2023). Weighted kappa measures for ordinal multi-class classification performance. *Applied Soft Computing, 134*, 110020. https://doi.org/10.1016/j.asoc.2023.110020