A Uniform Performance Index for Ordinal Classification with Imbalanced Classes

Wilson Silva, João Ribeiro Pinto, and Jaime S. Cardoso
Centre for Telecommunications and Multimedia, INESC-TEC
Faculdade de Engenharia, Universidade do Porto
Porto, Portugal
{wilson.j.silva, joao.t.pinto, jaime.cardoso}@inesctec.pt

***Abstract*—Ordinal classification is a specific and demanding task, where the aim is not only to increase accuracy, but to also capture the natural order between the classes, and penalize incorrect predictions by how much they deviate from this ranking. If an ordinal classifier must be able to comply with all these requirements, a suitable ordinal metric must be able to accurately measure its degree of compliance. However, the current metrics are unable to completely capture these considerations when assessing classification performance. Moreover, most suffer from sensitivity to imbalanced classes, very common in ordinal classification. In this paper, we propose two variants of a novel performance index that accounts for both accuracy and ranking in the performance assessment of ordinal classification, and is robust against imbalanced classes.**

I. Introduction

Classification consists on the attribution of a class to an object given a set of characteristics or features, using a model previously trained with similar data for that purpose. Unlike most cases, where classes are unrelated, ordinal classification problems blur the boundaries between classification and regression. In ordinal classification, although there is a finite set of possible labels like in any classification task, the labels present a natural inherent order among themselves like in regression problems [1]–[3].

It is incautious to objectively state that there is a natural definitive order among cats, dogs, and koalas, but it is undeniable that grade A is superior to grade B and grade C on an exam. While the first example pertains to a nominal classification task, the second illustrates the nature of ordinal classification problems, where labels typically present relationships of superiority and inferiority between them.

This generates extraordinary requirements for classifiers on ordinal contexts. Recalling the example above, misclassifying a cat as a dog or a koala is equally undesirable, but it is much worse to misclassify grade A students as grade C than to attribute them grade B. This means misclassifications should not be treated equally, and their influence should relate to the natural order between classes [3].

Similarly, if we attribute grade B to a grade A student, it would be more adequate and fair to misclassify a grade B student as grade C, than to give it grade A. This reveals another property of ordinal classifiers: misclassifications that preserve the natural order of the labels are more desirable than misclassifications that infringe it.

A good ordinal classifier should address these concerns [1], [4], [5], and a suitable ordinal classification metric should be able to adequately capture the degree to which the classifiers comply to them. Furthermore, the metric should also be robust against common classification issues, such as imbalanced classes [6]. Due to the natural order between classes, imbalanced classes are even more common in ordinal settings [7], with the first and last classes generally being under-represented in samples/datasets.

Furthermore, ordinal classification problems are currently present in all fields of research, from computer vision to social sciences [8], which magnifies the need for adequate performance measurement. In this paper, we aim to fill this void with two variants of a novel index, for performance assessment and comparison of ordinal classification in imbalanced settings, that more closely follows the explained desirable behavior.

II. Current Metrics in Ordinal and Imbalanced Classification

Several metrics are currently used for the measurement of performance of ordinal classifiers. However, each one presents its own weaknesses when dealing with this very specific and demanding scenario.

One of such metrics is the Misclassification Error Rate (MER) (1). Despite considering the accuracy of the predictions, it fails to account for the natural order of the classes by attributing equal cost to all misclassifications, which is undesirable for performance assessment in ordinal classification tasks.

$$MER = \frac{1}{N}\sum_{i=1}^{N} I(y_i \neq \hat{y}_i) \tag{1}$$

Mean Squared Error (MSE) (2) or Mean Absolute Error (MAE) (3) are two of the most common, where higher numerical differences between the actual and predicted labels are reflected on the error, resulting in higher penalization of

This work was partially funded by the Project “NanoSTIMA: Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics/NORTE-01-0145-FEDER-00001” financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

bigger mistakes (such as estimating class $\hat{y} = 5$ to an object of true class $y = 1$) over smaller mistakes (attributing $\hat{y} = 2$ for the same object). The error sum is then averaged over all N observations.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{2}$$

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{3}$$

Nevertheless, both metrics present the significantly disadvantageous dependence on the numbers arbitrarily assigned to each class. This can be fixed by defining the classes by their indexes on a confusion matrix, but MSE and MAE will still equally penalise ”forwards” (estimating a following class) and ”backwards” errors (estimating a previous class). In ordinal classification problems, where ranking plays a major role, this lack of distinction between errors is a significant flaw.

To attend to the relevance of ranking in ordinal classification, one common metric is the Spearman's rank correlation coefficient R_S [9], based on two rank vectors p and q, of length N, associated with the variables y and $\hat{y}$:

$$R_s = \frac{\sum (p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum (p_i - \bar{p})^2 \sum (q_i - \bar{q})^2}}. \tag{4}$$

However, as verifiable on (4), the Spearman's coefficient is still dependent on the values chosen for the ranks to representing the classes.

Kendall's τ_b [10], in turn, also takes into account ranking in the measurement of classification performance, but is independent from the values used to represent each class:

$$\tau_b = \frac{\sum q_{ij} p_{ij}}{\sqrt{\sum q_{ij}^2 \sum p_{ij}^2}}, \tag{5}$$

where q_{ij} behaves as follows:

$$\begin{cases} q_{ij} = 1, & \text{if } q_i > q_j \\ q_{ij} = 0, & \text{if } q_i = q_j \\ q_{ij} = -1, & \text{if } q_i < q_j \end{cases}, \tag{6}$$

and the same is true for p_{ij}.

In the same line of thought, and taking into consideration the high number of cases in which a tie happens, Pinto da Costa *et al.* [11] introduced r_{int}:

$$r_{int} = -1 + 2\frac{\text{card}(S_1 \cap S_2)}{\sqrt{\text{card}(S_1)\text{card}(S_2)}}. \tag{7}$$

Here, the total number of observations whose true class is y_i, $n_{i\bullet}$, is given by $\sum_{j=1}^{K} n_{ij}$, and the total number of observations whose predicted class is y_j, $n_{\bullet j}$, is given by $\sum_{i=1}^{K} n_{ij}$, and we get:

$$\text{card}(S_1) = \sum_{i=1}^{K} \sum_{j=i}^{K} n_{i\bullet} n_{j\bullet} - n, \tag{8}$$

$$\text{card}(S_2) = \sum_{i=1}^{K} \sum_{j=i}^{K} n_{\bullet i} n_{\bullet j} - n, \tag{9}$$

$$\text{card}(S_1 \cap S_2) = \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{i'=i}^{K} \sum_{j'=j}^{K} n_{ij} n_{i'j'} - n. \tag{10}$$

All three indices of similarity, R_S, τ_b and r_{int}, vary between -1 and 1.

However, it is fair to affirm that both Kendall's τ_b and r_{int}, by assuming that the only thing that matters is the order relation between classes, go too far in their quest for abstraction from class labels. The reliance on relative order is beneficial for robust ranking error measurement, but causes critical loss of information on absolute classification error.

The ideal solution would consider both the natural ranking between classes and the absolute classification accuracy on the performance assessment. Considering this, the Ordinal Classification Index was proposed by Cardoso and Sousa [2], fitted for accounting for both absolute classification error and ranking error. With r denoting a row and c a column of the considered confusion matrix, the OC_β^γ was defined as:

$$\begin{aligned} OC_\beta^\gamma &= \min \left\{ 1 - \frac{\text{benefit(path)}}{N + M} + \beta(\text{penalty(path)}) \right\} \\ &= \min \left\{ 1 - \frac{\sum_{(r,c)\in\text{path}} n_{r,c}}{N + \left(\sum_{\forall (r,c)} n_{r,c} |r - c|^\gamma \right)^{1/\gamma}} \right. \\ &\qquad \left. + \beta \sum_{(r,c)\in\text{path}} n_{r,c} |r - c|^\gamma \right\}, \end{aligned} \tag{11}$$

where the minimization is performed over the set of all consistent paths that can be traced over the confusion matrix, from entry $(1, 1)$ to entry (K, K). As defined in [2], a path is consistent if every pair of nodes is nondiscordant, which in turn means that the relative order of the true classes for that pair is not opposite to the relative order of the predicted classes.

Each path is characterized by a benefit and a penalty. The benefit will give advantage to paths that include the largest entries on the confusion matrix, rewarding paths that better follow the natural class order. The penalty will punish paths as they deviate from the main diagonal, effectively acting as a regularizer and including classification accuracy on the performance assessment. The parameter β will weight the benefit and penalty, allowing the metric to focus more on accuracy or ranking.

However, this metric suffers from two main setbacks. First, the freely tunable parameter, β, generates ambiguity as it allows users to choose its value for their own benefit. Second, and similarly to all aforementioned metrics, it is sensitive to imbalanced classes: the influence of each class is not necessarily uniform, and is instead linked to the number of instances of each on the considered population sample.

This implies that, if a class is significantly better represented in the sample than the others, it will have a much higher impact on the metric than the remaining classes, which is generally undesirable. To address this issue, some alternative metrics have been proposed.

If the imbalanced classification problem at hand is binary, then two metrics are commonly used: the F_1 (12) and the G-mean (13). However, they are largely limited by being solely applicable to binary classification.

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{12}$$

$$\text{G-mean} = \sqrt{\frac{TP}{TP + FN}(1 - \frac{FP}{FP + TN})} \tag{13}$$

For imbalanced problems with $K > 2$ classes, several adaptations of MAE have been proposed to more robustly assess the performance of the classifiers. Namely, Maximum (MMAE) [12], and Average Mean Absolute Error (AMAE) [13], presented in (14) and (15), respectively. With these, the error is determined separately for each class, and the metric assumes the maximum or average of the values obtained, effectively enforcing uniform influence of all classes, independently of possible imbalanced representations.

$$MMAE = \max\{MAE_k; k = 1, ..., K\} \tag{14}$$

$$AMAE = \frac{1}{K}\sum_{k=1}^{K} MAE_k \tag{15}$$

Nevertheless, these metrics, more devoted for imbalanced problems, present the major weakness of overlooking the importance of ranking for ordinal classification. Furthermore, the adaptations of Mean Absolute Error still depend on the values chosen as labels for each class.

Considering this current panorama in ordinal imbalanced metrics here presented, it is possible to conclude that there is still no metric that can adequately combine classification accuracy and ranking in the same metric, while remaining robust against the influence of imbalanced classes. In the following sections, two variants of a new index, based on the aforementioned Ordinal Classification Index, are formulated and proposed to fill this void.

III. The Proposed Performance Index

A. Conceptual formulation

The proposed index results of the adaptation of the aforementioned Ordinal Classification Index, OC^γ_β, proposed by Cardoso and Sousa [2], and aims towards the achievement of robustness against imbalanced classes, and the suppression of the freely tunable parameter β. First, to fix the weakness related to imbalanced classes, we start by re-interpreting the OC^γ_β in a stochastic formulation. Towards that goal, a simple algebraic manipulation of OC^γ_β gives:

$$OC^\gamma_\beta = \min\Bigg\{1 - \frac{\sum_{(r,c)\in\text{path}} \frac{n_{r,c}}{N}}{\sum_{\forall(r,c)} \frac{n_{r,c}}{N} + \left(\sum_{\forall(r,c)} \frac{n_{r,c}}{N}|r-c|^\gamma\right)^{1/\gamma}} + N\beta \sum_{(r,c)\in\text{path}} \frac{n_{r,c}}{N}|r-c|^\gamma\Bigg\}, \tag{16}$$

where $\beta \in \mathbb{R}_{\geq 0}$ and $\gamma \in \mathbb{R}_{>0}$.

Interpreting Y and $\hat{Y}$ as random variables, the normalized confusion matrix with entries $\frac{n_{r,c}}{N}$ can be understood as an approximation of the joint probability function between Y and $\hat{Y}$, $p(y, \hat{y})$. Adopting this stochastic view, OC^γ_β can be written as:

$$OC^\gamma_\beta = \min\Bigg\{1 - \frac{\sum_{(y,\hat{y})\in\text{path}} p(y,\hat{y})}{\sum_{\forall(y,\hat{y})} p(y,\hat{y}) + \left(\sum_{\forall(y,\hat{y})} p(y,\hat{y})|y-\hat{y}|^\gamma\right)^{1/\gamma}} + N\beta \sum_{(y,\hat{y})\in\text{path}} p(y,\hat{y})|y-\hat{y}|^\gamma\Bigg\}, \tag{17}$$

where $(y, \hat{y})$ is equivalent to the notation (r, c), used for confusion matrices.

Since, for any pair of random variables A and B, the joint probability can be written as $p(a, b) = p(a)p(b|a)$, we have:

$$OC^\gamma_\beta = \min\Bigg\{1 - \frac{\sum_{y:(y,\hat{y})\in\text{path}} \sum_{\hat{y}} p(y)p(\hat{y}|y)}{\sum_y \sum_{\hat{y}} p(y)p(\hat{y}|y) + \left(\sum_y \sum_{\hat{y}} p(y)p(\hat{y}|y)|y-\hat{y}|^\gamma\right)^{1/\gamma}} + N\beta \sum_{y:(y,\hat{y})\in\text{path}} \sum_{\hat{y}} p(y)p(\hat{y}|y)|y-\hat{y}|^\gamma\Bigg\}. \tag{18}$$

This is equivalent to:

$$OC^\gamma_\beta = \min\Bigg\{1 - \frac{\sum_{y:(y,\hat{y})\in\text{path}} p(y) \sum_{\hat{y}} p(\hat{y}|y)}{\sum_y p(y) \sum_{\hat{y}} p(\hat{y}|y) + \left(\sum_y p(y) \sum_{\hat{y}} p(\hat{y}|y)|y-\hat{y}|^\gamma\right)^{1/\gamma}} + N\beta \sum_{y:(y,\hat{y})\in\text{path}} p(y) \sum_{\hat{y}} p(\hat{y}|y)|y-\hat{y}|^\gamma\Bigg\}. \tag{19}$$

In (19), the dependency of OC^γ_β on the class distribution $p(y)$ is evident. When classes are highly imbalanced, classes with high probability dominate the result. Like AMAE brings robustness to the MAE metric in imbalance settings by replacing the original $p(y)$ distribution with uniform probabilities $1/K$ for each class, we propose to modify OC^{γ}_{β} using the same strategy. Thus, we propose the first variant of our index, the Uniform Ordinal Classification Index, UOC^{γ}_{β}, as:

$$UOC^{\gamma}_{\beta} = \min\Bigg\{1 - \frac{\sum_{y:(y,\hat{y})\in \text{path}} \frac{1}{K}\sum_{\hat{y}} p(\hat{y}|y)}{\sum_{y}\frac{1}{K}\sum_{\hat{y}} p(\hat{y}|y) + \left(\sum_{y}\frac{1}{K}\sum_{\hat{y}} p(\hat{y}|y)|y-\hat{y}|^{\gamma}\right)^{1/\gamma}} + N\beta \sum_{y:(y,\hat{y})\in \text{path}} \frac{1}{K}\sum_{\hat{y}} p(\hat{y}|y)|y-\hat{y}|^{\gamma}\Bigg\}, \quad (20)$$

which can be simplified to:

$$UOC^{\gamma}_{\beta} = \min\Bigg\{1 - \frac{\sum_{(y,\hat{y})\in \text{path}} p(\hat{y}|y)}{K + \frac{K}{K^{\gamma}}\left(\sum_{\forall (y,\hat{y})} p(\hat{y}|y)|y-\hat{y}|^{\gamma}\right)^{1/\gamma}} + \frac{N}{K}\beta \sum_{(y,\hat{y})\in \text{path}} p(\hat{y}|y)|y-\hat{y}|^{\gamma}\Bigg\}. \quad (21)$$

As β is, still, a user defined constant, it is possible to recast $N\beta$ as β, finally giving the proposed formulation to UOC:

$$UOC^{\gamma}_{\beta} = \min\Bigg\{1 - \frac{\sum_{(y,\hat{y})\in \text{path}} p(\hat{y}|y)}{K + \frac{K}{K^{\gamma}}\left(\sum_{\forall (y,\hat{y})} p(\hat{y}|y)|y-\hat{y}|^{\gamma}\right)^{1/\gamma}} + \frac{\beta}{K} \sum_{(y,\hat{y})\in \text{path}} p(\hat{y}|y)|y-\hat{y}|^{\gamma}\Bigg\}. \quad (22)$$

Following a procedure similar to [2], it is possible to show that for $\beta \geq 1$, UOC^{γ}_{β} in (22) results in a metric and the optimal path is always over the main diagonal. Thus, considering $\beta \in [0,1]$, for specific settings that require especial emphasis in either ranking error or instance-based error, the variant UOC^{γ}_{β} can be used with a user-defined β in the lower or higher end, respectively, of its range.

Nevertheless, the existence of β and γ remains a source of ambiguity in most applications. For γ, we propose the value 1, as used for OC, as the Minkowski distance is generally used for the values of 1, 2, or infinity, and the variation of the results with different γ values will not be significant [2]. For β, we propose its elimination through the formulation of a second variant, with the integration of UOC^{1}_{β} along β's aforementioned range of values, through:

$$A_{UOC} = \int_{0}^{1} UOC^{1}_{\beta}\, d\beta \quad (23)$$

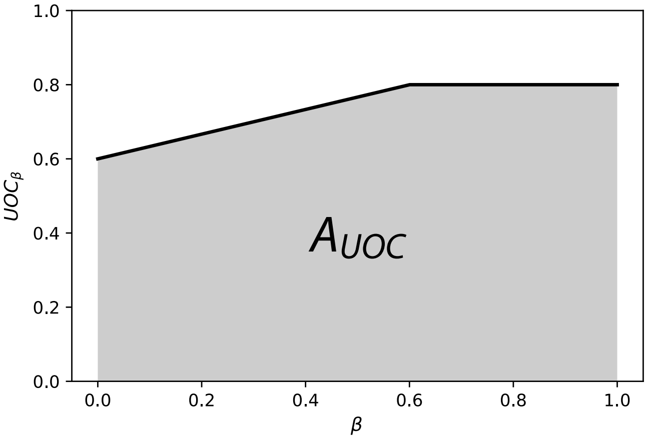

Fig. 1. Illustration of A_{UOC} and the values of UOC^{1}_{β} obtained from an example confusion matrix.

B. Application from estimates in a confusion matrix

When applied to a real scenario, $p(y,\hat{y})$ and $p(y)$ can easily be estimated through Maximum Likelihood Estimation, and UOC^{γ}_{β} can be applied from a confusion matrix.

$$UOC^{\gamma}_{\beta} = \min\Bigg\{1 - \frac{\sum_{(y,\hat{y})\in \text{path}} \hat{p}(\hat{y}=c|y=r)}{K + \frac{K}{K^{\gamma}}\left(\sum_{\forall (y,\hat{y})} \hat{p}(\hat{y}=c|y=r)|y-\hat{y}|^{\gamma}\right)^{1/\gamma}} + \frac{\beta}{K} \sum_{(y,\hat{y})\in \text{path}} p(\hat{y}=c|y=r)|y-\hat{y}|^{\gamma}\Bigg\} \quad (24)$$

$$UOC^{\gamma}_{\beta} = \min\Bigg\{1 - \frac{\sum_{(r,c)\in \text{path}} n_{r,c}/N_r}{K + \frac{K}{K^{\gamma}}\left(\sum_{\forall (r,c)} (n_{r,c}/N_r)|r-c|^{\gamma}\right)^{1/\gamma}} + \frac{\beta}{K} \sum_{(r,c)\in \text{path}} \frac{n_{r,c}}{N_r}|r-c|^{\gamma}\Bigg\} \quad (25)$$

With (25), it is also easy to integrate and obtain A_{UOC} from a confusion matrix. In Fig. 1, we illustrate A_{UOC} and the values of UOC^{1}_{β} obtained from an example confusion matrix.

C. Handling unobserved classes

One particular issue can arise from the application of UOC^{γ}_{β} to confusion matrices: it is not guaranteed that every class will be observed in the considered set/sample, especially if the latter is small, and N_r can, in some cases, be zero. We propose to generalise (22) to attend to these situations.

Let $\mathbb{1}_{\mathscr{O}}$ be the indicator function of the set $\mathscr{O}$ of the observed classes:

$$\mathbb{1}_{\mathscr{O}}(y) = \begin{cases} 1 & y \in \mathscr{O} \\ 0 & y \notin \mathscr{O} \end{cases}, \quad (26)$$

and $K' \leq K$ the cardinality of $\mathscr{O}$ (the number of observed classes). In order to make (19) robust in imbalanced settings and to address unobserved classes, we propose to fix the probability distribution for y to an uniform distribution over the observed classes only, with $p(y) = \frac{1}{K'}\mathbb{1}_{\mathscr{O}}(y)$. Introducing this proposed distribution in (19) and simplifying as before, one obtains

$$UOC_\beta^\gamma = \min\Bigg\{1 - \frac{\sum_{(y,\hat{y})\in\text{path}} p(\hat{y}|y)\mathbb{1}_{\mathscr{O}}(y)}{K' + \frac{K'}{K'^\gamma}\left(\sum_{\forall(y,\hat{y})} p(\hat{y}|y)\mathbb{1}_{\mathscr{O}}(y)|y-\hat{y}|^\gamma\right)^{1/\gamma}} + \frac{\beta}{K'}\sum_{(y,\hat{y})\in\text{path}} p(\hat{y}|y)\mathbb{1}_{\mathscr{O}}(y)|y-\hat{y}|^\gamma\Bigg\}. \quad (27)$$

For real scenarios, in place of (25), we can rewrite (27) in order to make it robust against unobserved classes while using estimates from a confusion matrix, and we obtain:

$$UOC_\beta^\gamma = \min\Bigg\{1 - \frac{\sum_{(r,c)\in\text{path}} (n_{r,c}/N_r)\mathbb{1}_{\mathscr{O}}(r)}{K' + \frac{K'}{K'^\gamma}\left(\sum_{\forall(r,c)} (n_{r,c}/N_r)\mathbb{1}_{\mathscr{O}}(r)|r-c|^\gamma\right)^{1/\gamma}} + \frac{\beta}{K'}\sum_{(r,c)\in\text{path}} \frac{n_{r,c}}{N_r}\mathbb{1}_{\mathscr{O}}(r)|r-c|^\gamma\Bigg\}, \quad (28)$$

that can be similarly used in (23) for settings that do not present a preferential value for β.

All this effectively amounts to ignore classes that are not observed in the considered sample. Alternatives such as considering uniform conditional probabilities on those cases presents the disadvantage of consistently penalizing performance because of each unobserved class. On the other hand, assuming perfect performance for each unobserved class is overly optimistic, as it rarely will be true. Our proposal avoids generating tendencies to either benefit or penalise performance due to unobserved classes, and instead bases it entirely on the classes that are observed.

IV. Experimental Study

A. Single Sample and Tridiagonal Matrices

As stated by Cardoso and Sousa [2], one of the weaknesses between τ_b, R_s, or r_{int} and MER, MAE, or MSE, is that the former are not applicable to performance assessment with a single observation.

Similarly to OC_β^γ, UOC_β is applicable to single observations, and its value increases monotonically from 0 to 1 with the increase of the sample's distance to the diagonal, and the rate is dependent from the chosen value of β. The value of A_{UOC}, although independent from β, presents similar behavior (cf. Fig. 2).

One other issue of r_{int}, R_s, and τ_b, is the result of their application to tridiagonal matrices (cf. Fig. 3). These confusion matrices present zeros on all entries except their three main diagonals, where the entries are 1. As K, the number of classes, increases, the three aforementioned metrics converge to 1. To affirm a certain score is the most appropriate for this situation would be reckless, as the relevance on the performance of the two diagonals (other than the main one) is subjective. Nevertheless, attributing a near-perfect performance to classifiers that, simultaneously, present a MER of $2/3$, is clearly inappropriate. The proposed variants, UOC_β^γ and A_{UOC}, present an intermediate behavior between the remaining metrics, while steering away from the undesirable behavior of r_{int}, R_s, and τ_b.

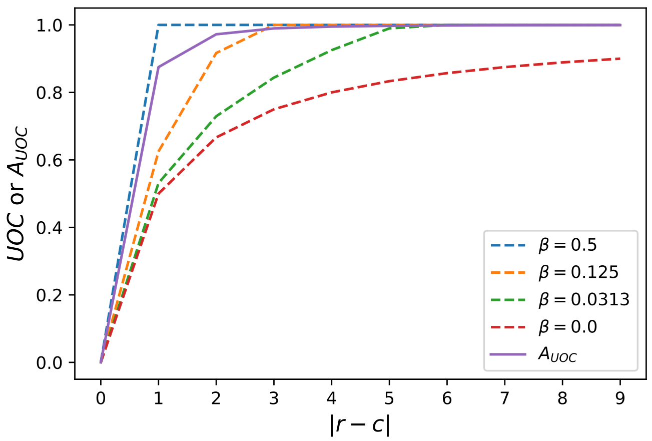

Fig. 2. Values of UOC, for several β values, and A_{UOC}, obtained with a confusion matrix with a single sample, according to its distance to the diagonal $|r-c|$.

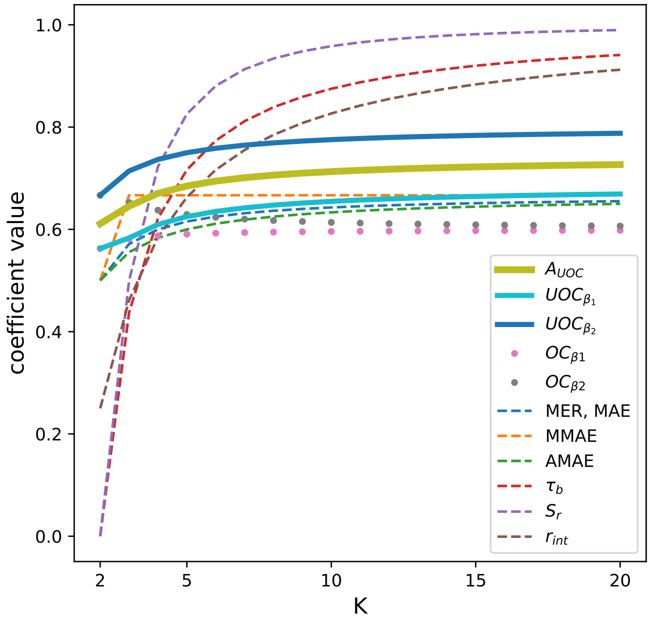

Fig. 3. Evolution of metric values with the total number of classes K, when applied to tridiagonal matrices.

TABLE I
RESULTS FOR THE SIMULATED CONFUSION MATRICES, WITH $\beta_1 = 0.25$ AND $\beta_2 = 0.75$

Classifier	**Accuracy-focused**					**Ranking-Focused**			**Mixed Focus**				
	Sensitive to imbalance			***Robust to imbalance***					***Sensitive to imbalance***		***Robust to imbalance***		
	MER	MSE	MAE	MMAE	AMAE	R_s	τ_b	r_{int}	$OC^1_{\beta_1}$	$OC^1_{\beta_2}$	UOC_{β_1}	UOC_{β_2}	A_{UOC}
A	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00
B	0.56	0.56	0.56	1.00	0.50	0.90	0.86	0.86	0.40	0.50	0.46	0.67	0.56
C	0.56	1.22	0.78	2.00	0.75	0.67	0.61	0.69	0.50	0.63	0.62	0.71	0.65
D	0.56	0.56	0.56	1.00	0.50	0.73	0.60	0.74	0.53	0.58	0.56	0.67	0.61
E	0.77	0.77	0.77	1.00	0.50	0.24	0.11	0.53	0.65	0.72	0.68	0.80	0.74
F	0.85	0.85	0.85	1.00	0.50	0.29	0.23	0.79	0.58	0.71	0.56	0.67	0.61

B. Simulated Examples, Missing, and Imbalanced Classes

To show that the proposed index variants combine both accuracy and ranking in the performance assessment, while remaining robust to missing classes and imbalance, the following simulated confusion matrices ($K = 4$), each one representing the behavior of a classifier, were considered. A comparison between the different metrics is also presented in Table I.

$$CM_A = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix} CM_B = \begin{bmatrix} 0 & 4 & 0 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

$$CM_C = \begin{bmatrix} 0 & 0 & 4 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix} CM_D = \begin{bmatrix} 0 & 4 & 0 & 0 \\ 6 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

$$CM_E = \begin{bmatrix} 0 & 4 & 0 & 0 \\ 6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix} CM_F = \begin{bmatrix} 0 & 40 & 0 & 0 \\ 6 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

All compared metrics attribute a perfect performance score to classifier A. This is expectable, as a perfect accuracy implies the absence of ranking error. However, the classifiers B, C, and D present classification errors.

In an accuracy perspective, B, C, and D present equal MER, but classifier C has higher MSE and MAE. Regarding ranking, B clearly resembles more closely the true order of the classes than the other two classifiers. Thus, the ranking-focused metrics, r_{int}, R_s, and τ_b, attribute worse performance scores to D than B, and τ_b goes even further and gives the lowest score to C as it presents lower ranking error than D, despite the lower accuracy.

All accuracy-focused metrics disregard ranking and give equal scores to B and D, and a lower score to C. UOC retains some similarity to OC, as both have the flexibility to resemble ranking-focused metrics for low-range values of β, and to focus on accuracy with higher β values. As expectable, A_{UOC} presents an intermediate behavior.

Finally, E and F can be considered similar to classifier D. However, classifier E was tested without objects of class 3, and the dataset used to evaluate classifier F is highly imbalanced. Most existing metrics present sensitivity to imbalanced classes, as they do not attribute equal scores to classifiers D and F, as they should. The exceptions are MMAE, AMAE, and the proposed variants UOC and A_{UOC}.

Nevertheless, while other metrics, including the proposed ones, will penalize classifier E due to the missing class, MMAE and AMAE assume an optimistic scenario (no error on the missing classes), which may be rarely true. The proposed index, as stated before, deals with missing classes in a balanced fashion, by ignoring them completely. In this case, this results on a slight performance penalisation, as two thirds of the classes do not conform to ranking order or accuracy, while for D it is only one half.

V. EXPERIMENTS ON REAL CLASSIFIERS

To showcase the behavior of the proposed index on real situations, and compare it with the aforementioned state-of-the-art alternatives, we trained a Support Vector Machine, a k-Nearest Neighbors, and a Random Forest classifier on 70% of the data of two real public datasets of ordinal classification problems, with imbalanced classes. The predictions of each classifier on the remaining 30% of each dataset were used to build confusion matrices (cf. Figures 4 and 5) and compute the metrics (cf. Table II). Here, our goal was not to assert the superiority or inferiority of any classifier over the others, but to showcase how each metric allows us to measure and compare their performances based on the resulting confusion matrices.

A. Wine Quality Dataset

This dataset relate eleven numerical features (such as acidity, sulphates, density, pH, and residual sugar) of Portuguese red wines with its quality ranking [14]. The dataset includes 1599 samples for quality classes three to eight, and is available on Kaggle datasets[1].

Analysing the regular confusion matrices of the classifiers, all three appear to have a very similar performance. Nevertheless, when inspecting the normalized confusion matrices, it can be clearly concluded that the predictions of the Random Forest classifier resembles much more closely the true ranking order of the classes, and the overall accuracy inside each class is higher. SVM clearly presented the worst results, while kNN showed an intermediate performance.

[1] Red Wine Quality - Kaggle datasets. Available at: https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009.

TABLE II
RESULTS FOR THE CLASSIFIERS IN REAL DATASETS, WITH $\beta_1 = 0.25$ AND $\beta_2 = 0.75$

Classifier	Accuracy-focused					Ranking-Focused			Mixed Focus				
	Sensitive to imbalance			***Robust to imbalance***					***Sensitive to imbalance***		***Robust to imbalance***		
	MER	MSE	MAE	MMAE	AMAE	R_s	τ_b	r_{int}	$OC^1_{\beta_1}$	$OC^1_{\beta_2}$	UOC_{β_1}	UOC_{β_2}	A_{UOC}
Wine kNN	0.38	0.58	0.44	3.00	1.10	0.57	0.52	0.64	0.46	0.48	0.76	0.82	0.79
Wine SVM	0.37	0.50	0.41	3.00	1.27	0.53	0.50	0.65	0.42	0.44	0.82	0.87	0.84
Wine RF	0.33	0.38	0.34	2.00	0.88	0.66	0.62	0.71	0.37	0.39	0.70	0.81	0.75
ESL kNN	0.37	0.39	0.37	1.00	0.50	0.91	0.84	0.81	0.39	0.40	0.54	0.72	0.63
ESL SVM	0.31	0.80	0.40	5.00	1.11	0.83	0.79	0.78	0.37	0.38	0.75	0.80	0.78
ESL RF	0.37	0.46	0.40	2.00	0.61	0.91	0.83	0.81	0.41	0.42	0.63	0.73	0.68

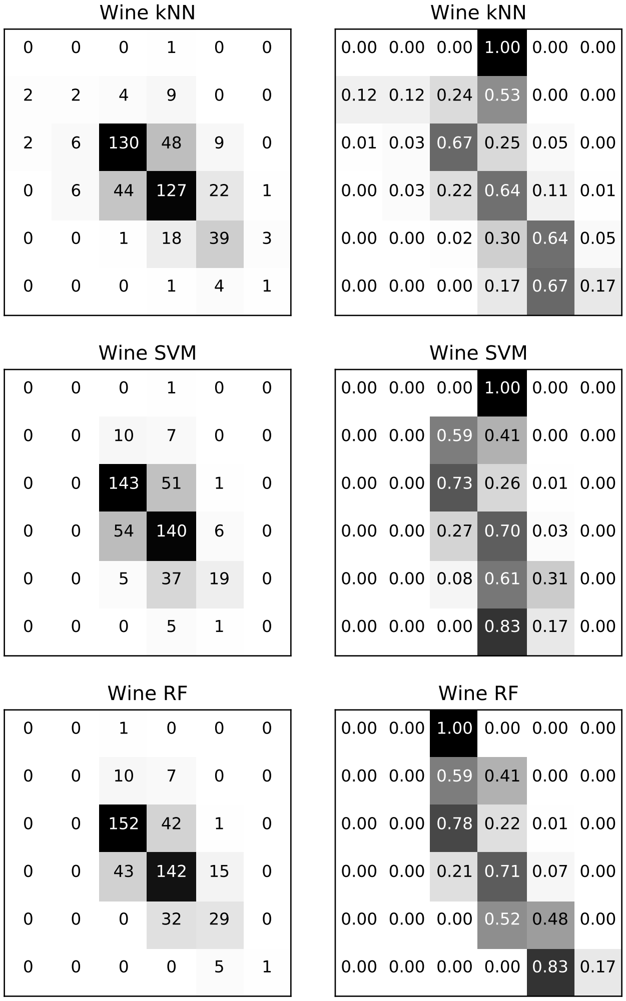

Fig. 4. Confusion matrices, regular and normalized, for the classifiers kNN, SVM, and Random Forest, used on the wine quality dataset.

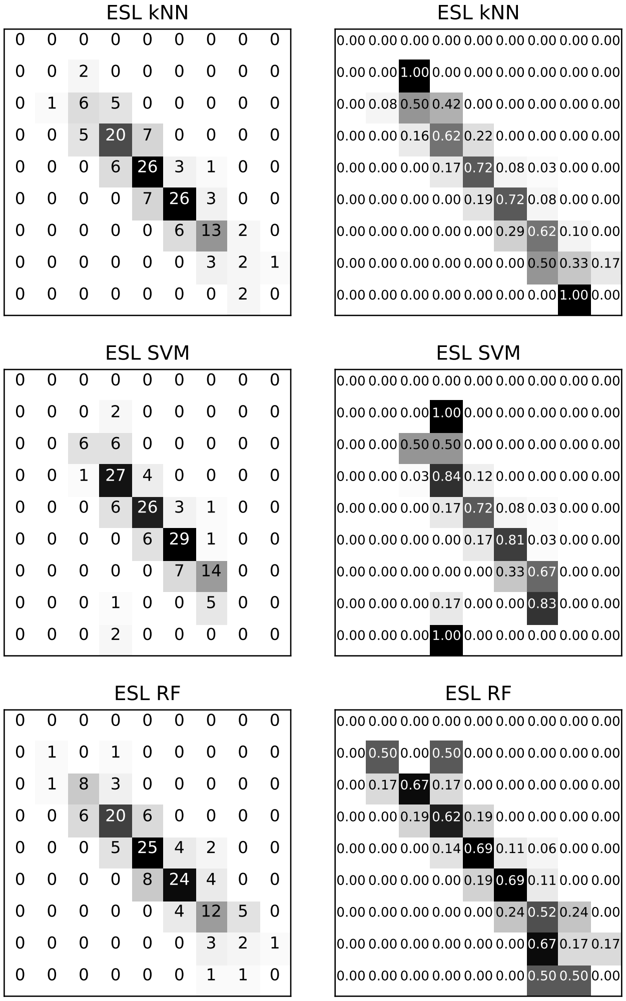

Fig. 5. Confusion matrices, regular and normalized, for the classifiers kNN, SVM, and Random Forest, used on the ESL dataset.

MER, by considering equal all misclassifications, attributed similar scores to all classifiers, with Random Forest only slightly better. MSE and MAE, sensitive to imbalance, attributed a worse score to kNN than SVM. The same was verified for the ranking-focused r_{int} and the mixed-focus metrics $OC^1_{\beta_1}$, and $OC^1_{\beta_2}$, denoting that this undesirable behavior is probably due to their sensitivity to imbalanced datasets.

MMAE presented flaws in its claim to be robust against imbalanced classes, since the scores it presents are clearly a major result of the worse represented class 1, with one object that is misclassified by all classifiers. On the other hand, AMAE shows the desired behavior, but the non-normalised values it takes are not fit for absolute performance assessment, and unfortunately limit its use to the relative comparison of classifiers in equal settings.

Both proposed variants UOC_β and A_{UOC} present the desired behavior. UOC_β presents the advantage of flexibility: with the lower value of β, the index favored ranking and the difference between the classifiers' performance scores was amplified; and with the higher value of β, the focus on accuracy increased and the behavior of UOC approached that of an accuracy-focused metric robust to imbalance. On the other hand, A_{UOC} presented an intermediate behavior, ideal for situations where neither ranking nor accuracy should be especially favored.

B. Employee Selection (ESL) Dataset

The ESL dataset includes numerical evaluations of 488 job applicants in four relevant psychometric parameters, and a final ordinal classification of the applicants according to their fit to the job (from 1 up to 9). The dataset belongs to the Business Administration School of the Tel Aviv University, and is available at Weka datasets[2].

Again, looking at the regular confusion matrices of the classifiers gives an impression of similarity between performances. However, the normalized confusion matrices dissipate this idea. Regarding the true ranking order of the classes, it is clear the distinction between SVM and the other two classifiers. In terms of pure ranking, kNN and RF have similar performance and both present better results than SVM, which has the best result when using MER, a completely accuracy-focused metric. Furthermore, MAE was not able to differentiate between the SVM and the RF.

Mixed-focus metrics $OC^1_{\beta_1}$ and $OC^1_{\beta_2}$ erroneously consider the SVM as the best classifier. This happens due to a class imbalance in the ESL dataset. On the contrary, UOC_{β_1} and UOC_{β_2}, being robust to class imbalance, acknowledge the best performance of the kNN. A_{UOC}, which represents an equilibrium between ranking and accuracy, is also capable of distinguishing the performances between the three classifiers and is in agreement with UOC_{β_1} and UOC_{β_2}.

VI. Conclusion

In this paper, two variants of a novel index for performance assessment of ordinal classification in imbalanced settings are proposed. The first, UOC^γ_β, is tunable to give preference to ranking or accuracy error, and thus allows for tailored performance assessment to fit settings that present such preferences. For a fixed, parameter-free performance assessment, A_{UOC} presents an intermediate behavior.

The proposed index was evaluated and compared with state-of-the-art alternatives in several simulated and real scenarios. The results show that its variants, unlike most other alternatives, are able to capture both ranking and instance-based error in the performance assessment, while remaining impervious against imbalanced classes, and presenting a desirable behavior when faced with unobserved classes.

Thus, it can be concluded that the proposed metrics are suitable to be applied in the complete absolute assessment of classification performance, as well as the adequate and robust relative comparison between sets of ordinal classifiers in imbalanced settings.

References

[1] E. Frank and M. Hall, "A simple approach to ordinal classification," in *Machine Learning: ECML 2001*. Springer, 2001, pp. 145–156.

[2] J. S. Cardoso and R. Sousa, "Measuring the performance of ordinal classification," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 25, no. 8, pp. 1173–1195, 2011.

[3] P. A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernández-Navarro, and C. Hervás-Martínez, "Ordinal Regression Methods: Survey and Experimental Study," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 1, pp. 127–146, Jan 2016.

[4] J. S. Cardoso and J. F. P. da Costa, "Learning to classify ordinal data: the data replication method," *Journal of Machine Learning Research*, vol. 8, pp. 1393–1429, 2007.

[5] J. Costa and J. S. Cardoso, "oAdaBoost: An AdaBoost variant for Ordinal Classification," in *Proceedings of the International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, 2015, pp. 68–76.

[6] R. Cruz, K. Fernandes, J. S. Cardoso, and J. F. P. Costa, "Tackling class imbalance with ranking," in *Neural Networks (IJCNN), 2016 International Joint Conference on*, 2016, pp. 2182–2187.

[7] M. Pérez-Ortiz, P. A. Gutiérrez, C. Hervás-Martínez, and X. Yao, "Graph-Based Approaches for Over-Sampling in the Context of Ordinal Regression," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 5, pp. 1233–1245, May 2015.

[8] K. Antoniuk, V. Franc, and V. Hlavác, "Interval insensitive loss for ordinal classification," in *Asian Conference on Machine Learning*, 2015, pp. 189–204.

[9] C. Spearman, "The Proof and Measurement of Association between Two Things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.

[10] M. G. Kendall, "A New Measure of Rank Correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, June 1938.

[11] J. F. P. da Costa, H. Alonso, and J. S. Cardoso, "The unimodal model for the classification of ordinal data," *Neural Networks*, vol. 21, pp. 78–91, 2008.

[12] M. Cruz-Ramírez, C. Hervás-Martínez, J. Sánchez-Monedero, and P. Gutiérrez, "Metrics to guide a multi-objective evolutionary algorithm for ordinal classification," *Neurocomputing*, vol. 135, pp. 21 – 31, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231213011399

[13] S. Baccianella, A. Esuli, and F. Sebastiani, "Evaluation measures for ordinal regression," in *2009 Ninth International Conference on Intelligent Systems Design and Applications*, Nov 2009, pp. 283–287.

[14] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems*, vol. 47, no. 4, pp. 547–553, 2009.

[2] Dr. Arie Ben David ordinal datasets - Weka datasets. Available at: http://weka.wikispaces.com/Datasets.